universität wien

DISSERTATION

Titel der Dissertation

„Parts of speech in English as a lingua franca: the POS tagging of VOICE“

Verfasserin

Mag.phil. Ruth Osimk-Teasdale

angestrebter akademischer Grad

Doktorin der Philosophie (Dr.phil.)

Wien, 2015

Studienkennzahl lt. Studienblatt:	A 092 343
Dissertationsgebiet lt. Studienblatt:	Anglistik und Amerikanistik
Betreuerin / Betreuer:	Univ.-Prof. Dr. Barbara Seidlhofer

Table of contents

Abstract

This PhD thesis addresses the conceptual and practical challenges involved in the part-of-speech tagging (POS tagging) of the Vienna-Oxford International Corpus of English (VOICE). There has not been an attempt ever before to POS tag naturally occurring, transcribed spoken interactions of English as a lingua franca (ELF) as in VOICE, which is not only characterised by features of spoken language but also by heightened variability and plurilingual language use.

After a general introduction to relevant issues to this thesis in chapters 1 and 2, chapter 3 provides a review of previous approaches to POS tagging spoken language. It is shown that these do not sufficiently address the issues involved in tagging VOICE. Rather, existing tagging methods tend to gloss over issues which are relevant to dynamic spoken discourse and, hence, needed to be extended and, in some cases, discarded entirely, for the tagging of VOICE. The issues which arose in the POS tagging of VOICE and the, sometimes unpreceded, ways of handling of these are discussed in chapter 4, which concludes with the main guidelines developed for tagging VOICE. Chapter 5 then illustrates the advantages of this unique approach to POS tagging by applying it to the specific issue of word class variation. It is shown that the tagging scheme developed for VOICE can yield insights into the nature of variable, spoken language data by investigating those cases in which the ELF users in VOICE go beyond conventional word class boundaries.

It is argued that the strength of the 'linguistic' approach chosen for tagging VOICE is that it views the annotation of such data as a discovery procedure and welcomes rather than glosses over, any challenging issues. As such, it brings out clearly complex aspects of spoken, plurilingual language in use and facilitates their investigation. Moreover, it links the practical task of POS tagging to conceptual issues regarding the categorisation of spoken discourse and can serve as a starting point for the investigation of similar types of data.

Acknowledgements

my_ PP$ supervisor_NNS Barbara_NP and_CC Prof._NP Henry_NP
for the amazing support in terms of helpful comments, feedback and discussion they have given me throughout the development of this thesis, often surprisingly promptly given their busy schedules and for being incredibly encouraging and generous in many ways,

proofreaders_NNS Dagmar_NP Ian_NP Julia _NP Marie-Luise_NP and_CC Nora_NP
for their indispensable help and excellent work,

my_PP$ colleagues_NNS and_CC fellow_JJ researchers_NNS
the original VOICE team **Angelika_NP Marie-Luise_NP Stefan_NP and_CC Theresa_NP** for letting me share their adventure and for friendships that exceed the office life as well as their manifold contributions to VOICE POS,
the_DT BS_NP team_NP for being such enjoyable colleagues and for all their support, and especially **Nora_NP** and **Cornelia_NP** for making sharing an office so enjoyable when I was the only 'voice' left, and their help with VOICE POS in the meantime, particularly Nora for working with us in the 'hot phase' of manual annotation.
Michael_NP for putting an awful lot of time and effort into the technical side of VOICE POS and
the_DT staff_NN at the English department for the interesting discussions, support when needed, the productive and friendly atmosphere and the enjoyable lunchtimes.

friends_NNS in different places: in Vienna – especially **Kat_NP** for enjoyable and stimulating conversations, **the_DT Salzburg_NPS crowd_NN**, and all newly found friends in Linz. You know who you are and I am so glad we are doing life together.

my_ PP$ 0_PA extended 0_PA family_NN, in particular

my_ PP$ parents_NNS Gabi_NP and_CC Fritz_NP
for believing that a good education is the best heritage, for their support in many ways throughout my studies (including some babysitting during supervision meetings) and always believing in me and what I wanted to do, even though linguistics might not have seemed like the most lucrative of all studies – and for always showing that they are proud of me. I love you!

Judith_NP and_CC Markus_NP
for giving us some very needed time out every now and then so we could go for meals, to concerts or weddings, and for always bringing cake ;)

Marie_NP and_CC Dave_NP
for looking after all three of us so well while we were in England, including doing our washing, as well as providing food and a home for several weeks, and thus adding to making the writing a much less stressful and more enjoyable experience.

the_DT Crees_NPS
for letting me use their house as a working place while in the UK and for making me dinner.
Katie_NP for allowing me to work in her room and
Rachel_NP for being a very pleasant (often invisible) roomie.

the_DT rest_NN of_PP the_DT Teasdale_NP clan_NN
for asking how I was getting on and caring.

Ian_NP
who was an incredible support and always believed I should do this. For sharing the parenting and giving me space to write even with a small baby and while he was finishing his own habilitation thesis, as well as for enduring the more stressful times and sharing the joys of the smaller successes and greater victories. I am one lucky lady!

Lotta_NP our_ PP$ amazing_JJ daughter_NN
for teaching me on a daily basis what matters most in life.

List of abbreviations

ACE	Asian Corpus of English
ANC	American National Corpus
BELF	Business English as a Lingua Franca
BNC	British National Corpus
BoE	Bank of English
CANCODE	Cambridge and Nottingham Corpus of Discourse in English
CASE	Corpus of Academic Spoken English
CEA	Computer-aided error analysis
CHILDES	Child Language Data Exchange System
CLAWS	Constituent-Likelihood Automatic Word Tagging System
COCA	Corpus of Contemporary American English
EAGLES	Expert Advisory Group on Language Engineering Standards
ELF	English as a Lingua Franca
ELFA	Corpus of English as a Lingua Franca in Academic Settings
ENL	English as a native language
GSLC	Gothenburg Spoken Language Corpus
HMM	Hidden Markov Models
ICE	International Corpus of English
ICE-EA	ICE East Africa
ICE-HK	ICE Hong Kong
KidKo	Kiezdeutsch-Korpus
L1	First language
L2	Second/foreign/additional/non-L1 language
LDC	Linguistic Data Consortium
LINDSEI	Louvain International Database of Spoken English Interlanguage
LINDSEI-Ger	German part of LINDSEI
LLC	London-Lund Corpus
MASC	Manually Annotated Sub-Corpus
MICASE	Michigan Corpus of Academic Spoken English
NECTE	Newcastle Electronic Corpus of Tyneside English
NOCE	NOn-native Corpus of English
NLP	Natural Language Processing

NNS	Non-native speaker
NS	Native speaker
OALD7	7th edition of the Oxford Advanced Learner's Dictionary
OANC	Open American National Corpus
Penn Guidelines	Part-of-speech tagging Guidelines for the Penn Treebank Project
PLINDSEI	Polish part of LINDSEI
POS tagging	Part-of-speech tagging
PVC	Pronunciation Variations and Coinages
SEC	Lancaster/IBM Spoken English Corpus
SLA	Second Language Acquisition
SPETs	Speech Event Types
StE	Standard English
STTS	Stuttgart-Tübingen TagSet
TBL	Transformation-based learning
TELF	Tübingen English as a Lingua Franca corpus
TL	target language
UCL	University College London
VOICE	Vienna-Oxford International Corpus of English

List of tables and figures

Preface

When Prof. Seidlhofer approached me in 2009 and asked if I wanted to investigate the issues and possibly write a PhD thesis on the part-of-speech tagging of VOICE, this came somewhat as a surprise to me. Coming from the field of general linguistics, and having not been in touch with the assignment of word classes to corpus data, being presented with the idea to write a dissertation from the field of computer linguistics seemed like a huge challenge. However, in this conversation, two points were established which contributed to my decision to take up this challenge: Firstly, it was agreed that the part-of-speech tagging enterprise would focus on the theoretical and practical challenges and issues of tagging this type of data. In addition, the tagging of ELF data should only be carried out if it was possible to display the data in a way that did justice to its complex nature. And this was by no means regarded to be a given.

However, in the course of the VOICE Project, the part-of-speech tagging of VOICE, which had not originally been the main focus of the project, reached unforeseen dimensions. In 2009, I started working full time for the VOICE project and in 2010, the team were joined by Michael Radeka, who had experience on the technical side of part-of-speech tagging. It became clearer in the course of time that the tagging of VOICE data was possible, if only through making the occasional detour and regularly stepping off the beaten tracks of established procedures (according to Widdowson 2012: 16 a feature being inherent to ELF usage itself) in order to find the most satisfactory, yet most feasible way to categorise the data. In short (and not-quite-so short over the next 200ish pages): this meant over 3 years of hard work, discussions of difficult issues and often arbitrary decisions, much sweat and some tears, the release of VOICE POS and finally, another two years later, a more detailed account of the challenges in this PhD thesis.

In the VOICE project, Michael Radeka implemented the POS taggers and was in charge of all things technical (with the continuous support of VOICE IT researcher Stefan Majewski), while myself (together with the rest of the VOICE team) focussed on the development of part-of-speech tagging guidelines suitable for an ELF corpus. Unsurprisingly, therefore, the development of these guidelines is the focus of this PhD thesis. I hope it can adequately represent the challenges faced by the tagging of VOICE, as well as demonstrate that the conceptual side of part-of-speech tagging is a little explored but much needed area of investigation, as Barbara Seidlhofer predicted some 5 years ago.

1 INTRODUCTION

This PhD thesis is concerned with the classification of spoken English as a lingua franca (ELF) data with word class categories, a process referred to as part-of-speech tagging (POS tagging). The POS-tagged version of the Vienna-Oxford International Corpus of English (VOICE) was published in 2013, and the aim of this thesis is thus to discuss in detail the practical questions and theoretical implications that emerged during the tagging process. ELF is essentially language in use in a globalised world and has been defined as "*any use of English among speakers of different first languages for whom English is the communicative medium of choice, and often the only option*" [original emphasis] (Seidlhofer 2011b: 7). This use is characterised by "hybridity and dynamism, fluidity, and flexibility […] heightened variability and a premium on mutual accommodation" (Seidlhofer 2011b: 110f.). Such spoken ELF interactions exhibit features of spoken language use which are challenging to categorise. These include not only features which characterise any spoken interaction like hesitations, re-starts, pauses, interrupted sequences and incomplete structures, but also features characteristic of the plurilingual nature of ELF such as code-mixing, the use of non-canonical forms or non-canonical form-function relationships. All of these characteristics are present in VOICE, a corpus of spoken ELF.

Arising from extensive research work within the VOICE project, this thesis has two major focus points. The first of these is to address the **practical question** of if and if yes how categories can be meaningfully and consistently attached to tokens in VOICE by the process of POS tagging, a type of annotation which increases the usability of corpora in numerous ways. However, the tagging documentation of existing spoken corpora, which are at least to some degree comparable to VOICE, is scarce, with practical issues and their solutions rarely described in detail and conceptual issues regarding the categorisation of language with parts of speech hardly addressed at all. Where documentation is available, the tagging schemes tend to gloss over, rather than display, precisely those features which are characteristic for spoken language in use. The insufficient amount of information provided in the documentation of existing corpora increases the workload for researchers attempting to tag their own spoken data and, without a doubt, limits the comparability between corpora. Since there is no precedent of automatic POS taggers being applied to spoken ELF data, and the tagging documentation of existing corpora does not address issues specific to this type of data, the tagging of VOICE was a highly challenging task. It does not only need to capture the spoken and highly interactive nature of VOICE data as accurately as possible and the variation of

language forms and functions in ELF but also be compatible with the VOICE annotation style.

The second focus of the present thesis is to enquire into the **theoretical implications** arising from the practical question of how word class categories can be assigned to a corpus of spoken ELF. Although the desirability of raising conceptual questions within tagging procedures has been suggested (Rahman & Sampson 2000: 301), this kind of enquiry is not common in the tagging literature (but see e.g. Rehbein, Schalowski & Wiese 2012a). For tagging VOICE, such an enquiry raises issues concerning the description and categorisation of spoken discourse in general and the question of what such categorisation can reveal about the variable and emergent nature of ELF in particular. As such, a 'linguistic apporach' is assumed, which is rather unusual in the generally more practically oriented tagging literature, but which will be demonstrated to provide a number of insights into more general issues underlying linguistic description of language in use.

The present thesis is structured as follows:

Chapter 2 serves as a general introduction to the main issues relevant to the POS tagging of VOICE. The first part provides a brief overview of the issues involved in the categorisation of language data, the process of part-of-speech tagging being an example of a type of categorisation. The practice of part-of-speech tagging, as well as its primary uses and challenges are then elaborated upon. It is argued that an assumed linguistic stability in such language categorisation is in conflict with the reality of language use, which is much more variable and unstable. ELF and VOICE data being an example of such language use, the second part of the introductory chapter provides a brief overview of ELF research and the relevant issues of this relatively new, but rapidly expanding field, as well as necessary background information to VOICE. The main two questions which arise from chapter 2 are firstly, how existing taggers and tagging practices operate on natural language use, as present in the ELF data in VOICE and secondly, how ELF data can be represented adequately within pre-defined linguistic categories.

In chapter 3, then, the questions arising in chapter 2 are investigated. A detailed review of available spoken POS-tagged corpora of both first language (L1) and second and foreign language (L2) data is given, as well as the POS tagging practices employed, followed by a discussion of the relevance of these for the POS tagging of VOICE data. Despite the advantages of using existing tagging standards, e.g. comparability, it is argued that existing POS tagging practices are in many ways insufficient as models for tagging VOICE: firstly,

the data in VOICE is of a much more varied and complex nature due to the highly diverse lingua-cultural backgrounds of the interlocutors and the detailed transcriptions of entirely naturally occurring speech situations. Hence, although the tagging practices of existing spoken POS-tagged corpora can serve as a starting point for some issues encountered in VOICE, such as recommendations on the size of the tagset, other issues, such as the handling of non-canonical language use, have only been discussed for very few corpora and usually rather superficially and hence do not provide sufficient guidance for the development of a tagging scheme for VOICE. Indeed, general issues concerning language variation, which are of central relevance to the tagging of VOICE, have not been previously addressed. Moreover, the hitherto adopted approaches to the tagging of spoken L2 corpora were not considered suitable for tagging VOICE due to the theoretical frameworks usually employed there. This represents L2 language in overt comparison with a native speaker model, whereas within an ELF framework, the language use of ELF speakers is represented in its own right.

In short, it was found that existing tagging procedures were not detailed enough in their descriptions, did not address a number of issues relevant to ELF data, and that the approach to the tagging of L2 corpora was considered inappropriate for VOICE data. Therefore, for tagging VOICE, new ways had to be found. The dealing with non-native data also raises questions regarding the distinction between ELF and L2 corpora as well as the data contained in these. The end of chapter 3 is concerned with how a distinction can be meaningfully drawn.

Chapter 4 of the thesis describes how the main issues and challenges encountered in tagging VOICE were handled, firstly, by establishing general criteria for tagging VOICE and secondly, by developing a tagging methodology for individual cases. Some of these issues are of a more theoretical nature, such as the feasibility of categorising VOICE data given the lack of existing tagging models appropriate for spoken ELF, and the use of external points of reference for ELF data. Other issues are more practical, such as the tagging of non-codified language use, ambiguities and variability. The chapter reports on how existing standards were applied and extended for VOICE and, which new and, in some cases, unconventional solutions were developed for the practical tagging procedure. The approach adopted in VOICE is shown to have its strengths in displaying, rather than glossing over complex aspects of spoken, plurilingual language in use. This, it is argued, is an approach that can be fruitful when applied to corpora dealing with similar kinds of language use. The chapter not only raises questions about the difficulties of categorising spoken language in use, but also shows how dealing with these difficulties points towards specific aspects of the data at hand which could easily be overlooked with existing tagging schemes.

In chapter 5 a case study on the nature of word class shifts in VOICE data is described. This study illustrates what the process of POS tagging, as well as the POS tagged data itself, can reveal about the nature of ELF and the nature of spoken language in general. During the POS tagging of VOICE, many cases were encountered whereby ELF users go beyond the commonly assumed boundaries of word class categories of Standard English (StE) by a process generally known as as 'word class change'. Word class change is understood as a change of word class category, or syntagmatic function, without alteration of the morphological form (cf. e.g. Balteiro 2007: 13, Pavesi 1998: 213). The analysis concerns those forms which resulted in a change of word class category due to a word being used with a non-codified function. Reviewing concepts commonly used to explain word class change, namely conversion and multifunctionality, it is argued that a number of presumptions inherent in these concepts cause difficulties when applied to the ELF data in VOICE. They are, hence, more appropriately termed 'word class shifts' or 'variation' for such data. In a case study, the most frequent types of such 'word class shifts' in VOICE are exemplified, analysed, and discussed with regard to the forms which are shifted, the directionality of these shifts, as well as the environment in which the shifts occur. Word class shifts in ELF interactions are shown to follow clear tendencies and this, it is argued, calls for a reconceptualization of the conventional concepts of conversion and multifunctionality with regard to ELF. This chapter shows, on the basis of the analysis of non-codified form-function relationships, that if tagging practices attempt to display, rather than disguise, initial 'problems' arising in the tagging process, this can help to gain useful insights into the nature of the data.

The thesis finishes with chapter 6, a general conclusion. It is argued that the findings of this thesis contribute to research into the description and annotation of spoken, plurilingual language use and shed light on a number of areas:

Firstly, it is shown in this thesis that VOICE can be meaningfully POS tagged to a large extent and that the usefulness of VOICE was considerably increased by the attachment of POS tags, but that there are also limits to the application of preconceived categories to naturally occurring language data. Secondly, it can be demonstrated that the linguistic perspective on tagging, as adopted for VOICE, is valuable in a number of respects, as it can connect the practical task of POS tagging with issues of the categorisation of language and other conceptual questions and is, in addition, valuable to the product of tagging itself. Three conceptual issues in particular were raised by the tagging of VOICE: Firstly, the issue of how to handle external points of reference for ELF data, as well as the lack of inclusion of typical ELF features in existing tagging standards, secondly, the distinction between ELF and learner

language data and their annotation, and thirdly, the consequences of the bias towards written language data in POS tagging. Moreover, it is argued that the linguistic approach used for the tagging of VOICE can serve as a possible strategy for other, similar types of data. While it does not necessarily represent the ideal, nor by any means the only possible solution to the annotation of such language data with parts of speech, it can certainly contribute to the discussion of the challenges and possible solutions of categorising language of a similar nature for research purposes. In addition, it ultimately adds to the discussion of the categories used for linguistic investigation in the light of actual language in use.

2 CATEGORISATION OF LANGUAGE INTO PARTS OF SPEECH AND THE ELF DATA IN VOICE

2.1 Introduction

This thesis is concerned with the part-of-speech tagging of transcribed, spoken data of English as a lingua franca interactions. This was, as will be elaborated on in detail, in many ways far from a straightforward enterprise. Before the tagging of spoken corpora in general, and specifically the tagging of VOICE data can be described, it will be necessary to provide some relevant background information. Hence, the purpose of this chapter is to review the main issues involved in the categorisation of language in general, particularly in part-of-speech tagging as well as issues regarding the categorisation with parts of speech of ELF data in this thesis (sections 2.2 and 2.3). Sections 2.4 and 2.5 provide an overview of the most important developments in the field of ELF research and introduce the corpus VOICE.

2.2 Definition and categorisation of parts of speech and the unique challenges of ELF data

2.2.1 The fuzzyness of linguistic categories

Traditionally, English has been described as having eight parts of speech (Aarts 2004b: 10). However, word classes have been defined in various ways, e.g. with regard to their notional characteristics (Aarts 2004b: 10)[1], by the way they act morphosyntactically (Aarts 2004b: 11), or with regard to semantic and pragmatic criteria (ibid.; Denison 2013b: 158). The number of word classes varies according to the way these are defined. There is little dispute that in linguistics, for doing any research, pre-defined grammatical categories are a necessity, as Aarts (2004b: 1) states. This view was also already expressed by Labov (1973: 342) who describes linguistics as a whole as a "categorizing activity". However, most linguists involved in practical tasks of language categorisation, such as part-of-speech tagging, would also agree that morphosyntactic categories cannot always be categorised easily and unambiguously. While classifications work well in many cases, the reality of actual language discourse does not always fit the conveniently pre-defined categories of linguistic descriptions. Labov (1973: 343) states that "we discover that not all linguistic material fits the categorical view: There is greater or lesser success in imposing categories upon the continuous substratum of reality". In the case of categorisation into parts of speech this might be because, as Denison (2013b: 181) phrases it, they are "abstract generalisations about the behaviour of […] thousands and thousands of words".

[1] Which is not unproblematic, see Denison (2013b: 155).

There is a long tradition of theorising about language in terms of non-fixedness, both of word class categories and of other types linguistic categorisations. In linguistics, this challenge has been accounted for with concepts such as vagueness, prototypes, gradience of categories, hybrid categories and category change. While a full account of these approaches would exceed the scope of this section, a few main strands and issues will be singled out. The section will focus on the concepts of vagueness, prototypes and gradience and some of the issues which arise from these concepts, as well as their relevance for tagging practices and the tagging of the ELF data in VOICE. Comprehensive overviews on the topic of the non-fixedness of categories are given by Aarts (2004b, 2004a and 2006), and also, a briefer account, by Sasse (2001: 495f.).

The idea that categories are not as fixed as they appear to be has been addressed as early as the 1920s, e.g. by Jespersen (Jespersen 1924, cf. Sasse 2001: 495) who mentions "hybrid categories", e.g. gerunds and by Russell (1923: 85) (see below). More generally, however, these issues were only more broadly acknowledged since the 1970s (Sasse 2001: 495). Aarts (2004b: 3f.) summarises that concepts of non-fixedness of categories gained further popularity with the rise of cognitive sciences, which combined the findings gathered in different disciplines. Categorisation was explained as a fundamental constituent of life in the cognitive sciences, and as such multi-facetted a highly complex one, suggesting that some, but not all categories had fixed boundaries and that not all constituents of a group "possess the same degree of category membership, with some members consistently being regarded as better examples of the category than other" (ibid.).

The concept of vagueness as an approach to the topic of categorisation was central to Russell's (1923: 85) article, in which the central claim is that things are not vague in themselves, as they "are what they are and there is an end of it" (ibid.). Rather, what is vague, according to Russell (ibid.), is their representation, e.g. in form of language. Russell (1923: 89) draws a distinction between the terms *accuracy* and *vagueness*, defining *accuracy* as

> one system of terms related in various ways is an accurate representation of another system of terms related in various other ways if there is a one-one relation of the terms of the one to the terms of the other, and likewise a one-one relation of the relations of the one to the relations of the other, such that, when two or more terms in the one system have a relation belonging to that system, the corresponding terms of the other system have the corresponding relation belonging to the other system.

Vagueness, on the other hand, is "when the relation of the representing system to the represented system is not one-one but one-many" (ibid.). Notably, according to Russell (1923:

91), a precise description is not necessarily truer than a vague one, in fact, quite the opposite, as a precise description is much more likely to be wrong than a vague one.

Labov (1973) approaches the topic of categorisation exploring the notions of *cup*, *bowl* and *vase* and argues that while linguistic categories work well in many cases, in some cases they cannot describe the vagueness inherent in language in use. According to Labov (1973: 353), vagueness is actually inherent in some objects themselves rather than "a property of our perception or the weakness of our instruments, [or] the abstractness of our objects". In his investigation, he found that the naming of cup-like pictures as cups, bowls and vases was highly dependent on the given context, e.g. imagining it containing coffee, food or flowers (Labov 1973: 355). As Aarts (2004b: 18) summarises, other scholars have argued along similar lines as Labov, stressing that while vague cases are indeed present in real life, the description of these cannot be fuzzy but needs to be precise (Joos 1950) and that "the challenge consists in portraying the vagueness inherent in natural language with precision" (Wierzbicka 2004: 474, cited in Aarts 2004b: 18).

Another view on the non-fixedness of categories is that of prototypes. This, as Sasse (2001: 495) summarises, was already mentioned in the 1960s by Crystal (1967: 46), as "centrality of membership" and Robins (1967: 213), who refers to "nuclear" and "marginal members". In essence, prototypicality means that some items in a group are more representative of that group than others. Aarts (2004b: 5f.) summarised Rosch's (1973) research, who investigated prototypicality with regard to the perception of colours and argued that categorisation works according to certain parameters: The first, 'Cognitive Economy' says that an item is analysed in the way it is similar to other items within a category but also different from items in other categories, resulting in a categorisation which is informative yet not very laborious to achieve, and also, that categorisation is not independent of the tasks it is intended to carry out (Aarts 2004b: 5f.). The second, 'Perceived World Structure', suggests that the world is recognised by humans as ordered and things as occurring in relation to each other (Aarts 2004b: 6). These principles, then, can explain effects such as "basic-level objects and prototype effects" (ibid.).

A third approach to word class assignment is to regard the categories themselves as gradient. A prominent example of this, as has been suggested by Ross (2004), is the 'squish' between certain categories. Ross (1972) proposes that the boundaries between verb, adjective and noun are gradient, with the following transitions: "Verb > Present participle > Perfect participle > Passive participle > Adjective > Preposition (?) > 'adjectival noun' (e.g., fun, snap) > Noun" (1972: 316). This 'squish' is demonstrated by on the basis of a number of syntactic processes, which apply most strongly to the left of the continuum, with verbs, which is also the most

productive end, and least to the right, for nouns, which is the least flexible (Ross 1972: 317). Other squishes that Ross discusses elsewhere are the 'NP Squish' (1973) or the 'Nouniness Squish' (Ross 2004). Dension (2013b: 180), argues differently to Ross (presumably in reference to Labov's 1973: 341 phrasing), that rather than being gradient or "slippery" (Denison 2013b: 180), parts of speech are "solid". This is based on the assumption that they are "morphosyntactic generalisations about the vocabulary of English, only imperfectly correlated with semantics" (ibid.). What is described as gradual, however, is the change of some words between different categories. Discussed are, for example, the diachronic shift of from noun to adjective, such as *fun* and *key* (and many other lexical items, cf. Denison 2013b: 165) which, having started out as nouns, have gradually adopted adjectival attributes. Other gradual transitions mentioned by Denison (2013b) are Adjective > Determiner (2013b: 167f.), Verb > Adjective, Adverb > Pronoun, Preposition > Adjective, Verb > Auxiliary/Modal (2013b: 168). Dension's idea of gradient change supports ideas of "genuine gradience between N and Adj" (2013b: 173) as well as the idea that some members of a class are more prototypical than others (ibid).

Another important point raised by Denison (2013b) is intra-speaker variability. Denison (2013b: 171) argues that although parts of speech are defined, their particular use cannot always be distinguished for individual speakers. Indeed, Denison (ibid.) argues that in cases which are underspecified and not ambiguous, it does not need to be distinguished by either speaker, listener or linguist. Denison (ibid.) gives the example of *It was fun* and argues that three different speakers could use the word fun as different parts of speech, i.e. varying from the more conservative use as a noun to the more recent adjectival use (ibid.), to everything in between, i.e. accepting some but not all adjectival uses of the word (Denison 2013b: 172). This point raises an important issue with regard to the difficulties of detecting gradience and fuzziness.

Various reasons are given for the mismatch between some pre-designed categorisation and what is observed in actual language data. One is, as has already been mentioned above, the fact that reality is much more complex than the simplified display of it in forms of categories. Sasse (2001: 506) on the other hand, sees the reason for gradience in grammatical categories in the complex interplay of different analytical factors regarding semantics, distinguishing criteria for categories, "lexical and grammatical categories established language-specifically on formal grounds", and the small number of categories for a large number of "semantic concepts" (ibid.) Another more encompassing, usage-based explanation for the design of parts of speech systems in languages, which is no "one-to-one mappin[g] of semantic categories

onto functional and formal categories", is given by Anward (2000). Anward (2000: 4) argues that forces which are responsible for the change of language in general are also in play when dealing with parts of speech. He proposes that part-of-speech systems come into existence because of such forces, as humans use the lexical repertoire of a language rather than designing POS-systems, and that the latter "are what 'happen', as language users strive to maximise meaning and minimise effort" (2000: 39). Languages achieve this "flexibility […] and contrast" (2000: 38) of parts of speech in different ways, which is how individual languages end up having different systems.

With regard to this thesis, a number of points can be singled out for the topic of categorisation. Firstly, it seems that the view expressed by Labov, in contrast to that of Russell, captures an essential question with regard to categories in language: Is language inherently vague or is it our descriptions of language which are imprecise? This issue is also applicable to the study of ELF: It is unclear whether the observed characteristics of ELF conversations are vague or if it is the conventional, codified descriptions of English which are not able to capture the reality of ELF in use unambiguously. While these are questions of a philosophical kind, they are nevertheless important to keep in mind when dealing with preconceived categories and actual language in use. It has, after all, been widely recognised that there is a certain mismatch between the way language works in everyday life and the way it is captured and described. This is clearly true for the analysis of communicative situations in ELF where a range of factors interplay, such as the speakers' linguistic repertoires and their cultural backgrounds, in addition to other factors at play in real-life interactions. The total number of parts of speech have been shown to vary according to how they are defined and are, as such, an artificial, albeit largely useful, system to describe language. However, they can never grasp language use in its fullness, or, as Denison (2013b: 181) phrases it:

> Like all good grammatical systems I have encountered, PoS make useful generalisations. Like all systems I've encountered so far, there's a residue which doesn't fit neatly—the bubble under the sticky plastic that you can't smooth out. Language resists being smoothed out and displayed as a neatly labelled specimen. It's a living thing (mixed metaphors and all!). That's why it's such/so fun.

A second question which emerges from the literature discussed above is whether descriptions of fuzzy boundaries should be vague or precise. Russell argues that a vague description is more likely to be true, implying that this may be preferable to a precise one, while scholars such as Wierzbicka (2004: 474) and Joos (1950) argue that a precise description of vague language data ought to be the aim, as Aarts (2004b: 18) summarises. For the assignment of parts of speech in tagging procedures, the first option, a vague description, might translate as

only displaying what can certainly be stated about the token in question, e.g. remaining vague in Russell's terms, and using e.g. vague, underspecified tags for nouns and verbs (N, V) in the VOICE Tagset (cf. section 4.3.2). The other option is the precise display of the vagueness, as the dual or the ambiguous tagging procedures in VOICE have attempted to do, in displaying more than one possible tag. In VOICE it was found that both approaches are usefully applied to different tagging issues. Finally, the point raised by Dension (2013b), about the difficulty (and lack of necessity) of identifying vagueness in some cases due to the fact that speakers may vary in the way they use language, indicates that the description of unclear boundaries might often not be solvable at all in part-of-speech tagging procedures. This is a topic which usually remains unaddressed in tagging literature, but is necessary to be considered.

2.2.2 Fuzzy categories in the POS tagging literature

On the other side of this discussion regarding the difficulties of categorising language into distinct parts of speech, is the way this is handled in linguistic works of reference and also, to a large extent, in part-of-speech tagging practices. For example, the section on parts of speech in Luraghi and Parodi's (2008: 147) *Key terms in syntax and syntactic theory* does not mention any of the uncertainty in categorising language and, instead, factually states that parts of speech can be defined as "classes that display the same morphosyntactic behaviour". Even though there is a mention of the possibility of other classification schemes, the underlying indication that this may lie in the difficulty of classification is not mentioned. Instead, the main word class categories are given with a list of morphosyntactic and semantic properties (Luraghi & Parodi 2008: 147ff.). Aarts (2004b: 3) suggests that the reason for such obliviousness is that it is less effort to assume *clear-cut* categories:

> The assumption that categorization is an unambiguous, clear-cut process turned out to be too attractive and convenient for most scholars to question its premises – as, indeed, it still proves to be for a new generation of computer scientists.

And although this quote is not explicitly related to computer linguistics or NLP, it is certainly true for the majority of the documentation of tagging procedures in the literature. The difficulty of assigning clear-cut categories to corpus data is rarely mentioned, let alone questioned, and the approach is usually a highly operational one (Güngör 2010: 206). Rahman and Sampson (2000: 301) criticise this and argue that theoretical questions in tagging ought to be given "a high priority", even though it is rarely practiced. Leech (2004)[2] is one of few scholars to explicitly acknowledge the difficulty of categorisation. In *Developing linguistic corpora: a guide to good practice*, Leech (ibid.) states that, "[t]here is no absolute 'God's

[2] Online version used, no page numbers given.

truth' view of language or 'gold standard' annotation against which the decision to call word *x* as noun and word *y* a verb can be measured". Leech (ibid.) argues that categorisation cannot be entirely uniform as there is no sufficient agreement on the definition of each word class, and even when the categories used in POS tagging refer to a "'consensual' set", borderline cases where agreement is hard to find, are likely to exist. However, in the literature in general, difficulties in part-of-speech categorisation tend to be are attributed to the complexity inherent in spoken language, as tagging schemes only work for "'textbook' example sentence[s]" (Rahman & Sampson 2000: 295) which are argued not to be easily transferable "to the messy constructions that are typical of real-life data" (ibid.). Alternatively, the differences are sometimes attributed or to 'erroneous' L2 output (cf., e.g. Jendryczka-Wierszycka 2009), as discussed later in this thesis.

Moreover, the vast majority of tagging literature does not mention or deal with issues such as vagueness and gradience, issues which form an integral part of what has been acknowledged to exist in the wider literature on the categorisation of language. There is rarely any in-depth discussion about how to present vague categories and other challenges which arise due to the difficulty of categorising language and hence, there is only a selection of suggestions regarding how to display such issues in tagging procedures at all. Denison (2007)[3], who views the preoccupation with tagging practices, "at the same time a boon, an obstacle and an object of interest in its own right" investigates how a number of vague categories have been displayed in tagged corpora. In his case study of 5 examples "which fall on or close to word class boundaries", Denison (2007), finds that these are not straightforward to deal with for taggers, due to the differences in tagsets and in the procedures specified for dealing with vagueness (ibid.). Denison (ibid.) also sees the non-uniform dealing with such cases as being due to the application of text-book descriptions of grammar to real-life communication and the view "that every word in every sentence belongs to exactly one category – no more or less – selected from a small number of parts of speech" (ibid.). However, according to Denison (ibid.), this is not to blame on the 'messiness' of spoken language use but rather on a "practical oversimplification of reality" which is "unworkable", especially for vague cases which are "a common part of everyday colloquial English".

The consideration of issues which pose difficulties for language categorisation within tagging procedures can also shed light on issues of general interest in linguistics. Dension (2007) suggests that part-of-speech tagging could "actually reveal language change […], rather than merely playing catch-up after the event", if vague cases were described systematically and

[3] Published online, no page numbers given.

taggers were trained to display these. However, for Denison (2013b: 175), deliberate "language play" or non-native language use do not seem to play a role in such changes. As the following quote (regarding internet data) indicates, Denison (ibid.) regards word creations by non-native users of a language as 'errors', stating that especially in the case of online data, problems are "the difficulty of knowing whether you're dealing with native speakers, whether something odd is a simple mistake and therefore whether one is observing the possible start of a change or just an error". However, he demonstrates that there are cases where the language "usage suggests that the [speakers] know they are pushing the boundaries of what is possible" (Denison 2013b: 175) and thus show that they are "quite aware of change" (ibid.). The pushing of language boundaries has also been shown to be an integral part of ELF communication (cf. Seidlhofer 2011b), as will be demonstrated later in this thesis (chapter 5). It is suggested therein that even though the situation is identical to that described by Denison as being irrelevant for language change[4], on the grounds that the examples presented are largely produced by non-native users of English, and hence, it is argued, it is often difficult to determine whether the novel use was intentional or a slip of the tongue, these 'shifts' or 'variations', as I will call them, show a certain systematicity and can reveal something about the way speakers exploit linguistic resources in online conversation.

A second issue of general relevance raised by the difficulty of categorising language with parts of speech, in Russell's use of the term of categorisation, is that it might serve as a pinpointer towards issues which are not well described with our representations of language. Whether or not these issues can be described in any better way or whether they cannot be grasped by our representation of language, remains, however, unclear. As Russell (1923: 87) argues, "the knowledge that we can obtain through our sensations is not as fine-grained as the stimuli to those sensations".

2.2.3 Fuzzy categories and the study of ELF

What does all of this mean for the study of ELF and the annotation of ELF data with parts of speech? In a way, ELF takes the critical discussion of fuzzy notions of the categorisation of language to a whole new level. Within the context of globalisation, Seidlhofer (2011b: 91f.) argues that established concepts such as community, and therefore also what communicative competence and language varieties are, are no longer what they were when used in reference to native speakers of a language, and therefore, "they represent a dynamic process as a fixed state of affairs" (2011b: 95). Seidlhofer (ibid.) continues to argue that this is problematic

[4] In fact, the concept of 'language change' might be more suited for relatively stable speech communities, and not for 'situationally contingent' settings as often encountered for ELF.

because “language is, of course, not fixed but continually in flux, always variable” and ELF “actually exemplifies this process” (2011b: 92).

In his 2012 article, Widdowson discusses several constructs used in language descriptions which are necessary to make sense of language reality, although, at the same time “convenient fictions” (2012: 7f., citing Seidlhofer 2011b: 70). For Widdowson (2012: 15), questions which need to be considered regarding such constructs are, “what are its limits? How convenient a fiction is it?”. As examples for such artificial constructs, Widdowson (2012: 8) discusses Chomsky’s “perfect knowledge of a language”, Labov’s “distinct speech community” (2012: 9ff.), the notion of a standard language (2012: 9) and its codifications in works of reference (2012: 13), the language captured by native speaker intuitions, as well as the transcripts in corpora, the notion of the native speaker (2012: 10ff.), native speaker competence (2012: 23) and native speaker varieties (2012: 9)[5]. He argues that neither the Chomskyian sense of ‘competence’, nor the ‘performance’ presumably collected by corpora, can capture actual language use. Despite often being “represented as the ideal” (Widdowson 2012: 14), the language output recorded in corpora is nevertheless only something that a particular group of speakers have produced (ibid.). However, it does not capture “why they produced [it] and to what pragmatic ends and purposes” (ibid.), as studies based on corpora “dea[l] with performance, but only with the form that performance takes, and abstracting that form from its natural communicative function in use makes it into an analytic construct, another kind of fiction” (2012: 15). The paper concludes with the proposal that research into ELF, where speakers have been shown to step off the “beaten track” (2012: 16) as they often do not conform to the various abstract pieces of ‘convenient fiction’, and nevertheless communicate successfully, sheds light onto these abstract notions of categorisation and has the potential to cause linguistics to reconsider them (2012: 24). ELF research, Widdowson (2012: 24) argues,

> is […], of particular significance in that it prompts a reappraisal of established, taken for granted ways of thinking about language, especially English. I have argued that, convenient though these ways may be for some purposes and for some manifestations of the language, they are an encumbrance when it comes to understanding how English is used as a lingua franca.

What Widdowson describes here as ELF being a test site for the validity of ways which have been used to describe language, is something which becomes very apparent when carrying out

[5] I would also argue that there needs to be caution about not committing the fallacy of generalising ELF users as abstract constructs, something which is not mentioned in Widdowson (2012) and indeed, much literature on ELF. ELF users vary greatly in terms of their linguistic repertoires, intelligibility to their interlocutors, etc. Just like the notion of the native speaker, or native language, it is necessary and useful to abstract the notion of an ELF speaker, but it is also be necessary to be aware just how convenient this construct is.

part-of-speech categorisation of ELF data. This is, of course, a very practical task which essentially seeks to describe the reality of ELF usage with pre-defined "convenient fictions" (Widdowson 2012: 7f., citing Seidlhofer 2011b: 70). And so, it is of little surprise that in the tagging of VOICE, numerous examples were found where categorisation was difficult, e.g. the case of tokens ending in *–ing* or many pronouns, (*none,* etc.). On the one hand, this might be due to the fact that certain categories in English are not clearly distinct, as described earlier, or at least that not all tokens which are attributed to them fit neatly into the existing categories. On the other hand, ELF itself and the way ELF users exploit the potential inherent in the language causes challenges in trying to apply conventional categories to the data in VOICE.

However, one should also not forget that these cases are in the minority compared to all the other tokens which were easily categorised. So with regard to VOICE data, there is certainly fuzziness for some items, however, the pre-defined categories do seem to work for a large number of cases which seems to suggest either that conventional categorisation systems are in line with the data, and/or that such a pre-defined categorisation is to some degree inherent in language. This resonates with Labov's (1973: 343ff.) idea of not one uniform type of categorisation for language, but the existence of different types of categories. He suggests three types of categories, namely fixed ones, internally variable ones and flexible ones. These are summarised in Aarts (2004b: 5) as follows:

> The first type conforms to the categorical view: categories of this type are bounded and non-graded, with distinct and invariable outer-limits, characterized by a fixed set of necessary attributes and simple yes-or-no membership. The second type of category is bounded, but at the same time graded. These categories have strict boundaries, but allow for within-category gradience, with some members being better examples of the category than others. The third type of category is both unbounded and graded. Categories of this type do not have boundaries; rather, they fade off into each other, with the more peripheral members having more in common with members of other (contrasting) categories.

So while Labov's different types of categories may provide an explanation for the fact that some items are more difficult to label than others, the inconsistencies and difficulties in the tagging of VOICE may also be accounted for in other ways. In fact, tagging challenges might point towards more general issues of language use. For example, according to Widdowson (2012: 23), ELF users may apply intuitively what is most communicatively effective, while omitting redundancies. Hence, I would argue that challenges in the tagging process of ELF data might well point to issues relevant for understanding the process of learning and using language. They might also be an indication for aspects in the English language which have a potential for exploitation and hence where categories are particularly fuzzy, as is the case with

the numerous tokens ending in *–ing*, which are employed creatively and not according to standard language codifications as verbs, nouns and adjectives and other word classes (see section 4.3.3.1.).

In a nutshell, what can be said is that categorisation is crucial to describe language. At the same time, however, it has its limitations, and these need to be recognised, not only theoretically, but also in practical representations of language categorisation, such as POS tagging. The study of ELF is, on the one hand, challenging, within these preconceived categories. On the other hand, it challenges precisely these as in many cases it transcends the boundaries established there. In order to fully understand the nature of the data which this thesis is concerned with, namely English as a lingua franca interactions within the VOICE corpus, ELF and VOICE are described in more detail in the last two sections of chapter 2. Before proceeding to this, however, the operational principles of POS tagging will be addressed in the next section, as well as the main principles, challenges and limitations of this annotation method.

2.3 Part-of-Speech tagging: principles, challenges and limitations

Part-of-speech tagging, as all corpus annotation, is an attempt to describe natural language with pre-formed categories. In this section, a concise overview of the definition, uses and methods of part-of-speech tagging for corpora is given. However, as the aim of this thesis is to demonstrate how spoken language has been dealt with within POS annotation schemes and how to devise a tagging scheme for spoken, plurilingual language data in particular, the technical aspects of part-of-speech annotation are not a primary focus and hence not described in detail herein. A discussion of the technical aspects of tagging VOICE is topic of Radeka (in prep.). Two other aspects, which are often related to the annotation of part-of-speech tagging but will not be addressed in this thesis are the theoretical guidelines and the technical procedure of lemmatising VOICE. An overview of the documentation of the lemmatization can be found in the VOICE Tagging and Lemmatization Manual (VOICE Project 2014a).

McEnery and Hardie (2012: 54) argue that little literature exists concerning corpus annotation, with the exception of papers on the annotation of specific corpora, and a very limited number of edited volumes on POS tagging (e.g. van Halteren 1999b) and corpus annotation in general (e.g. Garside, Leech and McEnery 1997). Indeed, literature on general procedures of part-of-speech tagging, as well as descriptions of tagging decisions, are rare, as will be demonstrated in later chapters of this thesis. However, some recent overviews on part-of-speech tagging are provided by Güngör (2010: 205ff.) and by Schmid (2008). An external

perspective on part-of-speech tagging from the field of mathematics can be found in Martinez (2012: 107ff.). Also, a concise and accessible summary is given in Ramsay (2010: 92ff.).

Atwell (2008: 501) defines part-of-speech tagging as "enriching a corpus by adding a part-of-speech tag to each word". However, tagsets for English usually consist of more than the 8 traditional parts of speech or word class categories (Lüdeling & Kytö 2008: 503), as additional aspects such as inflectional information or lexico-semantic properties tend to also be included (Voutilainen 1999b: 5), as do other aspects of language use, such as pragmatic markers. This is presumably why the annotation has also been referred to in broader terms as, for example, syntactic word class tagging (Halteren 1999b), grammatical tagging (Garside & Smith 1997) or morphosyntactic annotation (e.g. Rögnvaldsson 2005, Sagae et al. 2010). Despite these different terms, for reasons of consistency, the term 'part-of-speech tagging' (or 'POS tagging') is used in this thesis whenever referring to this annotation method.

Part-of-speech tagging is only one type of annotation which can be applied to corpora; there are many others e.g. semantic, prosodic (Leech 2007: 11). However, it is the most widely used form of tagging and often serves as a prerequisite for many other annotations, such as parsing or semantic tagging (Leech 2007: 11, McEnery and Hardie 2012: 34 and Schmid 2008: 541).

Numerous advantages have been attributed to corpora which have been annotated with parts of speech. [6] Apart from generally enhancing the usability of a corpus, it has been said to provide "'added value'" (Leech 2004) and, according to Ramsay (2010: 92), is a prerequisite for doing "any deeper analysis". Atwell (2008: 505) writes that part-of-speech tagging can have the purpose of "enrich[ing] the text with linguistic analyses to maximise the potential for corpus re-use in a wide range of applications". Some commonly mentioned applications are the preparation for using a corpus for the creation of dictionaries and grammars (Leech 2004, Leech & Smith 1999: 33), the creation of frequency lists with the added POS-information (which also means that the words can be added to different lists depending on their word classes) (Leech 2004), applications for Second Language Acquisition (SLA) research (Leech & Smith 1999: 33), as well as "many NLP applications" (Atwell 2008: 506), such as automatic "speech synthesis", in cases where the POS tag makes a difference to pronunciation (Leech 2004) or for the development of NLP software, such as new tagsets or taggers that can be applied to other corpora with similar data (Leech & Smith 1999: 33). Part-of-speech

[6] The annotation of corpora, such as part-of-speech tagging is, however, not entirely uncontroversial and has been seen as a disadvantage by some (e.g. Sinclair 2004). It has sometimes been argued that such annotation is not desirable as is might be error-prone and also mirroring "predilections" of the tagger Leech (2004).

tagging also renders searches for particular aspects faster and more accurate (Atwell 2008: 505). Moreover, an automatically tagged corpus (as opposed to an annotation which is done solely for one researcher's good) can be widely distributed and used (Leech 2004) and the uses of a corpus usually exceed the compilers' initial intentions, as "the annotations themselves spark off a whole new range of uses which would not had been practicable unless the corpus had been annotated" (ibid.). Atwell (2008: 505) names as such unpredicted uses of tagged text the already above mentioned lexicography and natural language processing.

The first corpora were tagged with hand-written rules by linguists (Voutilainen 1999a: 10), however, this method was both cost-intensive as it had to be carried out by expert personnel, as well as prone to errors (Ramsay 2010: 93). Hence, automatic taggers were developed, although a large amount of manual labour is still necessary today in order to achieve high tagging accuracy. Typically, automatic taggers work by, firstly, tokenising the text into individual parts, as a second step an "ambiguity look up", with a lexicon and a "guesser" for items not included there (Voutilainen 2003: 221), and thirdly, the "disambiguation" of tokens, which is done according to information about the individual tokens, as well as longer token and POS-sequences (Voutilainen 2003: 222). In essence, tagging is either done by rule-based or statistical (also called probabilistic) methods. Rule-based methods (cf. e.g. Brill 1992, Brill 1995, Voutilainen 1995), such as Transformation-based learning (TBL), apply rules to improve the tagging of text, e.g. via a comparison of a GOLD standard of a text (Ramsay 2010: 94). Rule-based taggers were used in the 1960s and 1970s (Schmid 2008: 541) and went out of fashion, but have been used again more recently, as Garside and Smith (1997: 102) report. Statistical methods, in a nutshell, estimate probabilities, e.g. transition probabilities, and choose the most likely tag for each token's syntactic environment (ibid.). The most common models are Hidden Markov Models (HMM) (e.g. Brants 2000, Church 1988, Cutting et al. 1992), the Maximum Entropy Approach (e.g. Ratnaparkhi 1996), Support Vector Machines (e.g. Giménez & Marquez 2003, Nakagawa, Taku & Yuji 2002), different neural network taggers (e.g. Roth & Zelenko 1998), or decision tree taggers (e.g. Schmid 1994)[7]. However, in practice, these probabilities will still involve errors, as "in almost any practical application these conditions are violated" (Ramsay 2010: 93). Hence, for both rule-based and statistical methods, a certain amount of manual annotation and correction is still necessary. For rule-based methods, the automatic tagging output has to be compared against an (often manually tagged) gold standard in order to create rules. For the estimation of

[7] A detailed overview of different rule-based and statistical taggers is given in Güngör (2010: 209ff.) and some references for a range of different taggers are provided by Schmid (2008: 541).

probabilities for tag sequences, a prerequisite is an already tagged corpus, which means that again, the initial work has to be carried out manually (Ramsay 2010: 93).

Methods of rule-based and statistical taggers are often combined. An advantage of TBL is its potential to be used in combination with other methods and taggers, e.g. as improvement after having used a statistical tagger (Ramsay 2010: 94). Garside and Smith (1997: 102) see the future in a "hybrid system", which "captures the best of both techniques", i.e. rule-based and statistical. They describe CLAWS4 as an example for such a "hybrid tagger, involving both probabilistic and rule-based elements" (1997: 103). Another possibility to combine the advantages of different types of taggers are 'stacked TBL' (cf. e.g. Wu, Ngai & Carpuat 2003) or a procedure referred to as 'voting' (Brill & Wu 1998).

The performance of a tagger is usually measured by correctness (or accuracy), precision and recall (for a more detailed account of these terms see van Halteren 1999a: 81ff.). However, as van Halteren (1999a: 84) argues "there is a marked decrease in the value of the performance measures" and what is more crucial is uniform and coherent tagging, i.e. "that the same word in the same context receives the same tag" (1999a: 85). Unfortunately, such coherence is more difficult to quantify. For this reason it is difficult to judge the consistency of taggers, except to say that rule-based taggers are more consistent than probabilistic ones (1999a: 85). Even though tagging is generally believed to be fairly accurate, Manning (2011: 171) notes even very recently that there is still much room for improvement in part-of-speech tagging as the tagging accuracies of just over 97% do not refer to whole sentences but individual tokens. For full sentences, accuracies are much lower with at about 55-57%.[8]

Although part-of-speech tagging has been around for a number of decades, a number of issues remain. The two major challenges in part-of-speech tagging which are commonly listed in the literature are the disambiguation of ambiguous forms (Ramsay 2010: 92), i.e. "distinguishing words which have the same spelling, but different meanings or pronunciation" (Leech 2004), as well as the tagging of word forms unknown to the tagger, i.e. "unknown words and unknown readings of known words" (Schmid 2008: 547). In VOICE, such *unknown words* were e.g. pronunciation variations and coinages. *Unknown readings* occurred for e.g. non-capitalised words (as in VOICE capitals are only used for signalling emphasis) such as 'i', which was capitalised in the tagger's training corpus. In some cases, ambiguities can be disambiguated by the surrounding environment (Ramsay 2010: 92). Ambiguities also arise as POS tags and tokens cannot always be matched one-to-one, as sometimes there is more than

[8] Note that tagging accuracy does not refer to how accurately the language reality of the data is represented in tags, which might be referred to as 'validity'.

one tag suitable for a token. Ambiguities in tag assignment and strategies to deal with this were also a major issue in the tagging of VOICE and are discussed in more detail in section 4.3.2.

Another issue mentioned are the challenges which occur when the data to be tagged is new, e.g. when taggers are applied to different types of data than they were trained on, in which case a decrease in the taggers' performance can be expected (Manning 2011: 171). In such cases, a large work-load goes into creating training material for a tagger which, as Schmid (2008: 541) writes, "[f]or optimal performance" means "up to a million words" (ibid.). This is an issue especially for small corpora of new data, such as VOICE, which only consists of little over a million new words and where hence, either a much smaller amount of training data of the same corpus has to suffice, or training data from other (often only remotely similar) corpora has to be used.

As a final issue, it seems important to mention that POS tagging will always contain a certain amount of errors and inconsistencies and will never be entirely accurate if only carried out automatically as "because of the complex and ambiguous nature of language, even a relatively simple annotation such as POS tagging can only be done automatically with up to 95% to 98% accuracy" (Leech 2004), and hence, manual annotation is a requirement, indeed "often on a large scale" (ibid.). In *The Routledge linguistics encyclopaedia* (Ramsay 2010: 92), this is phrased this as follows:

> [L]anguage engeneering techniques accept from the outset that their results are likely to be wrong. Given that the kind of text being analysed is certain to contain production errors, as well as often just being too complicated for more linguistic methods to work within a reasonable time, the potential for making mistakes is built into the task from the start. This moves the goalposts: what we want now are systems that are fast and reasonably reliable. Accepting the inevitability of error opens the door to approaches to linguistic analysis that would not have even have been thinkable in a more classical setting.

What I would like to draw attention to in the above quote is not only the stated point that while an annotation like POS tagging enhances the possibilities to work with the data, it is bound to have a certain error rate, which sometimes, in our experience within the VOICE Project, comes as somewhat of a surprise to external corpus users. Moreover, it is stated that these tagging errors occur due to "production errors", as well as "complicated" data. Both of these features describe, however (though phrased somewhat derogatively), the nature of spoken discourse. The cause for tagging errors can therefore be explained by attempting to describe such naturally occurring language in terms of pre-set categories, as discussed in the previous section. In the above quote it is argued that this is due to automatic methods, which

is the reason for Leech's (2004) point that manual annotation cannot be spared. However, in the tagging of VOICE we experienced that, even though possible in a number of cases, tagging errors cannot always be solved easily by manual annotation either. Instead, the difficulties which remain in the manual annotation of cases which were assigned wrongly by the tagger in the first place pointed towards exactly such problematic issues of categorisation. In these cases, we found that for VOICE the best solution is the development of a workable tagging scheme which brings these categorisation difficulties to the surface, but does not resolve them as this would, de facto, be impossible in many cases.

For a good quality of corpus annotation, such as part-of speech tagging, a number of recommendations have been made. One aspect is documentation. The part-of-speech tagging of a corpus should be well-described with regards to time, place and manner of the annotations, the theoretical framework, as well as people involved, the tagset and tag categorisation used and tagging accuracy achieved (Leech 2004). Additionally, the decisions taken on the choice of tags (in a tagset), as well as the "annotation practices", should be covered in an "annotation manual" (Leech 2004). Such a tagset can be created either by referring to standards, e.g. the EAGLES guidelines, or by the extension and changing of a tagset which has already been used on a different corpus (2008: 511). Other practical advice Leech (2004) gives on the annotation of corpora include the use of annotations which should be able to be isolated from the texts. Moreover, decisions taken should take into account a linguistically "'consensual' set of categories" (Leech 2004), though full agreement might never be achieved on some borderline cases. As such, Leech suggests the use of a set of pre-defined categories. This is in line with Atwell (2008: 507), who argues that a completely "theory-neutral" annotation is unrealistic in any case, as is demonstrated by the fact that there exist a large number of different tagsets (ibid.). Moreover, Atwell (2008: 510) recommends making the tagset fit the purpose it is to be used for, as "[m]any specific uses of corpora do not need delicate, detailed tag sets." A final recommendation of Leech (2004) is that "de facto standards" should be taken into account, i.e. referring to what other people have done successfully in the past and only to differ from it when this is necessary.

In general, for VOICE, POS tagging was carried out in order to firstly enhance the usability of the corpus by this standard annotation procedure and to facilitate other annotation forms, such as parsing (although there were never any concrete plants to carry out parsing within the frame of the project). However, the second, and perhaps more prevalent motivation for tagging VOICE was to see if and how the ELF data in VOICE could be categorised with codified categories for ENL. In particular, we were interested in the areas where problems

occurred as this meant that these pointed towards potentially relevant characteristics of the data. The very fact that VOICE is rather difficult to categorise demonstrates the fluidity of the data, which cannot easily be captured by conventional, rather restrictive descriptions of language and categorisations, as has been discussed in the preceding section.

2.4 English as a lingua franca explained in a nutshell[9]

It is widely acknowledged that the English language today is used in a multitude of situations all over the globe, with a significant role in various areas such as business and higher education, but also language politics or the media. Research into ELF has focussed on all of these discourse situations, some more intensively and some less (Jenkins, Cogo & Dewey 2011: 297). Rather than calling this 'English', which conveys the impression of being a single language variety, sociolinguistics acknowledges the variety of forms and functions English serves today by referring to it as 'Englishes' (e.g. Jenkins 2009, McArthur 1998). It has been recognised that various uses of English go far beyond those in inner and outer circle countries (Kachru 1985), where English has official status, to the expanding circle, where it has no such official status, without constituting distinct varieties. In expanding circle settings, English is often a communication tool between speakers who do not share a native language and therefore a lingua franca for all interlocutors. I would like to continue this section on ELF by letting two people give their definition of this term. The first is a quote by Barbara Seidlhofer, renowned ELF scholar and founder of the corpus VOICE, which contains the data this thesis concerns. According to Seidlhofer (2011b: 7), English as a lingua franca can be thought of as

> *any use of English among speakers of different first languages for whom English is the communicative medium of choice, and often the only option* [original emphasis].

The second definition is given by speaker S8 in a workshop discussion recorded in VOICE:

> so i think the lingua franca is actually when you put some of your own things (.) INTO another language. (.) and it's a TOOL that you use […] with (.) <5> (a) differ</5>ent person (.) <soft> who has not the same (.) mother tongue as you.</soft> (EDwsd303: 350-354, S8, L1= mac-MK)[10]

A few key issues in ELF research can be drawn from these two quotes. The first is that ELF is a way of communicating across languages, as is mentioned by both Seidlhofer and S8, and a very common one because there is frequently no other language to fulfil this role, as Seidlhofer states. As such, ELF is widely understood to include non-native, as well as native speakers of English (Jenkins, Cogo & Dewey 2011: 283, Seidlhofer 2011b: 7), a definition which differs from that which has been applied to other lingua francas (Seidlhofer 2011b: 7).

[9] Parts of this chapter have been published in Osimk-Teasdale (2013).
[10] For explanations on the format of data examples from VOICE see footnote 25.

Secondly, both definitions make clear that the main focus in research on ELF is primarily on issues to do with language **usage** in all kinds of contexts for real-life purposes, rather than on language **learning**. Hence, ELF is also often clearly differentiated from English as a foreign language and the speakers of ELF from learners of English (e.g. Jenkins, Cogo & Dewey 2011), a differentiation we will come back to in section 3.1.5.

A third, very important, point is raised by S8 in VOICE, namely that the language of ELF will necessarily contain *some of your own things*. This aspect, inherent in ELF usage, has been demonstrated by a large body of research which provides evidence that ELF speakers adapt English to their own needs and do not necessarily use it in the same way that is prescribed in codifications of English, or in the way that (idealised) native speakers of the language would. So while the language competence the ELF speakers have acquired will play a role in that they "will naturally make use of actualizations in English they have been exposed to, and instructed in" (Seidlhofer 2011b: 120), the speakers will also adapt their language output as needed and will use forms and functions which they have not learned and which are often non-codified. Seidlhofer (ibid.) argues that the speakers will exploit, what she calls "abstract virtual rules", and that in order "to regulate their performance in real time, actualizations that often go well beyond what they have been presented with as the English that they should emulate". Moreover, ELF users have been shown to employ their personal language repertoires in interactions, rendering ELF "a multilingual mode" (Hülmbauer & Seidlhofer 2013: 387). It has been shown that the forms ELF speakers use are functionally effective, though regularly non-conforming to codified English language use. In fact, Seidlhofer (2011b: 147) argues that this is the way it must be, as language in ELF interactions needs to serve other "needs and purposes" (2011b: 148) and maintains:

> The formal features of ELF, like those of any natural language, are motivated by the functions they are required to serve and in this respect they are not abnormal at all but on the contrary conform to certain basic communicative principles which are incompatible with conformity to native-speaker norms. (ibid.)

Another aspect which follows with regard to these *own things* that S8 mentions in his definition is that the forms and functions in ELF are dependent on the situation in which they occur. Such communicative settings might sometimes be more permanent and stable (i.e. in communities of practice, cf. Ehrenreich 2009) than other times (e.g. one-off encounters as many speech events in VOICE are) and the linguistic features may not necessarily be established beyond this situation (aspects of ELF communication referred to as 'situationality factor' by Hülmbauer (2009: 327), and 'situational contingency', by myself (see chapter 5). The use of ELF has hence been differentiated from concepts applied to other uses of English,

e.g. World Englishes. Concepts such as "variety, community, and competence" (Seidlhofer 2011b: 95) do not adequately capture language use as it has been described for ELF. Moreover, ELF has been described as closely connected to globalisation, being both its "consequence and […] principal language medium" (Jenkins, Cogo & Dewey 2011: 303) and it calls for a re-consideration of established concepts which function differently in a globalised world. Such a reconsideration is, for example, necessary with regard to the concept of communicative competence (Seidlhofer 2011b: 92), as "[w]hat it means to be communicatively competent in English can no longer be described with reference to norms of linguistic knowledge and behaviour that are relevant only to particular native-speaker communities" (ibid.).

What follows from this paragraph is that ELF users meet the challenges of global communication by adapting their communicative resources to the individual interlocutors, as well as the communicative situation, and that this is often carried out under a certain online pressure. It does therefore not come as a surprise that ELF necessarily contains a certain degree of variability and flexibility, and that ELF is not easily captured with conventional concepts and terms.

Historically, as Jenkins, Cogo and Dewey (2011: 282) summarise, the phenomenon of ELF was first mentioned in the 1980s by Knapp (1985, 1987) and Hüllen (1982), and up to the 1990s, only very few publications on the topic had been published (2011: 282). At the turn of the century, the number of publications increased significantly. Worth mentioning are, first and foremost, the ground-breaking monograph *The phonology of English as an International Language* by Jennifer Jenkins (2000) and Barbara Seidlhofer's (2001) "Closing a conceptual gap: The case for a description of English as a lingua franca", which called for a thorough description of the long recognised phenomenon of ELF. Given the crucial role of English as a global means of communication in the expanding circle, it is not at all surprising that the nature of this language use in ELF has been of increasing interest to researchers, and that since around the turn of the century, ELF has become a thriving research field. This is not only demonstrated by the increasing amount of research, including numerous MA and PhD theses on the topic, but also the annual conferences on English as a lingua franca since 2008, the launch of the discipline-specific *Journal of English as a Lingua Franca* in 2012, and the inclusion of the topic into a number of different EU projects, e.g. DYLAN and LINEE (Haser et al. 2012: 131).[11] The speed at which ELF as a research field has developed in recent years

[11] Excellent overviews of the recent developments of ELF as a research field are provided in thematic chapters written by Barbara Seidlhofer and her research team in *The Years Work in English studies* in Haser et al. (2012:

renders it unfeasible to capture or keep up with everything that has been published within this section. This section thus serves to give a concise overview of the main issues and developments of ELF research.

Research into ELF has primarily been concerned with spoken language, often highly interactive, and has focused on different domains of usage, largely in education (e.g. Björkman 2009, Breiteneder 2009, Ranta 2009, Smit 2009, Smit 2010a), and business (e.g. Ehrenreich 2009, Pitzl 2005, Pitzl 2010, Pullin Stark 2009). More recently, the use of ELF in other domains has also been investigated, such as in written business communication (e.g. Bjørge 2007, Böttger, House & Stachowicz 2013, Clayson-Knollmayr 2010, Kankaanranta 2006) and academic writing (e.g. Ingvarsdóttir & Arnbjörnsdóttir 2013), private conversations, e.g. couples' talk (Klötzl 2014, Pietikäinen 2014), as well as in asylum procedures (Dorn et al. 2014, Guido 2012) and language politics (e.g. Lo Bianco 2014).

It is noteworthy, though unsurprising, that some domain-specific differences have been shown which are likely to affect the way ELF is used, e.g. with regard to attitudes towards ELF in business in contrast to its use in educational settings (Ehrenreich 2009: 128). Moreover, ELF has been investigated in relation to its role in different parts of the world, e.g. Europe (e.g. Breitender 2009, Seidlhofer 2011a, Seidlhofer, Breiteneder & Pitzl 2006) but also Asia (e.g. Baker 2011, Kirkpatrick 2010, Kaur 2009, Murata & Jenkins 2009). In general, work on ELF has often been based on individual data collections of different sizes, including a small number of longitudinal studies (e.g. Smit 2010b). Through the compilation of a number of ELF corpora in recent years, e.g. the free-to-use VOICE (see also section 2.5), the Corpus of English as a Lingua Franca in Academic Settings (ELFA), the Tübingen English as a Lingua Franca corpus (TELF), and the Asian Corpus of English (ACE), corpus-based research has been enabled, which has doubtlessly contributed substantially to the description of ELF.

In their state-of-the-art article, Jenkins, Cogo and Dewey (2011: 286ff.) provide an overview of the different linguistic areas which have been investigated with regard to ELF and mention in particular phonology, lexis/lexicogrammar and pragmatics. Research into these three main areas showed, in essence, that in ELF communication, speakers purposefully adapt their language use (e.g. features of pronunciation) to the respective interaction (Jenkins, Cogo & Dewey 2011: 287). Hence, this shed light on pronunciation research from a different angle, demonstrating that factors other than the speakers' first language can play a role in ELF

119ff.), Haser et al. (2013: 124ff.) and Haser et al. (2014: 120ff.). The state-of-the-art publication by Jenkins, Cogo and Dewey (2011) provides a recent overview taking into account a longer period of time. An in-depth account of the key issues and implications of ELF can be found in Seidlhofer (2011b).

interactions (just as in any interaction among L1 speakers). Moreover, it was demonstrated that certain phonological or lexicogrammatical features which are, often laboriously, addressed in EFL teaching as being crucial, do not actually contribute to increased intelligibility (e.g. Jenkins 2000, Osimk 2010, Walker 2010) or to effectiveness of communication (e.g. Hülmbauer 2009).

Research into lexicogrammar has revealed that ELF users make use of a number of morphological and syntactic innovations (as summarised by Jenkins, Cogo & Dewey 2011: 291). It has been argued that the pre-requisites which are in place for ELF communication create particularly favourable conditions for such innovative language use, and hence occur more frequently in ELF than in native language use (Jenkins, Cogo & Dewey 2011: 291f.).

On the pragmatic side, Jenkins, Cogo & Dewey (2011: 293f.) summarise that the focus has been on the one hand, the topic of non-understanding (cf. Pitzl 2010), and on the other the strategies of reaching a shared understanding. This has shed light on a number of communication strategies ELF speakers have been found to use, such as "clarification, self-repair and repetition […] and paraphrasing" (Jenkins, Cogo & Dewey 2011: 293), as shown by e.g. Cogo (2009), Cogo & Dewey (2006), Kaur (2009), Lichtkoppler (2007), Mauranen (2006) and Watterson (2008). Another strategy listed by Jenkins, Cogo and Dewey (2011: 294) is the speakers' successful deployment of their plurilingual language repertoires, as demonstrated by, e.g., Klimpfinger (2009) or Hülmbauer (2013). A newer field within pragmatics is that of discourse markers, which was "relatively ignored" (Jenkins, Cogo & Dewey 2011: 294) at the time of their state-of-the-art article, but has since been explored more thoroughly with studies based on corpus data from VOICE (see section 2.5). These studies focus particularly on the functions of certain discourse markers in ELF and demonstrate that in many cases, the uses differ from those the markers fulfil in ENL. In general, an important development in ELF research was the change from the investigation of the specific characteristics of the forms of ELF in earlier papers, to functions of these forms, as well as the processes which cause these forms to occur. Jenkins, Cogo and Dewey (2011: 287) write,

> the focus of research […] shifted from an orientation to features and the ultimate aim of some kind of codification (an aim which, nevertheless, has not been dismissed out of hand), to an interest in the processes underlying and determining the choice of features used in any given ELF interaction (ibid.)

As elaborated on later in this thesis (section 3.1.5), a large body of research has been carried out which investigates various communicative functions of different linguistic 'building

blocks' of ELF interactions. Another research area has been that of attitudinal studies (cf. Jenkins, Cogo & Dewey 2011: 304 for a summary) of a number of different target groups.

The phenomenon of English used as a lingua franca was first recognised in applied linguistics (Seidlhofer 2001), before the underlying mechanisms of this language use were described. Now that research into ELF has begun to shed light on some of these mechanisms, it has also a number of potential implications, especially for language pedagogy, namely "the nature of the language syllabus, teaching materials, approaches and methods, language assessment and ultimately the knowledge base of language teachers"[12] (Jenkins, Cogo & Dewey 2011: 305). The relevance of the implications of ELF for teaching and testing (cf. e.g. Jenkins 2006, Jenkins & Leung 2013 and McNamara 2012), as well as other research fields, is now increasingly recognised in teacher education materials (Jenkins, Cogo & Dewey 2011: 305), and also illustrated by the fact that in 2014, the main focus of the *7th International Conference of English as a Lingua Franca* was on "Pedagogical and Interdisciplinary perspectives". However, as Jenkins, Cogo and Dewey (2011: 305) argue, suggestions of incorporating ELF into the classroom are still both scarce as well as controversial, and in any case, the main objective of ELF research is to be an enhanced awareness of the discrepancy between codifications of English and the reality of how English, and language in general, is actually used (2011: 306). Seidlhofer (2011b: 208) argues that another aspect which is transferable to teaching is the evidence drawn from descriptions of ELF. As such, what we know about ELF can inform teaching goals, as it can reveal what communicative competence means in international communication. In this light, a revision of traditional concepts is necessary, as they are "based on assumptions about the objectives and processes of learning that are outdated, and irrelevant, and unrealistic for most learners" (ibid.).

However, the implications of ELF are more far-reaching than merely concerning areas related to language teaching. As has been discussed in the present section and section 2.2., ELF frequently raises questions which are relevant to linguistics in general and calls for a re-consideration of traditional concepts used to describe actual language use. This is the case with the study carried out on word class shifts in chapter 5, which demonstrates that the notions of conversion and multifunctionality, which have been traditionally used to explain such cases, need to be reconsidered with regard to spoken, plurilingual discourse such as ELF. Another example of this is Santner-Wolfartsberger's (2012) paper on turn-taking which calls into question concepts such as 'party' which, she argues, need to be reconsidered in the light of ELF group interactions.

[12] Capitalisation removed from quote.

Moreover, in order to explain ELF more holistically, it has been viewed from a number of different angles and linked to other approaches and concepts. For example, Baird, Baker and Kitazawa (2014: 190) suggest that a fruitful way of explaining the variability of ELF is complexity theory and stress in particular, the importance of approaching a multi-faceted issue such as ELF from different angles. Trudgill (2014) argues that the comparison to other lingua francas, i.e. Greek, can also provide useful insights into the possible development of ELF.

In addition, being a global language and of enormous economic and social importance, it is not surprising that ELF is further linked and has engaged in research dialogues with a number of other disciplines and concepts. In particular, ELF has been seen as closely connected to issues related to globalisation (Jenkins, Cogo & Dewey 2011: 302), as a number of recent volumes show (Haser et al. 2012: 123), and even though ELF has not been specifically referred to in these, they show in general the "privileging of social activity over language structure: language events and experiences are the starting point, and observations about linguistic forms on different levels of language may follow suit as and when they are relevant." (ibid.).

Another concept which has been seen as being related to ELF contexts is 'superdiversity'. This was initially used to refer to the complex interplay and non-predictability of various factors apart from the traditional diversity with regard to ethnic background of migrants in Great Britain (cf. Vertovec 2007, Vertovec 2010) and has been regarded as more fitting than the traditional concept of multiculturalism (Vertovec 2007: 1027). The notion of 'superdiversity' is characterised "by a dynamic interplay of variables among an increased number of new, small and scattered, multiple-origin, transnationally connected, socio-economically differentiated and legally stratified immigrants who have arrived over the last decade" (Vertovec 2007: 1024)[13]. Cogo (2012) investigates the concepts of superdiversity within ELF business interactions, which she sees as "highly super-diverse under all categories considered" (2012: 289) and finds that the language practices at the internationally staffed company cannot be directly associated with being located in the UK, but that language is practiced democratically and that the participants regularly use their "multilingual resources for various purposes" (2012: 309) adapting them as necessary (ibid.). Cogo (ibid.) argues that "the distinction between the local and the global, the international and the domestic is certainly blurred in ELF contexts, with the resulting increased hybridity of linguistic repertoires".

[13] cf. also Vertovec (2010: 87).

ELF has also been investigated with regard to European multilingualism, especially as a work package within the DYLAN project (cf. e.g. Hülmbauer 2013, Hülmbauer & Seidlhofer 2011). Communication in ELF has been characterised as hybrid, multilingual, non-codified language use, transcending boundaries of what is codified in terms of English, as described earlier in this section, as well as the boundaries of English, involving other languages and displaying "flexibility beyond traditional language boundaries" (Hülmbauer & Seidlhofer 2011: 212). As such, ELF is a vital component of multilingualism, as it

> provides the possibility of extending the linguistics repertoire to account for the need for intercultural communication without undermining the role of diverse other languages and the expression of distinct sociocultural identities. From this perspective, ELF does not undermine the multilingual diversity but actually helps to sustain it. (Hülmbauer & Seidlhofer 2013: 399)

The issue of ELF, therefore, is a dynamic field of study with a growing body of empirical evidence. Moreover, it has far-reaching consequences for various areas in a globalised world which cannot be ignored, as mention made of it in other research areas demonstrates. The study of ELF poses great challenges not only when being described within the traditional boundaries of language categorisation, as will be shown in this thesis, but also when attempted to be captured within the boundaries of individual languages, or typical communicative situations. Hence, new concepts will have to be found and developed in order to describe the nature of global communication in English. Moreover, the precise effects of the study of ELF, especially with regard to language policy and planning, are expected yet to fully unfold.

2.5 VOICE and the nature of VOICE data

As has been demonstrated in the preceding section, research into ELF is now a thriving, interdisciplinarily approached and connected field of study. Among many other reasons, this is without doubt due to the accessibility of data, which has been provided in recent years by systematically collected ELF corpora, such as VOICE. This corpus was compiled and released at the University of Vienna with the goal to "close the 'conceptual gap'" (Seidlhofer 2001: 151) between the reality of the way English is used as a lingua franca and "an empirical basis for looking at the linguistic manifestations of ELF" (ibid.). VOICE provides such a basis for the investigation of transcribed, spoken ELF, providing easy access as it is freely available to all registered users of the corpus and as such, constitutes the first corpus of this kind.

The VOICE project ran between 2005 and 2013 and was financed primarily by the Austrian Science Fund and Oxford University Press. The data for VOICE was recorded, transcribed and prepared for publication between 2001 and 2009, which was when the first version of

VOICE was released. Further releases in 2010, 2011 and 2013 include updated versions (1.1. and 2.0) of VOICE Online, with minor corrections in the corpus texts and the release of 23 audio files, as well as downloadable XML versions of the corpus (VOICE 1.0/1.1/2.0 XML) and the online and downloadable POS-tagged version discussed in this thesis (VOICE POS Online 2.0, VOICE POS XML 2.0).

The corpus has a size of just over 1 million words, featuring transcriptions of 120 hours of recordings (VOICE Project 2013c). The transcripts cover 10 different speech event types (e.g. press conferences, service encounters, different types of discussions, meetings, panels, conversations) from the professional, leisure and educational domain and feature 753 speakers with 49 different first languages (VOICE Project 2013d). The 151 speech events, which are all included as a whole (Breiteneder et al. 2006: 166)[14], fulfil the criteria that they take place in ELF, are "[s]poken", "[n]aturally occurring", "[i]nteractive", "[f]ace-to-face", "[n]on-scripted" and that speakers took part in the interactions on their own estimation that they would be competent to do so (VOICE Project 2013d).[15]

The compilation of VOICE, the recordings and transcriptions of the data presented major challenges to the research team. In absence of available data or descriptions of ELF at the time, a definition for ELF data had to be created, in order to be able to select these in the first place (Breiteneder et al. 2006: 162f.). In order to avoid the potential circularity of criteria based on linguistic features, it was decided on an external, functional definition of ELF, "based on the characteristics and purposes of the speakers rather than on their linguistic output" (Breiteneder et al. 2006: 163). The seven criteria for inclusion of corpus texts are listed in the paragraph above. Other challenges described in Breiteneder et al. (2006) include theoretical questions of the definition of external criteria for data selection and of corpus balance, as well as practical issues such as the mode of the recordings and the capturing of additional information (Breiteneder et al. 2006: 170ff.), the challenges of representing spoken language in written form (Breiteneder et al. 2006: 171f.), e.g. human and computer readability of the transcripts, consistency through mark-up and spelling conventions, whether to choose British or American spelling for a representation of ELF, and how to represent the specific character of ELF in the transcripts through especially customised mark-up. For VOICE, special care was taken to ensure that the spoken language was represented as accurately as possible in the transcripts, and in that, to be as consistent as possible. This was ensured by

[14] However, some parts of speech events which did not meet the external selection criteria were excluded from the transcripts, e.g. large monologic parts (see Breiteneder et al. 2006: 167).
[15] For a more detailed characterisation and information on the availability of the different versions of VOICE see the VOICE Project Website (2013d).

very carefully designed and detailed transcription and mark-up conventions and through the transcription software *Voicescribe* (Breiteneder 2009: 23). Breiteneder et al. (2009: 22f.) write,

> Apart from the more conventional mark-up for, e.g., intonation, emphasis, pauses, repetition, word fragments, overlaps, or speaking modes, a fairly detailed set of descriptors was designed to account for frequently occurring features in ELF. These concern, for instance, code-switching, onomatopoeic sounds, pronunciation variations and coinages, as well as laughter.

The mark-up and spelling conventions were designed for the ELF data in VOICE, but were also intended to be adaptable for other types of data (Breiteneder 2009: 23). Moreover, a priority was that the data in VOICE could be accessed and used both long-term and flexibly, which was why the standardised formats such as XML and TEI were used (ibid.). However, as also user readability was a priority in the creation of VOICE, much work and effort was invested in the creation of an intuitively usable online interface with a variety of features.[16] A detailed account of the technical aspects of implementing the architecture for VOICE is provided in Majewski (2011). A number of examples have shown the applicability of, both the annotation scheme, as well as the architecture of VOICE. Not only has the VOICE annotation scheme been used for a number of individual research projects, as well as within academic courses in different European countries, e.g. Sweden and France. VOICE has also provided the basis in terms of architecture and methodology for the Asian Corpus of English under the supervision of Andy Kirkpatrick (2010, 2013), achieving two directly comparable corpora of English as a lingua franca. ACE is, as this is being written, in the process of being released for public usage.

Since its publication, VOICE has been widely received all over the globe, with over 7000 users, a total of 21 601 sessions and 476 733 page views (see Figure 1 and Figure 2).

[16] A detailed guide is provided in the online help file 'Using VOICE Online' VOICE Project (2009b).

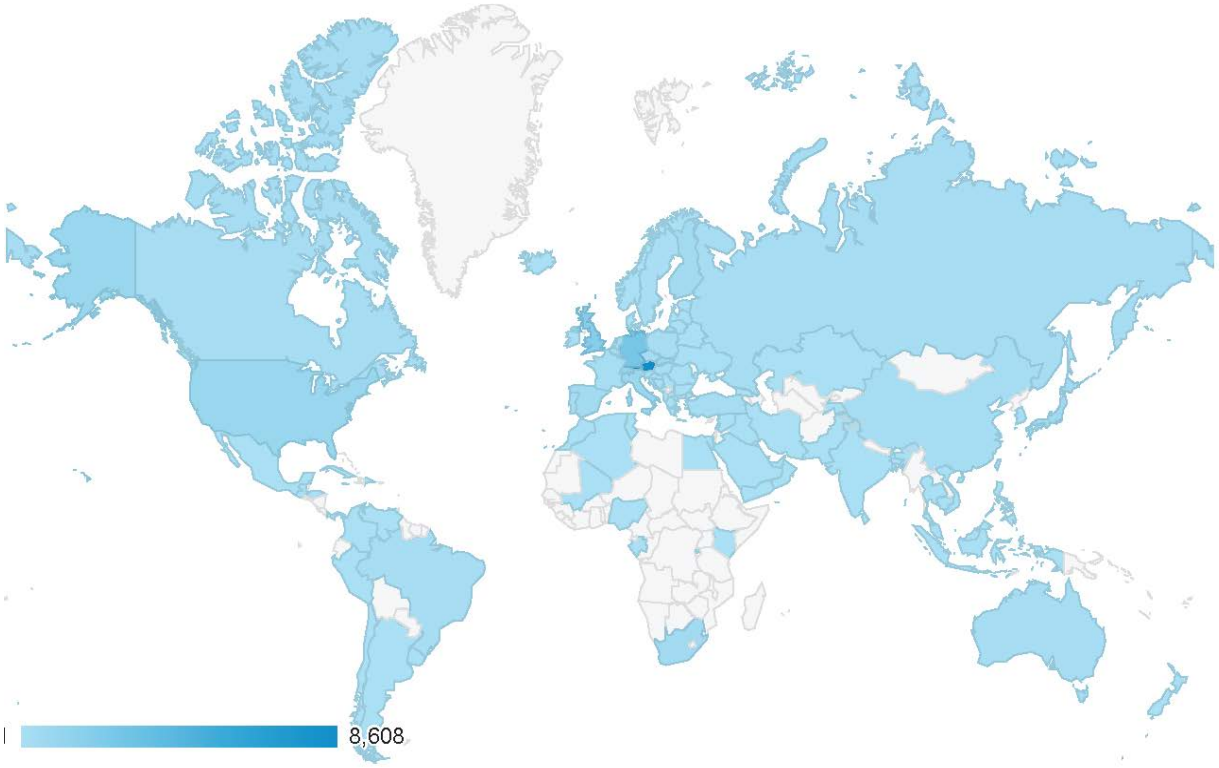

Figure 1: Map of all countries/territories with sessions on VOICE Online from 22 May 2009 to 21 September 2014. Source: Google Analytics

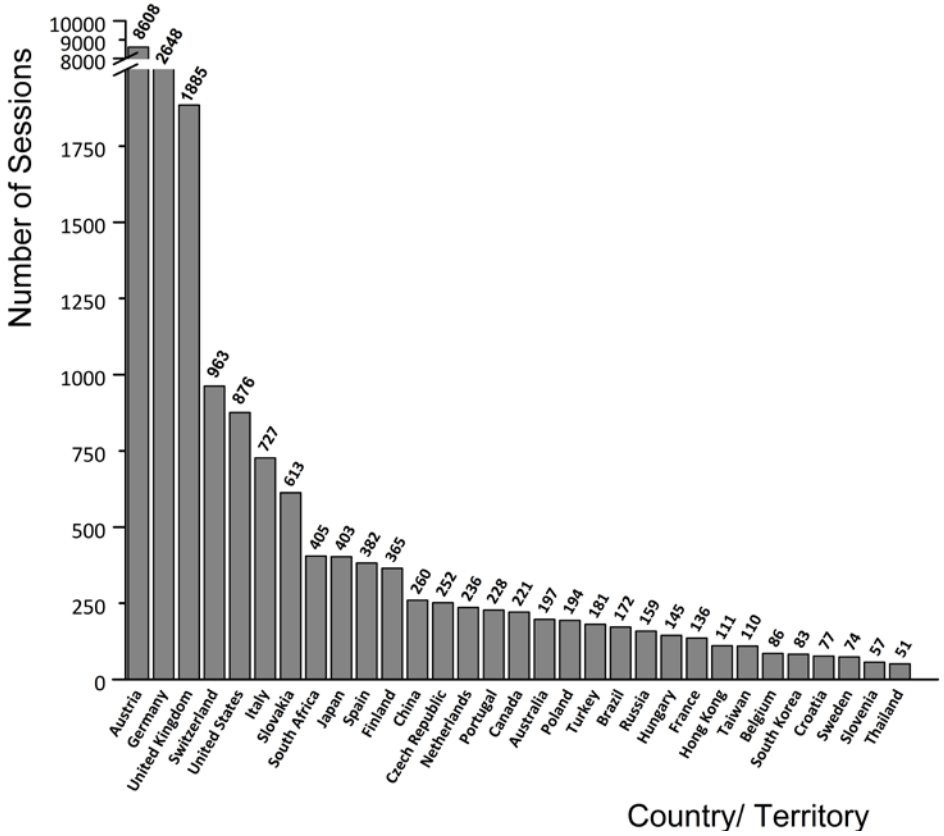

Figure 2: All countries/territories with more than 50 sessions on VOICE Online from 22 May 2009 to 21 September 2014. Source: Google Analytics.

Extensive research with VOICE has been carried out for MA and PhD theses at the University of Vienna, but also internationally, where VOICE has served as a basis for various studies on ELF. Aspects which have been investigated with VOICE include forms and functions of different lexico-grammatical areas, such as zero-realisation of 3rd person –s (Breiteneder 2009, Wacker 2011), phrasal verbs (Märzinger 2012), lexical innovations in ELF (Capone 2010), the discourse markers *you know* (Holzschuh 2013), and *so* (Reiter 2013), possessive markers (Vukovics 2013) or the progressive (Dorn 2011). On a macro-level, studies compiled have included pragmatic issues such as miscommunication (Pitzl 2010) and the use of idioms and metaphors (Pitzl 2011, Pitzl 2012) or broader questions regarding ELF discourse, e.g. relating to the negotiating and constructing of identity among the speakers (e.g. Bas 2010, Jenks 2013). VOICE has also been approached within different linguistic frameworks, such as, e.g. variation studies (Bürki 2013) or construction grammar (Pirk 2013).

The part-of-speech tagging of VOICE was built on the already carefully transcribed and processed data of VOICE 1.0 and 1.1. While sharing some challenges with the preceding transcription process, e.g. how to represent spoken English as a lingua franca adequately, some issues that arose were directly related to the fact that the tagging was done on the basis of the already transcribed recordings (see section 4.3.4) and some issues entirely new, i.e. unmentioned in previous literature, to the tagging of the data. This was due to the fact that part-of-speech tagging VOICE entails the description of ELF data in direct reference to existing word class categories, which present, I would argue, somewhat of an even tighter 'corset' than 'squeezing' spoken language into an orthographically represented form.

As has already been discussed, parts of speech are far from being straight forward to categorise or assign even for L1 language use, which part-of-speech systems were designed to describe. The tagging of VOICE, however, concerns ELF data and therefore an application of L1 categories to ELF language use. Hence, it also meant that "relying fully on existing English language tagging practices for VOICE would have constituted an attempt to apply a system of annotation to data it was not designed to account for" (VOICE Project 2014a: 5). The next three chapters of this thesis are concerned with how part-of-speech tagging has been carried out for other corpora of spoken, and to a very small extent, plurilingual language data, in how far these practices can be applied to VOICE and where they had to be adapted (chapter 3). This is followed by a discussion of the main challenges and solutions which emerged with regard to the task of categorising ELF data with codified language categories (chapter 4), and finally, a demonstration of how the part-of-speech tagging can point to interesting aspects in

the data when such challenges are being displayed, rather than glossed over in tagging procedures (chapter 5).

3 THE POS TAGGING OF SPOKEN LANGUAGE

3.1 Literature review: spoken POS-tagged corpora

3.1.1 Introduction

As already elaborated on in section 2.5, the tagging of VOICE is the annotation of naturally occurring, spoken and often highly interactive language data. With VOICE being the first corpus of spoken ELF to be part-of-speech tagged, there is no direct precedent to refer to for tagging methodology and procedure. However, there are a number of spoken corpora available, some of which are part-of-speech tagged. It is safe to assume that there are issues which occur when dealing with any naturally used, spoken language, e.g. regarding structure or particular forms inherent to spoken discourse. Similarities are expected to exist across different spoken usage situations, whether they take place in English or another language, in different dialects or standard varieties of a language, across mono- or multilingual and in inner, outer or expanded circle contexts. It is, therefore, useful to investigate how spoken language discourse has been dealt with in the tagging procedures of previous corpora and see how the challenges encountered by other corpus compilers are relevant for ELF data.

The literature on the POS tagging of spoken corpora is not very extensive and not always very informative, which makes it tedious for compilers of spoken corpora, especially those containing data from previously uncaptured domains, to gain insight into the issues which have already been encountered by others and the ways in which they have been handled. Leech (2000), for example, gives an overview of spoken corpora in English, but no detailed information on the tagging practices of the corpora is provided. Also, in the fast-changing world of corpora, 10 years are a long time and hence the information given by Leech (ibid.) is now partly outdated. In a more recent paper, McEnery and Hardie (2012: 71ff.) describe the development of English Corpus Linguistics, listing both written and spoken corpora. However, this review is restricted to English corpora, and while reference is made to corpora which have been compiled for other languages, these are not listed. Some ELF corpora, e.g. ELFA, VOICE, which had been compiled before 2012, are not mentioned at all in McEnery and Hardie (2012). Furthermore, no detailed information on tagging practices of spoken data is provided. While the reason for the unavailability of detailed, current accounts of part-of-speech tagging practices might lie in the fact that this annotation method has been seen as a "solved task" (Giesbrecht & Evert 2009), the large number of papers describing (often time-consuming) attempts to adapt the existing tagging practices and methods to new domains, written, e.g. websites (Giesbrecht & Evert 2009) or twitter (Gimpel et al. 2011), and spoken,

e.g. spontaneous language, e.g. new languages or multilingual communication (discussed in subsequent sections), show that significant issues in tagging remain. This has also been noted by other authors, e.g. Cresti and Moneglia (2005: 51), who write, "PoS tagging of a multilingual spontaneous spoken resource is still a new task and […] few antecedents can be found in the literature concerning the automatic PoS tagging of spontaneous speech". Hence, as Cresti and Monelia note for Romance languages, part-of-speech tagging is still of topical interest with regard to spontaneous speech, as there is a lack of tagging models. Even more so, this is true for the part-of-speech tagging of the data in VOICE, which is not only spontaneous and interactive but also inherently variable and multilingual. Numerous 'missing links' persist in tagging literature and standards with regard to such language data and hence presented a large challenge in the tagging process of VOICE.

The first part of this chapter will provide a brief overview of the types of spoken language corpora which have been part-of-speech tagged (sections 3.1.2, 3.1.3 and 3.1.4). Following on from this, the focus of the second part (section 3.2) is how particular issues relating to the tagging of spoken language discourse have been addressed in the literature and how they have been dealt with in tagging procedures. It will also emerge that many issues relating to spoken language use have not been discussed at all in the literature on part-of-speech tagging of spoken corpora and consider the reasons for the lack of detail of discussion of these features. Finally, the relevance of these findings for the part-of-speech tagging of VOICE will be addressed in section 3.3.

Before proceeding to review the available spoken corpora, it is necessary to mention that one reason why there are so few tagged spoken corpora is that there are relatively few spoken corpora to tag. The generally acknowledged bias in linguistics towards written language (cf. Linell 2005) is mirrored by ratio of written versus spoken language corpora, with the number of spoken, part-of-speech tagged corpora being relatively limited. Literature on spoken corpora makes mention of a clear bias on written language in corpus linguistics (Denison 2008), not only for English, but also for other languages, e.g. German (Rehbein & Schalowski 2012: 238). This is attributed to the time-consuming process of preparing spoken language for collection in a corpus. For example, Greenbaum and Yibin (1994: 35) write that, "[t]he spoken language […] will always be grossly under-represented because of the burdens in recording and transcribing speech". This is something which can be said to still apply to corpora two decades later, as Leech (2007: 9) notes more recently, "[s]poken texts must first be transcribed into written form, which means a continuing deficit of spoken (in comparison with written) corpus data."

Written corpora were also present before spoken corpora, e.g. the written Brown corpus which was compiled in the 1960s, while spoken corpora only started being compiled in the 1970s (Leech 2000: 678). However, apart from the many predominantly written corpora, e.g. the **Brown Corpus**, the **British National Corpus (BNC)** or the **Bank of English (BoE)**, often comprising millions of words, a number of spoken language corpora have also been compiled. While the majority of these are rather small projects (which is no surprise given the time-consuming process involved in processing spoken language data), some larger corpora of spoken language also exist. The largest and probably best-known spoken corpora for English are the spoken component of the BNC with 10 million words, **SWITCHBOARD**, comprising 3 million words (cf. Godfrey, Holliman & McDaniel 1992 for a description of the design of SWITCHBOARD), the **London Lund Corpus of Spoken English (LLC)**, containing "500.000 running words of spoken British English" (ibid.) and various sub-corpora contained in the **International Corpus of English (ICE)**. Twelve ICE corpora which are currently available (cf. ICE Corpora 2009a)[17] contain a spoken part of one million words each (ICE Corpora 2009a). Leech (2000: 681f.) additionally lists a number of other spoken corpora, e.g. **the Lancaster/IBM Spoken English (SEC)** or the **Cambridge and Nottingham Corpus of Discourse in English (CANCODE)**. Since this summary from over a decade ago, a number of other spoken corpora for English have been compiled, such as the **Bergen Corpus of London Teenage Language (COLT)** (Stenström, Andersen & Hasund 2002). However, while a considerable number of spoken language corpora exist, the majority of these have not been part-of-speech tagged. For example, out of the corpora that Leech (2000: 681f.) lists, only some, e.g. BNC, SEC, ICE-GB, Bank of English, have been part-of-speech tagged. This again, is doubtless due to the fact that the compilation and annotation of most kinds of spoken data is considerably more complex than the tagging of written data (Luzón et al. 2007: 3).

Still, even given the complexities of assigning word-class categories to spoken language, there have been a number of attempts to annotate spoken language data with part-of-speech tags, not only for English but also for other languages. In the review of literature on part-of-speech tagged, spoken corpora, we will, therefore, not restrict ourselves to English corpora, even though those have been very influential for corpus linguistics in general, but also include spoken corpora compiled for other languages. Some of these include a rather detailed

[17] According to the ICE website ICE Corpora (2009a), these are the ICE components for Canada, East Africa, Great Britain, Hong Kong, India, Ireland and SPICE Ireland, Jamaica, New Zealand, Nigeria, The Philippines and Singapore, which all contain a spoken part. For ICE-Sri Lanka and ICE-USA only the written part is available.

description of the issues encountered and solutions for tagging procedures developed for spoken corpora.

For the purpose of this literature review, we will sub-classify part-of-speech tagged corpora into those consisting of L1 data and then those containing L2 data. A further distinction is drawn between L2 and ELF corpora. In this thesis, the term '**L1 corpora**' is used to refer to those corpora which contain *primarily* monolingual first language data. The term '**L2 corpora**', then, refers to those corpora which contain second or foreign language data. Foreign language corpora are generally known as 'learner corpora' and are sometimes distinguished from 'user corpora', a distinction we will return to. A third distinction can be drawn between L2 corpora and **ELF corpora**. In L2 corpora containing second or foreign language use, speakers generally come from a single speech community. While the speakers are, by definition, multilingual in that they speak at least their first, as well as an additional language, the data in the corpora often only contain the L2 output of the speakers. Even if the speakers do share a first language, the data in the corpus consists only of the L2 and not the shared first or any other language. As such, the data in L2 corpora is in some sense also 'monolingual'. This is different for ELF corpora, which encompass the use of English as a means of communication by speakers of a variety of first languages (Seidlhofer 2011b: 7). Such communication does not exclude first language users of English and as most ELF speakers are multilingual, it will also include code-switches into languages other than English. As such, ELF can be defined as neither only L2 or L1, but as plurilingual language use (cf. Hülmbauer 2013).

I would argue that these three types of corpora all constitute, in essence, examples of natural language usage, some recorded in more interactive, 'natural' settings than others, and we can thus expect them to exhibit similar features, e.g. a certain level of variation and non-codified forms. However, it seems useful to roughly distinguish between these three types of corpora in nature as we do expect the quantity of these features to differ. This may be due to a number of reasons, e.g. Cook (2002b: 4) mentions, i.e. that an "L2 user is a different kind of person, not just a monolingual with added extras", who has additional uses for the language such as code-switching or translating (Cook 2002b: 4f.), a different knowledge of both the L1 and the L2 to monolinguals (Cook 2002b: 6f.), and his/her mind works differently from that of monolinguals (Cook 2002b: 7f.). For a corpus where speakers are multilingual, this means that for example, the number of code-switches into a different language are likely to vary. A second reason for distinguishing especially between L1, L2 and ELF corpora is that, despite the similarity of features they exhibit, they have often been dealt with rather differently in the

tagging procedures and the theoretical considerations thereof. This will be shown and discussed critically in section 3.1.5. It is clear, however, that a distinction between L1, L2 and multilingual data, as in ELF data, is often not clear-cut, as language situations in real life often include speakers with many different language backgrounds. Examples are the German corpus KiDKo and the corpus of spoken Irish (see below), which both contain L1 and L2 speakers of English. The way this multilingual data in these corpora is treated in tagging raises a number of questions, especially in relation to the annotation used in so-called 'learner corpora'. This will be discussed in more detail in section 3.1.5 and chapter 4 of this thesis.

3.1.2 L1 English corpora

English corpus linguistics has no doubt been highly influential in the field of corpus linguistics (cf. McEnery & Hardie 2012: 71). McEnery and Hardie (2012: 71ff.) give an overview of English corpus linguists and name universities where English corpus linguistics has been intensely developed. While the first corpus, the Brown Corpus, was compiled in the USA, universities in Europe were especially active in developing the field of corpus linguistics, including various forms of annotation and computerised corpora, beginning in the 1960s. In the UK, these were the University College London (UCL), Lancaster University, the Universities of Birmingham and Lancaster, in other parts of Europe for example the University of Louvain. In the USA, universities like the University of Pennsylvania, initiating the Linguistic Data Consortium (LDC), was influential, as well as Brigham Young University, the University of Michigan and the University of California Santa Barbara (McEnery & Hardie 2012: 90).

The corpora developed at these universities have been largely written corpora; however, some of them contain spoken material or are entirely spoken. A number of these spoken components have also been part-of-speech tagged. The **LLC** and **SEC** were the first corpora available for research and were annotated with prosodic annotation (Leech 2000: 682), and later also with part-of-speech annotation. The LLC, containing "500.000 running words of spoken British English" (Greenbaum & Svartvik 1990) was "analysed grammatically" (ibid.), though "[t]he full transcription and the grammatical analysis are available only on the slips at the Survey of English Usage at University College London" (ibid.). SEC was tagged with CLAWS1 and CLAWS2 (Constituent-Likelihood Automatic Word Tagging System, versions 1 and 2), a tagger developed at the university of Lancaster (UCREL 1993a-2014).The type of data in SEC is mostly monologic, though some dialogic material was also included (Taylor & Knowles 1988). The recording contained only data from speakers "whose accent was as close to RP as possible" (ibid.). Additionally, the recordings were cut to exclude "unintelligible

speech, background noise, or any feature that was felt to be unacceptable in the corpus" (ibid.) and hence it can be said that the data was to some degree adjusted in order to suit the methodology. SEC was tagged with CLAWS versions 1 and 2 (Taylor & Knowles 1988). Leech (2000: 682) observes that the type of data the LLC and SEC "suffer from restrictions", naming the variety of data they contain, which is for example "scripted, public speech" (Leech 2000: 683) for the SEC, and the fact that these are limited to British English data.

In the following decades, much larger corpora with a greater variety of spoken language data appeared, namely the monitor corpus **Bank of English**, totalling "524 million words of written and spoken English" (Xiao 2008: 394), the **BNC** with 10 million words of spoken material (Xiao 2008: 384) and the **CANCODE** with 5 million words of spoken material (Carter 1998: 55), including spontaneous data "covering a wide variety of mostly informal settings: casual conversation, people working together, people shopping, people finding out information, discussions and many other types of interaction" (Xiao 2008: 410). While the BNC and the BoE are, at least partly, available to the public (Xiao 2008: 385, 395), CANCODE remains unavailable for public use (Xiao 2008: 411) and while the BoE and the BNC have been annotated with parts of speech, no information on the tagging of CANCODE is available to date. The Bank of English was tagged automatically, and with the part-of-speech information containing about 5% rate of error, "should not be trusted blindly" (Bank of English user guide online, Titania). The spoken part of the BNC, which makes up about 10% of the whole corpus, includes formal, but also informal, speech situations (Burnard 2010). It is divided into context-dependent and demographic parts, e.g. "sales demonstrations, after-dinner speeches [as well as] conversations" (Garside 1995: 162). The BNC was tagged with CLAWS4 (Garside 1995: 162)[18]. However, Aston (1998) writes that the BNC is not suited "to study many features of spoken discourse: transcripts are orthographic, paralinguistic features are only roughly indicated, and situational description is limited." Included in the BNC is also **COLT** (Andersen & Stenström 1996: 133), which contains 0.5 million words (Stenström et al. 1998: 1) of conversation between teenagers between 13 and 17 years old (Stenström et al. 1998: 2) which were collected in 1993 (ibid.). COLT was annotated with CLAWS6.

Another well-known collection of corpora which have spoken, and to a large degree POS tagged, components is the **ICE**, which is composed of a number of individual, comparable corpora of about 1 million words (Greenbaum & Nelson 1996: 5). The data collected is

[18] Cf. Leech, Garside and Bryant (1994a) and Leech, Garside and Bryant (1994b) for a description of the tagging of BNC in general and Garside (1995) for tagging the spoken part of the BNC.

described as "standard or educated English" (Greenbaum & Nelson 1996: 5), the criteria on the ICE website for selection of participants being described as, "aged 18 or above, […] educated through the medium of English, and […] either born in the country in whose corpus they are included, or moved there at an early age and received their education through the medium of English in the country concerned" (ICE Corpora 2009b, cf. also Greenbaum & Nelson 1996: 5). The ICE corpora do not only contain data from inner and outer circle varieties, but also, e.g., material from Hong Kong (ICE-HK), which McEnery and Hardie (2012: 100) classify as belonging to the expanding circle. A number of these ICE components, e.g. Hong Kong, India, Singapore, Hong Kong, New Zealand, Canada, Jamaica and (the written component of) USA, have also recently been part-of-speech tagged with CLAWS7 (ICE Corpora 2009a). However, to the author's best knowledge, the only detailed account of the part-of-speech tagging procedure available to date is for ICE-GB (cf. Greenbaum & Yibin 1994).

SWITCHBOARD[19] is probably one of the best-known part-of-speech tagged spoken corpora compiled in the USA (Xiao 2008: 413). It was created in the 1990s and contains 3 million words of telephone conversations that were elicited and prompted (Godfrey, Holliman & McDaniel 1992: 518). Leech (2000: 716) writes that the data collected for SWITCHBOARD was "unnaturalistic in the sense that the speakers were assigned dialogue tasks which they then performed for the purposes of the recording" which he argues "cannot be easily applied to research relevant to human language learning and use of grammar, where the range of situations in which language might occur cannot be strictly controlled" (ibid.). Leech (ibid.) calls SWITCHBOARD a "task-oriented dialogue corp[us]". The part-of-speech annotation of SWITCHBOARD needs to be seen in the light of this description of the type of data it contains, which is, again, in many ways not comparable to the naturally occurring data in VOICE.

Three other corpora compiled in the USA, i.e. **the Michigan Corpus of Academic Spoken English (MICASE)**, **the Corpus of Contemporary American English (COCA)**, and the **American National Corpus (ANC)** have also been part-of-speech tagged. However, for MICASE, which comprises as size of 1.8 million words of various types of spoken, interactive data typical of academic settings, e.g. various presentations, meetings or discussion sessions[20], tagging with CLAWS (Tagset C8++) was only carried out "for in-house use" (Simpson, Lee & Leicher 2007: 10). COCA is the currently largest publicly and freely

[19] cf. Godfrey, Holliman and McDaniel (1992) for a description of the design of SWITCHBOARD.
[20] cf. Simpson, Lee and Leicher (2007: 15) for a list of the 15 speech event categories.

available corpus of English, with a size of 450 million words in June 2012 (Davies 2008-), out of which 95 million words are spoken language of "unscripted conversation" from radio and television (Davies 2008-, subpage: help-section). Although COCA has been annotated with part-of-speech tags, little mention is made of this in the documentation. Davies (2009: 164) merely remarks that, "[a]s a sidelight, we might mention that we use CLAWS-7 to tag the texts […]. Because the hardware for the corpus server is quite robust, we were able to tag approximately 25 million words per hour". The ANC contains 22 million words and has been tagged with various tagsets (e.g. CLAWS, Penn) (Xiao 2008: 386). It contains two sub-corpora, namely the **Open ANC (OANC)** and the **Manually Annotated Sub-Corpus (MASC)**, which is again part of the OANC. These are freely available and continuously being developed (American National Corpus Project 2012). OANC contains 15 million words, of which about 3 million are spoken and consist of the SWITCHBOARD corpus described above and Charlotte, which consists of about 200,000 words "narratives, conversations and interviews" (American National Corpus Project 2012). According to the project website, OANC has been annotated with the Penn Tagset. MASC has been richly annotated both with annotations from external sources as well as such which were "produced in-house" (American National Corpus Project 2012), and with regard to POS has also been annotated with the existing tools, i.e. Penn Tagset as well as with FrameNet (cf. American National Corpus Project 2012 and Ide et al. 2010).

More recently, a number of other, more specialised, spoken corpora for L1 English have also been compiled and part-of-speech tagged. A recent one is the **Newcastle Electronic Corpus of Tyneside English (NECTE)**. NECTE contains data from the two sociolinguistic projects 'Tyneside Linguistic Survey' and 'Phonological Variation and Change', and makes them available as a database long-term (Beal 2009: 35). The types of data are "loosely-structured interviews" (Beal et al. 2007) which were collected for purposes of sociolinguistic analysis (Beal 2009: 37f.). The data in NECTE is dialectal and the issues encountered in tagging this type of non-standard data have been discussed thoroughly in Beal et al. (2007).

Mention should also be made of **CHILDES**, a corpus of child data belonging to the TalkBank (cf. MacWhinney 2012a: 2375, CHILDES Website (see TalkBank), which has also been annotated with POS tags. The data in CHILDES are "60 million words of child-adult conversation across 26 languages" (MacWhinney 2012a: 2375). As taggers were not freely available and designed for written, rather than spoken language, the tool MOR was developed for part-of-speech tagging CHILDES and the TalkBank (MacWhinney 2012a: 2375). The tools used for the transcription and annotation of CHILDES come with a detailed

documentation (cf. TalkBank, subpage “Manuals”) and are available for download (MacWhinney 2012a: 2357f.). The issues encountered in the tagging process are discussed in detail in MacWhinney (2012a).

Finally, there is the **CHRISTINE Corpus**, for which spoken language data from different English L1 corpora, i.e. the BNC, LLC, and Reading Emotional Speech Corpus, was taken (Rahman & Sampson 2000: 298) in order to develop “a rigorous, predictable structural annotation scheme for spontaneous speech” (Rahman & Sampson 2000: 300). The CHRISTINE corpus, which adapts the annotation guidelines developed for SUSANNE, comes with a detailed documentation available online (cf. Sampson 2000).

In summary, it can be said that English corpora have no doubt been highly influential both in the development of spoken corpora and part-of-speech tagging practices. However, it also emerges that the part-of-speech tagging practices of the majority of the most well-known English L1 corpora with spoken data have a number of limitations. The data tend to come from genres which are more similar to written language, e.g. formal or controlled speech, and with speakers who allegedly produce RP, standard or educated English. In some cases, the data have often been edited, in essence to remove features typical of spoken language. This shows that, in fact, it is exactly the unpredictable, variable character of spoken language that has often been compromised. Of course, this is to some degree inevitable, as the only way to deal with spoken utterances is to convert them to written language, namely by transforming them into transcripts. However, it seems that the data has often been adjusted to suit the respective methodology, and to ease the task of annotating the corpora. The way the data has been adjusted to suit the needs of the corpus compilers by pre-selection of spoken genres as well as the removal of typical characteristics of the data is certainly something which needs to be taken into account when carrying out linguistic studies based on these sets of data, as well as for the tagging of other corpora containing different types of spoken data which are much more dissimilar to written genres.

Considering that the data have been manipulated in various ways beforehand, removing various sources of controversy, it is not surprising that tagging documentation, where available, often concentrates on the description of the tagset and distinction of different tags and that the discussions of other challenges in the categorisation of spoken data is rare. For tagging ventures of more interactive, variable, and, in case of VOICE, also plurilingual data, the methods used in these corpora for tagset design and tag categorisation are therefore only helpful as long as they conform to the rules of written, codified English.

Three corpora described in this section which pose exceptions are the specialised corpora NECTE and CHILDES, as well as CHRISTINE, all of which consider practical as well as theoretical issues arising from the part-of-speech tagging of their respective data, and as such pose valuable starting points for the tagging of VOICE.

3.1.3 L1 non-English corpora

As elaborated above, English Corpus Linguistics have been highly influential in the building and annotating of corpora (McEnery & Hardie 2012: 71). However, a number of spoken corpora for languages other than English have also been compiled and have been annotated with parts of speech. These are relevant for the part-of-speech tagging of VOICE as they, too, work with existing standards and tagging procedures, as well as available taggers, and apply them to other data than the often written, standard language they were designed for originally. Additionally, these corpora, many of which were compiled recently, generally display a rather open-minded attitude towards characteristics of spoken language, which is reflected in the tagging schemata developed. This section aims to list the most relevant spoken, POS tagged corpora for L1s other than English. However, it may not represent an exhaustive list as there are a number of spoken corpora which have been compiled and have reportedly been part-of-speech tagged, but for which no detailed documentation of the tagging procedure is available[21]. Other corpora, such as the C-ORAL-ROM, a corpus of spontaneous speech of various Romance languages, does contain a description of the POS tagging, however, this was done automatically and could not address important issues concerning the nature of spontaneous speech (Cresti & Moneglia 2005: 51ff.). Hence, in this section, priority is given to those corpora which describe the Part-of-speech tagging of their spoken data in some detail and/or which have some relevance for the tagging of VOICE in other respects.

In recent years, a number of spoken corpora were part-of-speech tagged for Scandinavian languages. Especially worth mentioning are the **Nordic Dialect Corpus** and the **Gothenburg Spoken Language Corpus (GSLC)**. The **Nordic Dialect Corpus** (cf. Johannessen et al. 2009), compiled between 2006 and 2008, is an example of a recent project (ScanDiaSyn, subpage: Data collection) which largely applied tagging procedures developed for English to other languages. It is a collection of spoken language data from six different North-Germanic languages, i.e. Danish, Faroese, Icelandic, Norwegian, Övdalian and Swedish (ScanDiaSyn, subpage: Data collection) which is availabile online and contains "2,8 million words from

[21] Spoken POS-tagged corpora which have reportedly been part-of-speech tagged but for which no further information can be traced include corpora of Asian languages, e.g. Japanese (Maekawa et al. 2000) or Mongolian (Dawa et al. 2006) or European languages, e.g. Dutch (Son & Pols 2001), Greek (Goutsos 2010), Ingrian Finnish (available online via Metashare) or Italian (Bellini & Stefan 2006).

conversations and interviews by dialect speakers" (ScanDiaSyn). The type of data includes interviews, as well as informal conversations (ScanDiaSyn), as recording "spontaneous speech" was a priority (Johannessen et al. 2009: 74). Existing taggers based on written language were adapted for the tagging of the individual components of the Nordic Dialect Corpus, using different taggers (Olso Bergen Tagger, TreeTagger, TnT tagger) as is explained in the Nordic Dialect corpus website (ScanDiaSyn, subpage: Technical Solutions). The tagging procedures of the individual components are described in, e.g. Johannessen and Jørgensen (Johannessen & Jørgensen 2006), Jørgensen (2007) and Nøklestad and Søfteland (2007) for the Norwegian part, Kokkinakis Johansson (2003) for Swedish, and Trosterud (2009) for Faroese. Nivre and Grönqvist (2001: 48) report on the tagging of the **GSLC**, which, among corpora of other languages, contains a corpus of spoken Swedish with data from a wide range of different speech situations, e.g. shop encounters, therapy sessions, interviews, bus driver-passenger conversations or auctions (Allwood et al. 2000: 3). The Swedish part was part-of-speech tagged by extending a tagger trained on written Swedish for the spoken data. This obtained very good results (up to 97%, cf. Nivre & Grönqvist 2001: 69), an even "higher accuracy for spoken language than for written language" (Nivre & Grönqvist 2001: 48). In order to achieve this, the tagset was extended with tags for "feedback" and "own communication management" (Nivre & Grönqvist 2001: 56). Other tags, such as a tag for foreign words, were adapted to suit the data (ibid.). The way the tagging was done, e.g. the algorithms used, is explained in detail in Nivre and Grönqvist (2001), even though theoretical questions on the tagging are not raised.

The **Spoken Dutch Corpus** is another large spoken corpus which was lemmatised and part-of-speech tagged (van Eynde, Zavrel & Daelemans 2000). It was compiled between 1998 and 2003 and contains over 10 million words of spoken language (Oostdijk 2000). It seeks to respond to the "great demand for spontaneously spoken language data; this explains the overall bias towards unscripted language" (Oostdijk 2000) and the great majority of texts feature two or more speakers. The type of data is "spoken standard Dutch, not dialect data" and there is no online availability (Johannessen et al. 2009: 79). Out of the ten million words only one million were annotated with parts of speech, including a range of different types of data, e.g. more formal or scripted texts, e.g. lectures and news reports, but also interactive texts, such as discussions, debates and face-to-face conversations (Oostdijk 2000). The Dutch spoken corpus was tagged with TnT and a combination of different taggers in order to improve the tagging result (van Eynde, Zavrel & Daelemans 2000). The lemmatisation and tagging of the Spoken Dutch Corpus is described in detail in van Eynde, Zavrel and Daelemans (2000).

The **corpus of spoken Irish** reported on in Uí Dhonnchadha (2012: 5), is a diachronic collection of currently 140 000, but aiming at a total of 2 million, words of spoken Irish, ranging from 1930 to present day Irish. The majority of the data (70%) will be dialogic (Uí Dhonnchadha, Frenda & Vaughan 2012: 3), such as "face-to-face conversations, phonecalls [sic] and interviews", the remaining 30%, which are monologous, are divided between scripted and non-scripted recordings, i.e. "broadcast discussions and interviews, parliamentary debates, classroom lessons, business meetings" with "at least two participants" (Uí Dhonnchadha, Frenda & Vaughan 2012: 4). An interesting aspect is the multilingualism contained in the data of the corpus. This is due to the situation in Ireland, where the two official languages are English and Irish, and not all speakers have Irish as their first language. Hence, in the corpus, there are NS and NNS mixed, including "(a) 'traditional' native speakers, (b) non-native speakers and (c) 'non-traditional' native speakers" (Uí Dhonnchadha, Frenda & Vaughan 2012: 2). A part-of-speech tagging of the already available data has been carried out, but the outcome has not yet been evaluated (Uí Dhonnchadha, Frenda & Vaughan 2012: 5).

Finally, mention should be made of the **Kiezdeutsch-Korpus (KidKo)**, a corpus of "a variety of German spoken by adolescents from multiethnic urban areas" (Rehbein, Schalowski & Wiese 2012a: 29). The main corpus contains only 333 000 tokens (KiezDeutsch-Korpus, subpage: Main corpus and complementary corpus). However, the tagged set of data described in Rehbein, Schalowsi and Wiese (2012a: 31) comprises only 1,265 tokens. Thus, both the data volume as well as the amount of tagged data are considerably less than those in VOICE. Irrespective of this, the type of data recorded for KiDKo is in many ways similar to the ELF data in VOICE: it is characterised by its spoken, spontaneous nature which the teenagers (14-17 years old) recorded themselves (KiezDeutsch-Korpus, subpage: Main corpus and complementary corpus), but also by the fact that is contains many nonstandard structures (Rehbein, Schalowski & Wiese 2012a: 29). The speakers in KiDKo are to a large percentage (84.4% for one group, 4.8% for the other) plurilingual, i.e. the language spoken at home was a different one than German, yet the speakers converse in German (KiezDeutsch-Korpus, subpage: Main corpus and complementary corpus). KiDKo differs from VOICE in that the speakers of KiDKo all speak one 'variety', namely Kiezdeutsch (cf. Wiese 2013) whereas ELF cannot be considered a variety (cf. Seidlhofer 2011b: 25), and, though the speakers of KiDKo may have different cultural backgrounds they can also be assumed to share cultural knowledge, as they all grew up in the same geographical area. Due to these characteristics of the multilingual language backgrounds, it is difficult to classify KiDKo either as a typical L1 corpus or an L2 corpus but it is somehow a hybrid, comparable to some extend also to the

corpus of spoken Irish described above. Unlike the corpus of spoken Irish, however, for which no discussion of the details of the tagging procedures is available yet, the description of KiDKo addresses a number of theoretical issues about the tagging of the data. That a number of issues arise is hardly surprising, considering the variation encountered and the fact that the compilers, too, try to apply systems developed for standard written data to spoken, multilingual data.

This section focussed on L1 corpora for other languages than English. Although it was mentioned that a number of L1-non English corpora exist for which little documentation is available, or the tagging has been done merely automatically, a number of corpora have been listed which deal with interesting aspects regarding the tagging of spoken data. In general, it can be said that the corpora listed in this section all place priority on the annotation of spontaneous, often interactive, and in two cases (the corpus of spoken Irish, KidKo) even multilingual of spoken language, including typical characteristics of such data. It does not come as a surprise that the categorisation of such data with traditional word class categories based on written language, is challenging in many ways and prompts the discussion of these issues. This is evident in the documentation of some of the corpora described in this section. Such issues include not only the adaptation of taggers and tagging schemes originally developed for written language, but also more general issues of the categorisation and description of spoken discourse. As such, these corpora can serve in many ways as guidance for the tagging of VOICE.

3.1.4 L2 corpora

As demonstrated in the preceding sections, there are a number of L1 corpora which have been annotated with parts of speech, the data incorporating not only monolingual but also multilingual individuals, as is the case for KidKo, and sometimes non-native speakers of the language which is the main language of the corpus, as in the corpus of spoken Irish or components of ICE, e.g. ICE-HK. The situation is very different for spoken corpora for which primarily other languages than the speakers' L1 have been recorded, i.e. data from language users who communicate in (a) language(s) other than their first, henceforth referred to as 'L2'. For the purpose of this section, a fairly general use of the term 'L2' is assumed. I am, as is Mauranen (2011: 156), using 'L2' to mean "foreign, second, third, nth and any non- first language", thus referring to all manifestations of English beyond the Inner Circle. Hence, the term 'L2' in this section covers both corpora of second language speakers, i.e. where a language [i.e. English in the following quote] is "an official (i.e. institutionalised) second language (L2) in fields such as government, law and education" (Jenkins 2005: 4) and foreign

language corpora, typically referred to as ‘learner corpora’. However, the boundaries between second and foreign language have been argued to be blurred (cf. Jenkins 2005: 15ff.) and second language corpora often also contain foreign language use. For example, ICE-HK contains data that would otherwise be classified as ‘learner language’ (McEnery & Hardie 2012: 100), as in Hong Kong, “English is spoken mainly as a foreign language (what Kachru 1986 dubs the ‘expanding circle’ of world English)” (ibid.). Note, however, that data of ELF corpora is not included in my definition of L2 language. The terms *L2*, *Foreign* or *Learner English* and *ELF* or *User English* have often been distinguished in the literature. The distinction itself, however, has been drawn with regard to different parameters and is open to debate, as will be examined more closely in the subsequent section (3.1.5). Moreover, ELF corpora are not discussed in this section as there were, to our knowledge, no POS-tagged ELF corpora available at the time that VOICE was being tagged.

L2 corpora and ELF corpora have often been compared, particularly because of the large proportion of non-L1 language in both types of corpora. Therefore, for POS tagging VOICE, an evaluation of approaches taken in spoken L2 corpora was useful in order to assess whether the annotation methods used in L2 corpora could be adapted for ELF data. The aim of this section will be to describe those attempts at part-of-speech tagging L2 corpora, as well as comparing the approaches adopted there to the theoretical approach adopted in VOICE in order to determine similarities and differences.

The part-of-speech tagged corpora containing spoken L2 English are those in the collection of **ICE corpora**, e.g. ICE India, Singapore, Hong Kong, Jamaica, which are downloadable from the ICE Homepage (2013), ICE East Africa (ICE-EA) and ICE Philippines (ICE Project 2013, subpage: POS-tagged and semantically tagged Corpora). The type of data recorded there corresponds, in essence, to that of other ICE corpora described above, apart from when practical issues in the data collection prevented this, as described e.g. for the ICE-EA (Hudson-Ettle & Schmied 1999: 7ff.). As described above, all ICE corpora have been tagged with CLAWS7 (ICE Project 2013, subpage: POS-tagged and semantically tagged Corpora) and the UCREL CLAWS7 tagset has been used. Unfortunately, no documentation with regard to the adaptation to the specific characteristics of the type of data in the ICE corpora containing L2 data can be found (an exception being ICE-EA, see below). With regard to non-standard language use, for example, the general ICE tagging manual (Nelson 2005) only makes mention of an UNTAG and a tag ‘?’ which were used in cases of unclear tagging, however, no further explanation is given on whether specific issues in L2 corpora were dealt with in a particular manner. An exception is the manual for ICE East-Africa (ICE-EA) for

which is it briefly mentioned that a strategy in general corpus annotation was 'normalisation', which meant that "deviations, mistakes, either lexical, idiomatic or grammatical" (Hudson-Ettle & Schmied 1999: 14) were approximated to "a first-language norm" (ibid.). In general, however, it seems that these were tagged according to the same principles as the ICE corpora containing primarily first language use (e.g. ICE-GB).

The second type of L2 corpora, including essentially speakers of the expanding circle, usually referred to as learner corpora, have often been annotated with 'error tags' rather than part-of-speech tags (cf. Gilquin & De Cock 2011: 151f.). This annotation "consists in marking the errors [with regard to a target language – ed. note] in a corpus and, usually, providing a corrected form for each error" (Gilquin & Cock 2011: 151). It is clear from this description that error tagging and part-of-speech tagging are different levels of annotation (cf. Gilquin & Cock 2011, Pravec 2002: 97). It has been acknowledged that the process of error tagging is far from unproblematic (Gilquin & Cock 2011: 151), as it is sometimes difficult to "identify", "reconstruct" and "correct" forms, as well as to "interpre[t] what the learner meant to say" (ibid.). Moreover, the procedure of error tagging has also been criticised from an SLA perspective for bearing the risk of committing the 'closeness' and 'comparative' fallacies[22]. From an ELF perspective, the 'error tagging' of L2 data can also be said to display major differences regarding the perspective which it assumes, as in error tagging procedures, the data is compared to a presumed native speaker ideal, as opposed to viewed in its own right (cf. e.g. Seidlhofer 2001). This is very clear in the way, e.g. Jendryczka-Wierszycka (2009) describes the uses of learner corpora, which, according to the author, can serve to conduct "over- and underuse studies, and computer-aided error analysis (CEA)" and in sum to show "significant differences against NS corpora." (ibid.). What is increasingly being called for within learner corpus research is a different approach, which is turning away from merely encoding differences to NS performance, and stressing the exploration of learner systematicity when annotating learner data, as it is has long been carried out in SLA studies. This is a view that is, for example, expressed by Meurers and Wunsch (2010). They write,

> [l]earner language is typically viewed as a linguistic system worth characterizing in its own right, so-called interlanguage. Thus learner corpora require systematic linguistic annotation of both correct and incorrect structures for them to effectively support the empirical questions addressed by SLA research.
>
> (Meurers & Wunsch 2010: 1)

As a consequence, it has been proposed that annotation, as it is generally used for corpora which deal with L1 data, should also be carried out for L2 corpora (cf. Díaz-Negrillo et al.

[22] cf. Rastelli (2009: 61) for an explanation of these terms.

2010: 2) and a number of learner corpora containing written data have indeed reportedly been annotated with parts of speech (e.g. Díaz-Negrillo et al. 2010, Granger & Rayson 1998, Meurers & Wunsch 2010, Rastelli 2009, van Rooy 2003 and van Rooy & Schäfer 2002). However, these concern written, not spoken data. In this process, a number of theoretical questions were raised, similar to those which were encountered in the part-of-speech tagging of ELF data in VOICE. For example, Meurers and Wunsch 2010 are interested in the question of how to best annotate learner corpora with parts of speech and suggest to consider different sources of evidence, e.g. syntactic and morphological (see section 3.2.3.4), and add different tags for each source separately. Similarly, Rastelli (Rastelli 2009) proposed an approach which he terms 'SLA-tagging' for a corpus of L2 speakers of Italian.

There are only a small number of *spoken* learner corpora (e.g. Granger 2008: 262) and of these, to the authors knowledge, two accounts of attempts to partially POS-tag these. These attempts are reported in Mukherjee (2007) and Jendryczka-Wierszycka (2009). Both tagged parts of components of the **Louvain International Database of Spoken English Interlanguage (LINDSEI)** with the CLAWS7, as reported in Mukherjee (2007: 372) and Jendryczka-Wierszycka, Rayson and Hoffmann (2009).

Mukherjee (2007: 371ff.) describes an attempt to annotate a part of the German part of LINDSEI (**LINDSEI-Ger**) with different levels of annotation, including part-of-speech tagging. The article states that the aim is for the entire LINDSEI-Ger component to be annotated in this way (Mukherjee 2007: 371). However, to the author's knowledge, this has not yet been completed. The annotated data consisted of "50 interviews with advanced learners of English" (Mukherjee 2007: 366) who were university students. Each interview consisted of three parts: firstly, a "prepared narration on a particular topic", secondly "a dialogic part", and thirdly "a monologic picture story" (Mukherjee 2007: 366). For evaluating the envisaged annotation schemes, one of these interviews of LINDSEI-Ger (No. 049) was annotated with a multi-level annotation, consisting of POS tagging, parsing, and error annotation (Mukherjee 2007: 371). In the paper, a number of examples of the tagging are given, however, the performance of the tagger and issues encountered in the tagging or the other annotations are not described.

Jendryczka-Wierszycka, Rayson and Hoffmann (2009) tagged 1000 words of the **Polish part of LINDSEI** (**PLINDSEI)** with CLAWS7. The primary goal of this experiment was, however, not a neatly tagged version of the data, but "[f]irst, to investigate the errors in spoken learner data and second to find ways of improving CLAWS' tagging accuracy on this difficult type of input." (Jendryczka-Wierszycka 2009: 12). Hence, this cannot be seen as a

genuine attempt to tag L2 data with the aim for it to increase searchability of the corpus. Rather, its aim was to investigate the data and evaluate in how far "errors and disfluencies" (Jendryczka-Wierszycka 2009: 12) of the L2 data influenced the performance of the tagger. Only a very short paragraph is dedicated to the actual experiment of tagging the data and the issues encountered, relating to the tagger inconsistency, problems due to the transcription format of the data, and "speech and learner-related issues" which the authors say "were closely connected with disfluencies so commonly present in learner speech". Unfortunately, no details are given on which L2-related issues were encountered, or on how these issues could be dealt with in the subsequent tagging process. As in Mukherjee (2007), the type of data in the Polish component of LINDSEI consisted of "rather informal interviews […] in a Question-Answer format" (Jendryczka-Wierszycka 2009: 2) which were carried out in English, between speakers of Polish (Jendryczka-Wierszycka 2009: 4).

In general, it emerges from the description of the L2 corpora that the strategies for annotating L2 data are scarce and only rudimentarily applicable to VOICE data. Another issue which arises are the different ways in which available L2 data have been treated in corpus annotation procedures, and how this, in turn, varies from strategies of annotating L1 data. In particular, differences have been made between data from expanding circle contexts, e.g. learner data, and that of user data. These differences are also relevant to learner vs. ELF corpora. In the available literature, ELF and L2 corpora (learner corpora in particular) and the data contained in these have been compared and related to each other. These correlations have implications, both for the research fields dealing with second, foreign language and ELF data, as well as for the annotation of these. However, in the tagging of VOICE the question arose what exactly constitutes the difference between learner and user data. It emerged that perhaps the difference made between learner and user data might not be as clear cut as is generally assumed. Therefore, before continuing to discuss the issues raised by L1 corpora relevant to tagging VOICE, we will consider how differences between learner language and ELF have been perceived, as well as the various implications of these differences and similarities on the investigation and annotation of these types of data.

3.1.5 ELF and learner corpora: between 'learners' and 'users'[23]

When dealing with L2 data, especially foreign language data, and ELF data, the question often arises how these are to be distinguished. Questions which arise are: Are the types of data different? Are the speakers different? And if yes, what are the distinctive characteristics, e.g. is it proficiency? Before the discussion of how individual aspects relevant to the tagging of

[23]Parts of this chapter have been published in Osimk-Teasdale (2013).

the ELF data in VOICE have been treated in the literature, I will therefore devote this section to the distinction of the notions of 'language learner' and 'language user' and that of the data available for these two categories of speakers.

The distinction between ELF corpora and learner corpora has often not been clear cut. Sometimes, ELF and learner corpora are kept quite separate; sometimes they are grouped together. Often, ELF corpora, such as **VOICE**, **the Asian Corpus of English,** the **Corpus of English as a Lingua Franca in Academic Settings, the Tübingen English as a Lingua Franca corpus** or the **Corpus of Academic Spoken English (CASE)** (cf. Diemer forthc.) are not mentioned at all in literature about learner, or indeed other second language corpora. For example, Gilquin and De Cock (2011) do not mention ELF corpora in their article which features various learner corpora. While McEnery and Hardie (2012) list the most important of native corpora which were developed at different universities in Europe and America, as well as the learner corpora ICLE and LINDSEI, there is no mention of the of ELF corpora ELFA, VOICE, TELF or CASE, which constitute an entirely new field in English Corpus Linguistics. On the other hand, Lee (2010) makes a distinction between learner and lingua franca corpora. However, he refers to them under the same heading "Learner Corpora, Lingua Franca Corpora" on his website which aims to list the most important corpora of English. What though, if any, are the differences between ELF and other data including speakers from the expanding circle? With regard to corpus annotation the answer to this question is not a trivial one, as it has implications for the way the data is annotated and numerous other issues, for example the way external points of reference are used in the annotation procedure.

Granger (2008: 260) addressed the distinction between learner corpora and ELF corpora and sees the foci of learner corpora and ELF corpora as merely "two sides of a coin" (ibid.). She draws the line according to **level of proficiency**. Granger (2008: 260) understands learner corpora to deal with those "speakers who are still in the process of learning the language" whereas ELF corpora featuring "proficient non-native speakers of English" (ibid.). In the literature, the difference between various L2 data and ELF data is often referred to with the terms 'learner' for the former, versus 'user' data for latter. Cook (2002a: 2) sees this difference between language users and language learners according to the **circumstances** under which the language is used: language learners "are acquiring a system for later use" (ibid.), for example by memorizing lexis and using it in role plays in artificial circumstances, so to speak, whereas language users "are exploiting whatever linguistic resources they have for real-life purposes" (ibid.). A similar view is proposed by Mauranen (2011), who, while acknowledging that "learner and L2 user corpora […] also have very much in common"

(2011: 157), such as "[t]he differences between first and second languages" (ibid.), assumes that the data itself is different. She discusses the differences between learners and users – ELF being the example of a user language – and differentiates the groups with regard to "social, cognitive, and interactive parameters" (2011: 157). With regard to social situation, Mauranen (2011: 157) argues, learners typically have the same L1 whereas users of ELF do not. Moreover, for learners, this L1 background comes with an awareness of the own, versus the target language cultures, which serve as "models for social appropriateness" (2011: 158), whereas in ELF, the culture connected to English as a native language is not of importance, as what is appropriate needs to be established in the respective communicative situation (ibid.). The language classroom constitutes a specific "social environment" (2011: 158) in which roles are assigned to the students as 'learners'. For Mauranen, 'users' and 'learners' are also social roles between which individuals alternate according to the situation, i.e. "in the classroom they are learners, but as soon as they get outside it they may turn into users of the same language" (2011: 158).

Another difference between learners and users Mauranen (2011: 159) sees is that while learners do not have a say in the development of the language they are acquiring, users have the potential to influence it with their language production. On a cognitive level, "ELF speakers as L2 users do not orient to their linguistic environment as a setting for language learning, but focus their efforts on making sense and making themselves understood" (2011: 161), and in this are focussed on conveying a message rather than on delivering 'correct' language forms. They also are confronted with unpredictable language use, e.g. "their interlocutors' accents, transfer features, and proficiency levels" (2011: 162). Finally, regarding interactional aspects, the rules of the discourse are in flux and need to be established by participants in ELF, while for learners these are a matter of following any ENL guidelines. According to Mauranen (2011: 162), "[i]n a lingua franca context, linguistic authority is not given, and participants are not seeking to learn the language from each other."

Again, a different categorisation is proposed by Seidlhofer and Widdowson (2009: 105f.), who view using and learning a language as ***processes***, whereby the users either have to either bridge the gap between "conceptual barriers between languages – between English and their own" or "communicative barriers between language speakers who only have English available to them as a common resource" (Seidlhofer & Widdowson 2009: 105). Therefore, in a usage situation what is learned has to be adapted to the fluid communicative situation of this particular setting, whereas in a learner situation the process is to learn the language (ibid.). In both situations, learning takes place, but this process is either defined by the dictates of

teaching and testing or it constitutes a natural learning process in a usage situation (Widdowson 2014).

It seems, then, that while the need to draw a distinction between the two notions of language learners and users seems to be present, no uniform definition exists in the current literature. I would argue that this is the case because the differentiation between the notions of 'language learners' and 'language users' is anything but straightforward and very difficult to draw at all. First of all, Granger's differentiation of language users being more '**proficient'** than language learners does not hold for ELF users, as ELF, as communication between non-native speakers for real-life purposes or "to go about their normal business" as Granger (2008: 261) words it, can happen at all levels of proficiency. The notion of 'language use' is to a certain degree 'self-selected' and happens when a speaker is confronted with a situation which requires the use of the language for a real-life purpose[24]. As such, the situations range from, i.e. a Spanish tourist with elementary knowledge of English ordering a coffee at Moscow airport to Turkish and Swedish business people discussing complex financial issues. Disregarding their level of proficiency, all of these people will be users in the situations described above.

Cook's and Mauranen's positions, in which differentiation between language learners versus users is drawn according to the **situation** in which the communication happens, is not unproblematic to apply in practice either. Even though one might agree that classroom situations differ from real-life use in principle, in that the latter are more unpredictable and 'real' than the former, I would argue that this differentiation paints a picture too simplified for the reality both in- and outside the classroom. Moreover, it is not directly transferrable to the data contained in learner and user corpora. For example, Mauranen (2011: 162) assumes a homogeneous, monolingual and -cultural classroom situation with native speaker orientation. Multilingual classes are mentioned (ibid.) but not included in her argument about learners and users of a language. Mauranen (2011: 162) proposes that if heterogeneity in a classroom exists, "the variation becomes familiar in the classroom, as speakers get used to each other's ways of speaking". I would argue that the assumption of such a homogenous classroom may be possible in some cases, but is also highly simplified. The motivation of students for learning a language may vary according to various factors, such as age and educational setting. Especially in English as it is often taught in adult education, which is often the setting for learner corpus data, students might have a variety of motivations, but these will often be practically oriented, e.g. to use English in their jobs with other NNSs of English. Moreover, in

[24] "Self-selected participation […] i.e. the speakers decided for themselves that they are capable of using ELF to accomplish specific participant roles in the speech event they are taking part in" (VOICE Project 2014b, subpage: 'Corpus Information') was also a criterion for the selection of speech events in VOICE.

adult education, speakers often have mixed language backgrounds, and the language of education is often the lingua franca in the classroom even at lower levels, as I have experienced myself when teaching lower levels of English for adults at the *Sprachenzentrum* at the University of Vienna. On the other hand, user situations, also in ELF are not always completely unpredictable and heterogeneous, as for example, it has been shown that ELF speakers often participate in communities of practice (cf. e.g. Ehrenreich 2009, Seidlhofer 2007, Smit 2010b), where there is also a more or less closed group of people who are used to each other's language habits, not dissimilar to what Mauranen describes for a classroom situation, cited above. Hence, it cannot be maintained that the learner and user situations can be strictly divided with regard to homogeneity or heterogeneity, respectively. I would therefore agree with Mauranen's (2011: 158) view, that individuals assume different roles, i.e. that of learners and users, and that "we do not have single or simple identities, but assume them situationally as is relevant, foregrounding and backgrounding our different identities and their elements, and drawing on them as the need arises in response to a social environment." (ibid.). However, contrary to Mauranen's (ibid.) view that the classroom necessarily triggers the speakers to assume 'learner roles', and the environment outside the classroom 'user roles', I would also argue that the role of a learner is not bound to the classroom, just as the role of the user cannot only happen outside the classroom, as Mauranen seems to argue, "when people enter an educational context as language learners, their position shifts from that of a user to whom a given language is the relevant means of communication." (2011: 158f.). Rather, I argue along with Seidlhofer and Widdowson (2009) that learning and using a language are manifestation of the same **process** which are inherently intertwined.

In the same way as has been argued above, the data which is recorded for learner corpora does not reflect a distinction of learners and users according to language proficiency or the communicative situation. This means that both learner and ELF corpora feature speakers with varied levels of language proficiency. Also, the distinction with regard to the communicative situation does not hold as learner corpora are not typically composed of classroom interactions. For example, in the components of the learner corpus LINDSEI, the language recordings were in an interview setting rather than a typical classroom setting. One could perhaps argue that the situation is remotely similar to an exam situation in a homogeneous classroom, e.g. speakers had time to prepare for the interactions in LINDSEI, and their interlocutor(s) shared the same first language, and as such, a small component of classroom reality is represented. Regardless, however, the situations in LINDSEI do not represent a 'typical learner situation' of, e.g. classroom setting, but, though prepared, a 'real conversation'. This makes it difficult to maintain that the data in learner corpora is data

typical of a learning situation. Moreover, though the data in VOICE clearly differs from data in LINDSEI in that it is entirely naturally occurring and speakers generally do not share an L1, there are also a number of interview situations (16 in total) which are not entirely dissimilar in terms of number of speakers and speech event type to those recorded in LINDSEI. Even more so, data in e.g. TELF are discussions in which speakers have been given a topic beforehand and which they are asked to debate – a situation not very much unlike those presented in LINDSEI (cf.TELF Project). Granger (2008: 261), in fact, suggests "to restrict the term 'learner corpus' to the most open-ended types of tasks", thus approximating the data in learner corpora language usage situations.

A second problem with the strict distinction between learners and users of a language according to communicative setting, it can be argued, is that learning and using are intertwined and cannot be easily separated according to situation in- or outside the classroom, or other (more or less) restricted situations such as those recorded for learner corpora. In fact, the distinction only holds if we are exclusively taking into account relatively restricted language situations in a classroom, which would typically count as 'learning' situations, e.g. restricted language practice activities. Once we start thinking about less restricted activities, and especially speaking activities in a communicative language setting, where functional language (e.g. a request) is practiced, as well as activities in which the students communicate with each other about their interests, exchange ideas, or discuss opinions, it becomes rather difficult to argue that this is very different from a user setting such as e.g. some of the speech events in VOICE, in which ideas are discussed. Even though the discussion may have been prompted by a teacher, rather than, e.g. a discussion leader of a company meeting, and the language situation in the classroom might create different pressures, the language output cannot be expected to alter all that much. Mauranen (2011: 162) argues that "[c]ommunicative simulations" are, albeit "realistic and meaningful […] not authentic in its basic sense of being real" (ibid.). As such, "[t]hey can be simulated in the classroom, but not performed there" and "cannot be assessed in class for their success or effectiveness in achieving their goals outside it" (ibid.). Although I would agree with this in principle, it can be argued, as above, that if communicative situations are realistic, the boundary to actual usage situations becomes very fuzzy and far from clear to draw. The argument becomes even more difficult to maintain if one imagines 'real-life' usage situations in a classroom taking place, e.g. when asking the person next to you for a pen or a piece of paper, i.e. when English is used as a lingua franca in adult education. Whichever way one looks at it, classroom settings cannot be reduced to simply 'learning' the language, but will contain some genuine language usage as well. On the other hand, learning can be said to be involved in typical

'usage' settings. This is what Ehrenreich (2009: 146) argues for the ELF speakers of two companies which she based her study on. She (ibid.) writes,

> By applying the "community of practice" framework of Wenger's social theory of learning, the managers I studied can usefully be conceptualized as life-long learners, moving in and out of different ELF-speaking CofPs [i.e. Community of Practices] as well as adapting to and actively shaping their communities' socially shared repertoires.

There are also plenty of examples in which speakers in 'real' situations in ELF ask each other for words. Klimpfinger (2009: 362 f.) names "*appealing for assistance*" [original emphasis] as one of the functions of code-switching in ELF talk. As Seidlhofer (2001: 144) argues, what has generally been called a **learning** strategy in learner corpus research, can also be viewed as a **communication** strategy. I would argue that these are two sides of a coin, and that we can also see learning taking place through a communication strategy, such as a request for help with a word, as demonstrated in Klimpfinger (2009: 362 f.). Mauranen (2011: 158) writes that sometimes users assume a learner role by precisely asking for words, for instance by "asking […] interlocutors about the correctness of our language". She maintains, however, that "these borderlines are not often transcended" (ibid.). To the author's best knowledge, no detailed studies on this issue have been carried out. However, a simple search for *word* in VOICE yields numerous results in which people ask for words, such as *what is the word i'm looking for* (EDwgd6:57), *can't think of a word you merge yourself in a culture and a language completely … do you know what i mean?* (EDwsd304:299 and 302), *do you know what's the english word for <L1ger> zecke? {tick} </L1ger>* (LEcon420:13), *what is the dutch word?* (LEcon560:1079). However, the usage situation seems to be different in the way that interlocutors are in a non-hierarchical position, so as opposed to being student and teacher, where the student asks for a word and the teacher gives the 'correct' answer, in the ELF situations, the word is more often 'put on the table' and everyone is invited to help with finding the correct answer, as demonstrated by the example below:

(1) S1: so: (.) to start (.) overcoming our differences and language barriers **just in case (.) e:r any of you (.) FEEL like (.) they cannot FULly say <soft><fast> whatever they would have liked to say** </fast></soft> in english (.) but there is a way of **HELping each other in: (.) in translation** o:r (.) er yeah then (.) <soft> then <fast> **i think we should we should try to help each other** so </fast></soft> just in case someone feels at some point a little bit (.) annoyed by not finding a word in english or something yeah {parallel conversation between S3 and S12 starts}<soft> (that should yeah) </soft> **say it in your language SOMEone around the table will figure it out in the end** {parallel conversation between S3 and S12 stops} (POwsd266:18; L1= rum-RO)[25]

[25] **POwsd266**: ID of speech event; **18**: Utterance number in the speech event**; L1=rum-RO**: First language of speaker, here: Romanian, **RO**: corresponding country, here: Romania (abbreviation of languages according to

Sometimes, speakers give a direct translation, e.g. in LEcon420:14, LEcon560:1090, other times, speakers offer translations in languages other than English, as in, e.g. EDwsd304; 299ff., LEcon229:146ff. What can be seen in any case, though, is that the usage situations, in which speakers are achieving real-life goals, clearly sometimes create the necessity to use words which are not part of the speakers' English repertoire and create situations in which learning of language-related aspects – of a new word or phrase, takes place. In example (2) below, S7 offers an alternative word (*retaliation*), to the one S3 uttered (*rehabilitation*), and S3 repeats this three times as if to acknowledge and remember it.

(2) S7: assessing damage and (.)
S3: th- th- i i there i would have two i would have (1) er (.) **rehabilitATION** one? (.) and two: (2) er not revenge but (1) <4> counter-attack </4>
S7: <4><soft> **retaliation** </soft></4>
S3: **retaliation retaliation**. that's the word. (2) <soft> **retaliation**.</soft> (2) <soft><un> xx </un></soft> or (.) now what do we (.) (d-) need for all this three. (.) for all this three.<fast> all these three what do they have in common </fast> (that) intelligence (1)

(EDwgd6:366-369, S7: L1= eng-US; S3: L1=alb)

These examples show that communication settings traditionally assigned to 'learning' or 'using' a language cannot be easily separated. Therefore, the notions of language learners and users cannot be clearly distinguished according to the situations in which they find themselves, as learning and using might occur in both situations. Moreover, the data found in learner corpora is not 'classroom' data as such, but somewhere in between a stereotypical learner and user situation. The distinction between language learner and language user therefore seems to be difficult to draw, and neither the notions proficiency, nor the communicative setting help with this distinction.

If the distinction cannot be clearly drawn, this means that the data in so-called 'learner' and 'user' corpora will not be dissimilar in some respects, as learner and user aspects overlap. Therefore, the data present will never only be very 'restricted' to 'learn' a language, nor will it always be merely 'used' without 'learning' going on. Seidlhofer (2011b: 189) trenchantly

---

the ISO 639-2 codes, abbreviation of corresponding countries according to the ISO 3166-1-alpha-2 codes). Examples taken from VOICE are either given in VOICE style (cf. VOICE Project 2009a) as, e.g. in examples (1) and (2), or in a reduced mark-up, where relevant tokens are given in the format used for VOICE POS Online, as, e.g. in examples (3) and (4) (cf. VOICE Project 2013b) for general information on the differences between the styles used in VOICE Online and VOICE POS online). Emphasis in bold, as in examples (1) and (2) is always added and not part of the original transcription.

argues that using and learning a language are two sides of a coin, which is especially visible in ELF:

> The wide use of ELF brings home the simultaneity of learning and using in a particularly striking way. I would suggest that learning and using English are best thought of as *aspects of the same process of exploiting the meaning-making potential* virtually inherent in the language. It is not surprising to find, therefore, that ELF usage bears a resemblance to learner language. It would indeed be surprising if it did not.[my emphasis]

The similarities in data and the different ways in which these are described and annotated lead to the assumption that the distinction between 'learner' and 'user' data is mostly a matter of a different **approach**. And indeed, this is visible with regard to the purpose of the two types of corpora. As Seidlhofer (2012: 144) states,

> ELF corpora consist of data produced when speakers of different languages draw on the formal resources of English to communicate with each other in naturally occurring contexts and their purpose is to provide evidence of how they do it.

For learner corpora, on the other hand the "purpose is to document the nonconformities, predicated on the assumption that they constitute errors that learners will wish to correct" (ibid.). In ELF contexts, these "nonconformities", Seidlhofer (ibid.) stresses, "far from being in need of correction, represent the effective communicative exploitation of linguistic resources". As mentioned above, 'learner corpora' are built with the goal to show differences to native speaker data, while 'user corpora' such as ELF corpora, aim at investigating the language at hand in its own right, as Seidlhofer (ibid.) argues. We have seen, therefore, that the difference in 'user' and 'learner' corpora cannot be said to be due primarily to differences in the data, which is similar, or at least overlapping, but rather to a **difference in perspective**, something which is also argued by Seidlhofer (2001: 143f.). Seidlhofer (2001: 143) writes, "[t]he main difference lies in the researchers' orientation towards the data and the purposes they intend the corpora to serve". This difference in perspective is, however, far from trivial and has certain implications with regard to analysis and annotation of the data. There are, it seems to me, five main implications, which will be discussed below.[26]

The first implication refers to the **description of language output of users and learners**, respectively, in the literature. Learner corpus research is often viewed in direct comparison to the output of native speakers, e.g. learner language is described in terms of under- or over-usage with regard to native speaker data (e.g. Gilquin & De Cock 2011: 154), with the aim to

[26] It may be added that the different perspectives are not something negative per se – learner corpora can, for example, be used for EFL teaching, with regard to an ENL target. However, the different perspective within ELF research has consequences for an ELF corpus, which will be discussed further in chapter 4 – as we had to consider new ways of applying POS tagging to L2 data.

determine "the gap between learner language and the native speaker norm […] in an empirically sound way" (Mukherjee 2007: 368). Or, in the description of part-of-speech tagging, the learner data is referred to as "difficult type of input" which is expected to "reduce the accuracy of POS taggers trained on native language data." (Jendryczka-Wierszycka 2009: 12). Similarly, Rooy and Schäfer (2002) write that the problems for POS tagging the Tswana Learner English Corpus were due to 'learner errors'. On the other hand, such derogatory formulations are generally absent in descriptions of corpora where the speakers are viewed as 'language users' e.g. in literature on KidKo, the corpus of spoken Irish or the relevant ICE corpora (but cf. ICE-EA as discussed above). ELF is generally described in a similar way as other user corpora, the assumption being that the language output is natural (see e.g. Breiteneder 2005), following certain functions.

A second implication is that **different points of view illuminate different aspects** in the data. Hence, some aspects will be visible from one perspective but not from another. For example, Smit (2009: 202), based on her longitudinal investigation on tertiary classroom discourse, argues that an approach which focuses on the description of non-native language as deviating from L1 actually misses out on describing features of ELF discourse. Therefore, the advantage of viewing ELF speakers as language users rather than learners is that it sheds light on a number of aspects which might otherwise be overlooked or disregarded. Similarly, as has been outlined above, Seidlhofer (2001: 144) argues, that strategies of learning can also "be regarded as communication strategies" (ibid.). So here, rather than e.g. dismissing instances of L2 speakers' language as "misspelled, badly uttered, incomprehensible and non-interpretable items" (Rastelli 2009: 57), unconventional forms are taken to be interesting manifestations of language in use as they shed light on the underlying communicative functions they serve. Seidlhofer (2009a: 241) argues for ELF that,

> the crucial challenge has been to move from the surface description of particular features, however interesting they may be in themselves, to an explanation of the underlying significance of the forms: to ask what work they do, what functions they are symptomatic of.

There are numerous examples in ELF research which have suggested that particular forms serve certain functions and ELF speakers have been shown to use their multilingual repertoires successfully in order to go about their daily business and achieve their intended (e.g. communicative) goals. For example, Klimpfinger (2009: 352, 359ff.) identifies four main functions of code-switching in ELF, namely "specifying an addressee, signalling culture, appealing for assistance, and introducing another idea". Another area, the non-codified use of certain morphological forms, such as 3rd person singular zero marking, has

been shown to serve the function of accommodation between interlocutors (Cogo & Dewey 2006: 84), and, more generally, "to acknowledge understanding, ensure the smooth development of the conversation, the synchrony of its delivery, and alignment" (Cogo 2009: 259)[27]. Other examples include research into the function of the progressive (Dorn 2011), of lexical innovation (Pitzl, Breiteneder & Klimpfinger 2008), of different pragmatic markers, e.g. Holzschuh (2013) and Reiter (2013) or the investigation of the function of filled pauses (Böhringer 2009), to name just a few. These functions cannot, of course, be accessed by an analysis which focusses solely on the 'over- and underusage' of various aspects. Another aspect which learner corpus research is generally not interested in is the fact that language users do already communicate with the resources they have, and do so very successfully, through "exploit[ing] the possibilities of the virtual language to their own ends, appropriat[ing] it for their own purposes" (Seidlhofer & Widdowson 2009: 103). This is the reason why Seidlhofer and Widdowson (2009: 102) argue that, while it is generally assumed that NS competence is the end of the journey, it can be shown by the way ELF is used, that what the learners learn is in itself the goal, since "what learners learn can be put to use as an end in itself" (ibid.).

As a third implication, the perspective adopted will naturally affect the **type of conclusion** drawn. For example, for Gilquin and De Cock (2008: 143), stating that learners are using hesitation markers in an inferior manner to native speakers, the view expressed is that the speakers need to be mainly intelligible to native speakers, e.g. to "giv[e] an impression of fluency" (Gilquin & De Cock 2011: 159, referring to Hasselgren 2002). The conclusion drawn there is that it may be "important to incorporate hesitation strategies into the foreign language curriculum, since they make it possible to increase fluency and help deal with planning pressure" (ibid.). If on the other hand, the view is that speakers are users, already competently applying hesitations in discourse, it is actually possible to shed light on certain functions of hesitations. For example, Böhringer (2009) was able to demonstrate that pauses have the functions of allowing more time for planning, weakening hierarchical structures and organising the discourse. The outcome of this research, i.e. the design of better exercises to train speakers to adjust to native speaker performance vs. the insight into how speakers deploy whatever language is available successfully, is therefore quite different.

Another implication of the different perspectives is that the **research done in both fields is often not mutually taken into account**. Whether this is because of the lack of awareness of the similarity of the data or of the other research (e.g. ELF being a relatively new research

[27] See also Breiteneder (2005) for an extensive account on the functions of zero-realisation of 3rd person –s.

field), or an entirely different reason is not clear. For example, Gilquin and De Cock (2011: 158) summarise from the literature that an analysis of errors which affect intelligibility would be useful for classroom practice, as these errors could then be focussed on. However, within the examination of 'user' language, such as, e.g. in ELF, there has been a large body of research on the intelligibility and effectiveness of lexical items and phonology and phonetics, e.g. Jenkins (2000), Hülmbauer (2010), Osimk (2010) and Osimk-Teasdale (2009), which is not mentioned in Gilquin and De Cock (2011). Cases such as these are unfortunate, as I would argue that both 'parties' could profit from the research done in the other field[28]. For example, the suggestion that spoken language strategies should be included in language teaching could profit from the fact that it has already been shown that users of English can very successfully employ various features of spoken language, such as hesitation markers, without probably ever having been taught them in a classroom setting. On the other hand, ELF studies can profit from insights gained from second language acquisition, something that was demonstrated in Osimk-Teasdale (2009), where I used experimental methods usually employed in SLA studies to test the intelligibility of different sounds in ELF communication. This is in line with Mauranen's (2011: 157) view that due to similarities between user and learner data "any research findings from either group are of interest to those studying the other". Regarding the ways in which both can profit each other, Mauranen (2011: 164f.) draws particular attention to exactly these differences between L2 (defining this term broadly to include outer and expanding circle, as outlined above) and ELF data:

> While learner corpora can inform ELF study about what linguistic deviations from ENL are common in SLA and thus likely to have learning-based explanations if found in ELF, ELF can in return enlighten SLA research by showing how L2 actually works in ordinary life outside educational contexts and what might be worth focusing on in teaching. Together, learner and ELF language research can get deeper into the nature of languages other than the first.

The fifth implication of the different perspectives on user and learner data concerns the **procedure of part-of-speech tagging** itself, assuming a different form of annotation depending on whether the perspective on the speakers in the data views them as learners or users of the language. As mentioned above, learner corpora have usually been 'error tagged', while other corpora containing second language 'user' data, e.g. KidKo, ICE-HK or the corpus of spoken Irish have been annotated with part-of-speech tagging, the procedure also carried out for VOICE. The spoken L2 data in the ICE corpora have been tagged no differently to L1 corpora. As in other L1 corpora, the data have been adjusted to be more

[28] It is noteworthy, that the relation between 'errors' and intelligibility has been a subject of much discussion under the heading of 'error gravity' in the literature of Error Analysis, cf. e.g. Burt (1975). However, the data collected in this field focusses mostly on written data, such as students' essays.

similar to codified language use. As mentioned earlier, for example for the ICE-EA, non-codified language use was changed to represent that of L1 standard language descriptions. The LINDSEI corpora, for which there have been the two POS tagging attempts mentioned, were tagged from a learner perspective. The issues encountered and discussed result in the conclusion that the tagger had problems dealing with the data because of the erroneous language of the speakers. Neither of these perspectives, i.e. assimilation of the data to L1 standard, nor selecting 'problematic' aspects related to the non-native nature of the data are compatible with the approach adopted in VOICE, where the aim is to annotate the data in its own right, from a user perspective, not seen as a faulty version of ENL. Hence, even though similar issues to those in learner and second language corpora may arise in the tagging of VOICE, the few solutions given, i.e. assimilation of the data to codified language use, are not satisfactory for tagging ELF data. It seems more appropriate to look for guidance at other corpora, i.e. the L1 corpora described in sections 3.1.2 to 3.1.4, which tag data from a language user perspective. If looking for solutions for the issues encountered in tagging VOICE therefore, how to categorise and display the data there in its own right (cf. e.g. Rehbein, Schalowski & Wiese 2012a: 31), we need to look at corpora which actually seek to display spoken language in its own right. The following section (3.2), will therefore primarily consider how different features encountered in the tagging of VOICE were dealt with in L1 corpora, and then how these are applicable to the tagging of VOICE in chapter 4.

In this section, we have taken a closer look at the distinction between learner and ELF corpora and the data contained in these. Three commonly drawn distinctions are listed, namely language proficiency, communicative situation and cognitive processes. The first two distinctions are shown to be problematic, as they present a simplified picture of language reality, assuming that learning and using can be neatly separated. The opposite is shown to be the case. The fact that learning and using are not easily distinguishable is also mirrored by the data contained in learner and user corpora, as it could be shown that both types of corpora contain data which could fall in either category. Hence, an understanding of learning and using as being two intertwined processes is argued to present a more accurate picture. The section goes on to suggest that, in essence, the notions of language learners and users, and their data is in essence a matter of perspective, one that has, however, far-reaching implications. These refer to the manner in which language output of users and learners is depicted in the literature, the way different perspectives lead to different insights of the data and the conclusions drawn. While learner data tends to be described in terms of how it deviates from NS data, highlighting the challenges which this kind of 'difficult' input presents for corpus compilation and annotation, research on ELF and other types of user data views

this language output as natural, focussing on the functions which such language serve. In addition, research on learner and user data is often not mutually taken into account and hence, it is argued, misses out on valuable findings from both fields. As a fifth implication, it is explained that these different perspectives also have an impact on part-of-speech tagging the data, as different annotations (error-tagging vs. word class tagging) are applied depending on whether the data has been categorised as learner or user data. It is concluded that for the tagging of VOICE, guidance is better sought from user than learner corpora, as will be discussed in the following sections.

3.2 Issues arising from POS tagging spoken language data

In the previous section, part-of-speech tagged corpora for both L1 as well as L2 data have been described. In this section, the issues which have been addressed with regard to the nature of spoken language will be described in more detail. When considering the way spoken discourse has been categorised in part-of-speech tagging procedures, there are three main questions which are relevant to consider: **Firstly**, to what extent are issues to do with part-of-speech tagging spoken language discourse **discussed at all** in the literature. **Secondly**, the question of which **general attitude expressed** towards issues which occur is addressed, i.e. one which views them as problems, or in a positive light, e.g. as interesting instances which might lead to insights into the nature of the data. The attitude towards the issues encountered is, in turn, likely to influence the solutions proposed for dealing with these issues in the individual tagging procedures. For example, if features of spoken language such as hesitations are viewed as disturbance, the approach might be to ignore them (as is done in, e.g., SWITCHBOARD), if they are viewed as interesting, they are more likely to be represented in the tagset. The **third** aspect of interest is **which issues** and difficult aspects have been addressed, and how these have been **dealt with in the individual tagging** procedures, i.e. how they are reflected in the tagging guidelines, in the tagsets, etc. Subsequently, it will be of interest to which degree the approaches adopted compare to the one that was chosen for VOICE and hence, which aspects are relevant to the tagging guidelines for VOICE.

However, before continuing the discussion of how aspects relevant to the tagging of VOICE were treated in accounts of other L1 data, it is necessary to explain that the discussion is limited to corpora containing L1 data, and, to a small extent, written L2 corpora. It emerged in section 3.1.4 that the strategies applied to tagging spoken L2 data can be regarded as either insufficient or unsuitable for ELF data. With regard to ICE, it can be said that in general, for the tagging of VOICE, the ICE tagging manual served as a starting point for some tagging issues, e.g. the sub-categorisation of formulaic items (VOICE Project 2014a: 30). However,

for issues specific to L2 language use, e.g. a critical reflection of the relevance of external, normative points of reference, or solutions for the variable language use of multilingual language users, there are unfortunately no guidelines to be found in the description of ICE corpora. As regards learner corpora, the two experiments on POS tagging the two spoken learner corpora PLINDSEI (Jendryczka-Wierszycka 2009) and LINDSEI-Ger (Mukherjee 2007) cannot serve as suitable starting points for VOICE either, for a number of reasons: First, the communication setting for the type of data collected is much more 'set up' and has indeed been artificially created: the speakers share their first language, yet converse in a language which is not their first. Moreover, the type of data is a rather restricted interview format: after a choice of one topic from a selection, and a few minutes preparation to think, the interviewees give a monologue about the topic. The interviewer then prompts a few more questions and finally, the interviewees narrate a story based on images which were shown to them (Jendryczka-Wierszycka 2009: 5). These artificially created communication settings are very different to the nature of VOICE data, which consists of a variety of naturally occurring speech event types and involves usually more than two of speakers with a variety of language backgrounds. The interactions would have taken place without our intention to record them and were determined by external tasks. Secondly, the tagging of the LINDSEI subcorpora was not primarily carried out to develop meaningful tagging strategies for L2 data, but to test the tagger's initial performance on a small section of data. The challenges encountered are not described in detail and there is no mention of possible solutions for these. Finally, what is of course also very different in the annotation of learner corpora, as opposed to annotating the ELF data in VOICE, is the general perspective adopted. While learner corpora seek to provide a basis for comparing learner language to an assumed NS target, the approach in ELF research is to seek to investigate the language of L2 speakers in their own right. The goals and approach to the two types of corpora are very different, which makes the attempts to tag learner corpora unsuitable as models for the tagging of VOICE. And finally, some documentation on written learner corpora, e.g. by Rastelli (2009) or Meurers and Wunsch (2010) does raise some issues and present possible solutions potentially also relevant to VOICE. However, at the same time, they do not cover all the complex issues involved in tagging natural, plurilingual, interactive and spoken language data as contained in VOICE. In the tagging of VOICE, advice was hence sought primarily from available tagging documentation for spoken L1 corpora (of English and other languages) and the small number of relevant written L2 corpora mentioned above.

3.2.1 Detailing of tagging issues in the literature

Regarding the **first** issue, the **extent to which tagging issues** of spoken language discourse are detected in the literature, one needs to say that for many spoken English L1 corpora, no detailed documentation is available, as has already been discussed. For some corpora which are not publicly available, but where many publications on the data exist, e.g. CANCODE, no documentation is available at all. Also, the literature on a number of corpora for which the existing tagging system CLAWS was adapted to tag different types of spoken data, e.g. teenage language, American English or newer components of the ICE corpora, contains no documentation of the changes in the tagging process. For example, for COLT, MICASE, COCA, and a number of tagged components of ICE, unfortunately, no detailed documentation of the issues encountered in the part-of-speech tagging procedure has been made available. Similarly, the tagging of SEC is based on the written LOB corpus and is described to contain only very minor differences to this (Atwell 1996)[29]. The implication is that for these corpora, either no changes have been made, or that they have not been seen as important enough to be reported. For the tagging of other corpora, e.g. SWITCHBOARD, some documentation has been made available by the LDC (1999) for the spoken part of the Penn Treebank, listing new tags and describing their uses, e.g. for the tag UH. Again a detailed description of e.g. the adaptation from written to spoken medium is not available.

A small number of more detailed reports on the issues encountered for tagging spoken L1 corpora can be found for e.g. the BNC, ICE-GB, NECTE and CHRISTINE. The problems encountered in tagging the spoken part of the BNC are briefly mentioned in Leech, Garside and Bryant (1994a: 624) and are discussed more thoroughly in Garside (1995). The tagging procedures for ICE-GB have also been documented (cf. Greenbaum & Yibin 1994). For NECTE, a very detailed account of the issues encountered in tagging dialectal data is reported (cf. especially Beal et al. 2007, but also Allen et al. 2007: 24f.). Another very useful annotation scheme was developed within the CHRISTINE project, as already mentioned. It discusses, and gives possible solutions for, various aspects of tagging spoken language data, such as how to deal with word fragments, discourse markers, the way spoken language is structured and non-standard constructions (cf. Rahman & Sampson 2000: 300).

As regards non-English L1 corpora, the situation is similar to that of L1 English corpora: Many non-English L1 corpora which have reportedly been part-of-speech tagged (cf. footnote 21) contain no detailed account of the procedures applied. Again, however, there are some

[29] See Atwell et al. (2000) for a description of the AMALGAM project in general within which a number of POS tagging systems were analysed and tools for using these systems in comparison were developed.

cases which do provide a rather detailed account of the problems encountered. Among these are the components of the Nordic Dialect Corpus, the GSLC, the Spoken Dutch Corpus and KidKo. The corpus of spoken Irish briefly mentions the part-of-speech tagging (Uí Dhonnchadha, Frenda & Vaughan 2012: 5), which has, however, not yet been finished. It is hoped that more detail on the tagging of this multilingual data will be provided in future publications.

Overall, the amount of documentation in the tagging literature is therefore rather scarce and not all sufficient as a guideline for the tagging of VOICE. In addition, it emerges that the focus in part-of-speech tagging has – apart from the few mentioned exceptions – been largely on the practicalities of assigning word classes as quickly and conveniently as possible, rather than being held up by descriptive and theoretical problems that arise in the tagging process. This contrasts starkly with the approach adopted in VOICE, where the main concern was how the challenges in tagging ELF usage raise issues about the very nature of the data.

3.2.2 Attitudes expressed towards issues of spoken language

Features of spoken language have often been neglected in the tagging of spoken corpora, with the exception of a few corpus projects which did attempt a more prominent display of features of spoken language in their tagging procedures (see section 3.2.3).

When reviewing the tagging literature, it can be also noticed that the terms used when referring to spoken features tend to be of a derogatory nature. An example for this is the use of the term ‘disfluency’ (or ‘dysfluency’), which is commonly used in the literature, e.g. in Hamaker (1999), Johannessen et al. (2009: 75), Leech (Leech 2000) and Rehbein, Schalowski and Wiese (Rehbein, Schalowski & Wiese 2012a: 32). In tagging literature, the term ‘disfluency’ (or ‘dysfluency’), is often used summarily for structural markers in spoken discourse (see section 3.2.3.2), implies that spoken language is interrupted and that these structural markers disturb the flow of speech and conversation, making tagging a tedious process. For example, Hamaker (1999) writes, “[i]t is these disfluencies, along with the profuse occurrence of monosyllabic words that make SWITCHBOARD an extremely difficult recognition task” and “much more complicated […] due to the variability of conversational speech and a wide range of issues associated with conversational speech such as disfluencies, restarts and repetitions, hesitations, etc.” (ibid.). In NECTE, certain features of spoken language are called “‘cockroaches and pesky critters’” (Beal et al. 2007). Garside (1995: 162) refers to the characteristic of spoken language as “tricks of speech” which have not been transcribed for the BNC. And Stenström and Svartvik (1994: 242) write regarding repeated words, that they “are not accounted for in grammars (which are usually based on written

language) but have a nasty habit of cropping up with some frequency in the impromptu spoken performance of human speakers". The wording in these papers reflects of course the challenges faced when tagging spoken corpora, but it also demonstrates the status of spoken language as opposed to written language. Gilquin and De Cock (2011: 148) describe how the terms for typical markers of speech have often been of a negative nature. Mauranen (2009: 217) writes that the "real-time nature" of speech is something which is often seen as a disadvantage in comparison to written data and that the typical characteristics of speech are often viewed as a necessary evil when working with spoken data:

> Speech tends to get described in terms of deviations from the written language, which at least tacitly sets the standard. So we talk about 'dysfluencies', 'hesitations', 'dislocations' using these and many other negative terms for phenomena that are common features of normal speech (see, for example, Biber et al. 1999: 1066–68). In this way, the temporal character of speech is treated as one of the unavoidable hardships of life, which can account for certain less desirable manifestations of behavior, but is not worth serious attention in itself.

In the tagging literature, the naming of the tags is another area in which in some cases a derogative attitude towards features of spoken language is expressed. In ICE-GB an 'UNTAG' is assigned for incomplete words if "the speaker fails to finish a word but then repeats it in full" (Greenbaum & Yibin 1994: 37). In the TOSCA tagset, which served as a starting point for ICE-GB, spoken markers come under "miscellaneous", subdivided into 'MISC(discourse), 'MISC(foreign)' and 'MISC(interjec)'(Kaszubski 2003). Similarly, the tag 'unclassified' used in BNC refers to a number of categories for spoken language (Garside 1995: 165). As has already been mentioned, many of the corpora elaborated on in section 3.2.3.2 assume relatively broad categories for spoken language (especially in the case of BNC, SWITCHBOARD, and SEC with extremely broad categories), which of course mirrors the (un)importance attributed to the features of spoken language in use. A tendency for a more neutral or positive wording for features of spoken language is displayed by those papers on tagging corpora that also pay great attention to the tagging of the features of spoken language in use: e.g. KidKo, NoTa, CHILDES and the corpus of spoken Irish. Jørgensen (2007: 239), for example, distances himself from the term 'disfluency', referring to "so-called 'disfluencies'". Garside (1995: 163) refers to "laughs and pauses" as "various phenomena", Nøklestad and Søfteland (2007: 246) to "properties", such as "pauses and hesitations" (ibid.) and Rehbein, Schalowski and Wiese (2012a: 29) remark in a neutral manner that "filled pauses, self-corrections and aborted utterances […] pose a challenge for syntactic annotation". However, some inconsistencies can be found even in these publications, as in the same paragraph, Rehbein, Schalowsi and Wiese (2012a: 29) state that the data in KiDKo

"deviat[es] strongly from canonical written German", as 'deviate' can be interpreted as having a negative connotation. Uí Dhonnchadha, Frenda and Vaughan (2012: 2) write of spoken language as "not […] as orderly" as written language, but it being "natural" for typical features of spoken language such as "repetitions, [and] false starts" to occur. A rather neutral mention of spoken features is also made in MacWhinney (2012a: 2376) for CHILDES, calling "interjections, communicators, and dialect forms" "informal forms". Also, the reference to "non-canonical" as opposed to "ungrammatical" structures in Hirschmann, Doolittle and Lüdeling (2007: 1) can be perceived as a favourable treatment of features of spoken language.

It appears an obvious conclusion that the way features of spoken language are perceived, and therefore referred to, by corpus compilers also has an influence on the way tagging categories are assigned and will, at least to some extent, influence how tagging is carried out. In particular, it will have an impact on how spoken features are handled, e.g. whether spoken features are eliminated, put into broad 'dumping' categories or are handled with differentiation and care. Many of these decisions are of course already taken at the transcription stage, which then can limit what is possible in the part-of-speech tagging process. In turn, these decisions determine how the corpus can then be used by researchers. The next section will now elaborate on the handling of different spoken features in the tagging of the corpora discussed.

3.2.3 Issues which have been addressed

Even though the attitude towards features of spoken discourse differs various spoken corpora, some issues concerning the tagging of spoken language have been addressed within the individual tagging procedures. These will be discussed in the following sections. One aspect worth noting at the outset is that the issues which are mentioned are similar across the different types of corpora, e.g. standard or dialect corpora, children or adult speech, which is why they will all be discussed under the same heading. The issues are, as we will see, also relevant for tagging ELF data. In a way, this is not surprising as this is all natural language in use in various shapes and forms. However, because these issues have not at all been dealt with in L2 corpora, or in a way that is not compatible with an ELF approach (as discussed in the section 3.1.4), we are looking at relevant corpora which deal primarily with L1 data.

Five main **groups of issues** relevant to the tagging of VOICE data could be identified in the literature on the tagged corpora described in section 3.1. These are:

1. Differences between written and spoken data
2. Representation of spoken language in the tagging procedures

3. Size of the tagset
4. Non-standard language use
5. Other theoretical considerations in the tagging process

3.2.3.1 Differences between written and spoken data

The **first group** are issues concern the **differences between written and spoken data**. The spoken language is rarely dealt with in its own right, but usually in comparison to written data. This is closely linked to the fact that most taggers were designed for and trained on written data. For the part-of-speech tagging of spoken data, the taggers and descriptions designed for written data are usually used on the spoken data, both for corpora which contain written and spoken components, e.g. CLAWS for the BNC (Leech & Smith 2000), and for spoken-only corpora, e.g. various components of the Nordic Dialect Corpus, e.g. NoTa (Jørgensen 2007), the GSLC (Nivre & Grönqvist 2001), the Corpus of Spoken Icelandic (Rögnvaldsson 2005). The same is true for tagsets where versions developed for written language have often later been adapted to deal with spoken data. For example, Rehbein and Schalowski (2012) describe how they adapted the Stuttgart-Tübingen TagSet (STTS) for the tagging of the spoken language in KidKo. SWITCHBOARD was also based on the tagging guidelines originally designed for the Penn Treebank (Hamaker 1999), SEC was tagged with the tagset first used on LOB (Atwell 2008: 511), and the CHRISTINE tagging scheme was developed from the tagging scheme for the written SUSANNE (Rahman & Sampson 2000).

However, as Nivre and Grönqvist (2001: 55) observe, even though the adaptation of tagging procedures from written data for spoken language has been a common tagging practice, the implications of this have often not been described in detail. One problem which Nivre and Grönqvist (2001: 47) name is the lack of suitable, i.e. tagged and spoken, training material available for the taggers which can be then used for tagging new spoken data. Nivre and Grönqvist (2001: 47) write, “researchers dealing with spoken language often find themselves in the situation of having to tag a corpus with no tagged corpus available with which to train the tagger”. Moreover, for the tagging of the CHRISTINE corpus it was found that categories of written language cannot be applied easily to spoken language and that categories often overlap: “[I]n annotating speech, whose special structural features have had little influence on the analytic tradition, ambiguities of classification constantly arise that cut across traditional category schemes” (Rahman & Sampson 2000: 309). As we found for tagging VOICE, these issues are still relevant over a decade later, as many later corpora are still tagged with taggers trained on written data (Radeka in prep.).

With written and spoken language differing fundamentally, it is not surprising that tagging procedures and guidelines suitable for written corpora are often unsatisfactory when tagging the different and complex phenomenon of spoken language, as they for example, tend to "fail to capture many of the characteristics of spoken language" (Rehbein, Schalowski & Wiese 2012a: 30). When these taggers are used for spoken data, this naturally results in a number of problems which are due to such differences. The following three topics, which can be subsumed under the heading of 'differences between written and spoken data', are regularly addressed in the literature:

The first problem named frequently in tagging is the **different structure** of written and spoken language. As Rehbein, Schalowski and Wiese (2012a: 31) note, "the sentence concept", which is the basic structure in written corpora, cannot be applied to spoken data. Spoken language is better described in terms of utterances as organising elements (2012a: 31). For KidKo, the authors provide a detailed description of the utterance as their choice of a basic structural unit, taking into account also "speech act type and discourse function of the segments" resulting in "the principle of the *smallest possible unit*"[original emphasis] (Rehbein, Schalowski & Wiese 2012a: 32). Nivre and Grönqvist (2001) also used utterances as basic units rather than sentences, as they say that "since grammatical dependencies across utterances are normally rather weak (at least at the level of parts of speech), utterances provide a convenient unit for part-of-speech tagging (and utterance boundaries an important contextual feature)" (Nivre & Grönqvist 2001: 50). The specific structure of spoken language is also addressed as an issue for CHRISTINE. Rahman and Sampson (2000: 304) mention "[i]ll-formed utterances […] in which each element coheres logically with what immediately precedes but the utterance as a whole is not coherent". This raises two issues for the tagging of VOICE, where the utterance is also the basic structuring unit. First, we found that in spoken, interactive, overlapping speech, dependencies extend the boundaries of a single utterance, continuing to other utterances, and hence, solutions had to be found for this (for a discussion see section 4.3.4.3). Secondly, while I would refrain from calling spoken utterances "ill-formed", as the concept of ill-formedness is based on the concept of written sentences and does not easily lend itself to the nature of spoken language, we did encounter the phenomenon of 'incoherent elements', as I would call them, mentioned in Rahman and They will be explained in section 4.3.2.

Another problem which is occasionally addressed is the fact that the tagging has to be done on the **basis of already transcribed data**. While in tagging written data the tags can be directly applied to the original source, spoken data has to be pre-processed, namely, transcribed, in

order to be able to be part-of-speech tagged. These transcriptions then form the basis for the part-of-speech tagging of the data, another interpretative process. In the transcription process, the audio recording of the data is interpreted by the transcriber to represent certain words. The transcriptions then also include passages which could not be clearly identified in the audio recording, and are generally marked as 'unclear' by the transcribers in some way. Such sections can influence the tagging of surrounding tokens (Rahman & Sampson 2000: 308). Another problem, addressed by Gilquin and de Cock (2011: 147), is that characteristics of spoken discourse, such as pragmatic markers and hesitations, as well as non-standard language use, tend to be ignored by the transcribers (cf. Lindsay & O'Connell 1995). Gilquin and de Cock (2011: 148) also mention a tendency of a negative attitude towards spoken language phenomena such as hesitations. It seems fair to argue, therefore, that the attitude towards typical characteristics of spoken language may influence the way these are rendered in the transcription. This is to say that if these characteristics are "dismissed as 'messy' coding details" (MacWhinney 2009: 54), they are more likely to be ignored by the transcribers, if they are seen as essential and characteristic, they are more likely to be considered systematically in the transcription. In VOICE, the transcribers were especially trained to pay attention to the spoken features of the audio data (Breiteneder et al. 2006: 172). These were considered as important from the beginning. As a result, various spoken features were systematically captured in the VOICE transcription scheme, which is especially rich with regard to these features.

A **third issue** often addressed is that the spoken data differ from written training material in terms of distribution of categories for individual tokens. Gibbon, Mertins and Moore (2000: 27) write:

> the fact that the same tagset can be applied to spoken and written data should not lead us to ignore the fact that frequency and importance of word categories vary widely across the two varieties of data, or that at a given level of generality tags may be applicable to related but different categories.

This creates problems for the estimation of probabilities of certain word classes for spoken data, which cannot be done on the basis of the probabilities for written data, as they would otherwise result in a lower tagging accuracy and wrong tag assignment. As a way around this, Nivre and Grönqvist (2001: 60) estimated new probabilities for their spoken Swedish data. The problem of certain tokens, for example "I, well, and right" (1994a: 624), being distributed differently for writing and speech, and the resulting cause of tagging mistakes is also mentioned for the tagging of the BNC. As a solution to this, the CLAWS4 resources were adapted, with the help of an additional lexicon and an "idiomlist" (ibid.). Of course, not only

is the distribution of word class categories different in spoken language in use, but spoken discourse also reveals aspects which do not occur in written language at all, as will be discussed below.

Despite these issues, some good results have been gained through the adaptation of taggers trained on written language to spoken data: e.g. Nivre and Grönqvist (2001) report good results for tagging the GSLC by extension of the tagset for typical markers of spoken language, e.g. "feedback" and "own communication management" (2001: 56). Often, however, these results are achieved by "normalisation" of the spoken language, meaning that **spoken** data is **made more similar** to **written** data for the part-of-speech tagging process. For example, Rögnvaldsson (2005) reports a tagging accuracy of 92.5% for the Corpus of Spoken Icelandic, which was "much better than we expected, and […] as least as high as for the written texts."[30] This was achieved by tagging a spoken corpus with a tagger trained on written Icelandic. The data, however, had been normalised before, for example by excluding overlaps and interrupted speech (Rögnvaldsson 2005).

3.2.3.2 Representation of spoken language in the tagging procedures

The **second group of issues** addressed in the literature concerns the tagging of features typical of spoken language. During the tagging of VOICE, we came across a number of characteristics of spoken discourse and it was of interest to us how these had been dealt with in previous tagging procedures. Such issues were, for example, discourse markers, filled and unfilled pauses, response markers, truncated words, repetitions, onomatopoeia, formulaic expressions, multi-word expressions, various non-verbal items, such as laughter, silence or significant use of breath as a non-verbal expression of meaning, words which occur in speech but not so much in writing, e.g. "innit", "gonna", "wanna", code-mixing and switching, foreign language use, etc. As has already been mentioned in section 2.3, part-of-speech tagging in general goes well beyond the traditional eight word classes generally described for English and includes various other morphosyntactic aspects, but also sometimes markers of pragmatic functions. However, in general, such tagging descriptions of spoken features are rarely found, as has also been observed by others, e.g. MacWhinney (2012a: 2375), who notes, "[t]he lexicons used by the taggers focused primarily on written, rather than spoken speech, seldom including methods for tagging interjections, onomatopoeia, babbling, code mixing, and many other forms in natural spoken language."

[30] No page numbers indicated in pdf.

In general, there are three different strategies of how to deal with features of spoken language, as listed by Johannessen and Jørgensen (2006: 11). They summarise that the structuring elements of spoken language have either been completely disregarded, partly disregarded, or taken seriously in dealing with spoken data. For the annotation of SWITCHBOARD, a bracketing procedure was developed, which, as Johannessen and Jørgensen (ibid.) report, ignored all features of spoken language structure. Hamaker (1999)[31] tagged a subcorpus of SWITCHBOARD sized 11 000 words, with the aim to "provide robust tagging procedures" for this type of data, and deleted typical markers of spoken language for the tagging process, writing, "[t]he tags marking silence and noise were removed from the data in order to preserve what little grammatical structure there is." Similarly, for the BNC "unplanned repetitions" were ignored for the tagging process (Leech, Garside & Bryant 1994a: 624). In a few selected, newer corpus annotation projects, a conscious effort was made to represent the spoken features of the data in the tagset as they had been captured in the transcription (cf. e.g. Johannessen & Jørgensen 2006, Rehbein & Schalowski 2012). Below, a more detailed account of the treatment of spoken features by various tagging strategies will be given and discussed.

As mentioned above, aspects of spoken language naturally also become an issue in the comparison of spoken with written data, as the characteristics of spoken data do not occur to the same extent in written data, and therefore the taggers do not have suitable strategies for dealing with these, and spoken features tend to be underrepresented in taggers lexicons (MacWhinney 2012a: 2357). Often what can be found is not just a lack of information on features of spoken language, but also that they are referred to in rather broad categories that are used inconsistently. As already mentioned in the previous section, a frequently used term for features of spoken language is 'disfluency'. This term vividly illustrates the lack of consistency when spoken features are addressed in the tagging literature. 'Disfluency' seems to cover a broad range encompassing all sorts of spoken phenomena. Stenström and Svartvik (1994: 242) list the following "kinds of non-fluency" in spoken language: pauses (silent and filled), repeats, false starts, corrections, interjection, stutters and slips of the tongue. While Hamaker distinguishes the spoken features "disfluencies, restarts and repetitions, hesitations" in the quote above, Gibbon, Mertins and Moore (2000: 27) name "pause fillers [and] word fragments" as "dysfluency phenomena" and suggest separate categories for, e.g. "hesitator" and "word fragment" (2000: 28), rather than applying broader umbrella categories for these features and propose not considering these features in the part-of-speech tagging procedure at

[31] No page numbers indicated in pdf.

all, “but to mark them in the orthographic transcription as non-word vocalisations comparable to laughs and snorts” (ibid.). According to them, this is in line with considering “dysfluency phenomena as extraneous to the grammatical annotation of speech, on the assumption that they belong to a distinct level of dialog control” (ibid.).

Because the spoken features are often written about in general terms and not differentiated in the tagging literature, it is difficult to establish how individual aspects of spoken language are dealt with in tagging, e.g. hesitations as opposed to response markers. Moreover, it is virtually impossible to compare different methods for dealing with individual features, as the detail of description is usually not given and terms such as ‘dysfluency’ (or ‘disfluency’) are used rather loosely. Looking at the literature more closely, it becomes clear that in fact, in many cases, no differentiation has been made between individual spoken features but rather, that they are dealt with quite summarily. Spoken features are often not given individual tags but one uniform tag, or a small number of different tags for all the features which can be regarded as peculiar to spoken language. There are only a few exceptions, especially among some very recent corpora.

In early corpora, e.g. the **BNC**, the tagset C5, which was used for tagging the spoken (as well as the written) part of the BNC, only includes two tags for spoken phenomena: one tag “ITJ” called “interjections”, which is used for tokens such as “*oh, yes, mhm, wow*”[original emphasis] (Leech 1997) and a tag “UNC”, standing for “unclassified” and referring to “items which are not appropriately classified as items of the English lexicon” (ibid.). This includes a number of spoken items, i.e. “(non-English) words, special typographical symbols, formulae, and (in spoken language) hesitation fillers such as *er* and *erm*” [original emphasis] (Leech 1997), as well as “truncated words” (Garside 1995: 165) The C7 Tagset, according to which 1 million written and 1 million words spoken data of the BNC were tagged, only includes the tag “UH” for interjection but no tag for “Unclassified” (Leech 1997).

The case is similar for **SWITCHBOARD**. Even though the Part-of-speech tagging Guidelines for the Penn Treebank Project (henceforth: Penn Guidelines) were extended for SWITCHBOARD, broad categories for markers of spoken discourse remain. The tag referring to ‘interjections’ (UH) encompasses a wide range of phenomena, e.g. “fillers” such as “uh-huh, uh, um, […] oh”, “[e]xclamations” such as “wow, boy, […], bam”, response items such as “yeah, […] nope”, “greetings and adieus”, such as “bye-bye, bye, hello”, “[c]ontinuers and assessors”, such as “exactly, really, […], right, yeah, sure”, “[d]iscourse markers” including “well, actually, see […], sure, like […], so” (Linguistic Data Consortium 1999: 2). Merely one tag “XX” for “partial words” was added (Linguistic Data Consortium 1999: 1). Other

than that, guidelines for annotating repeated words are given, i.e. to annotate identical, repeated words with the same tag (Linguistic Data Consortium 1999: 2) and a policy for handling typing errors is introduced (Linguistic Data Consortium 1999: 2).

To give an example of another early corpus, **SEC**, which was tagged with CLAWS1 and CLAWS2, only contains a single tag for "interjection[s]", "UH" used for e.g. "oh, yes, um ..." (UCREL 1993b-2014), to represent spoken phenomena. The CLAWS2 tagset additionally contains so-called 'ditto tags', which "signif[y] that the tag occurs as part of a sequence of similar tags, representing a sequence of words which for grammatical purposes are treated as a single unit" (UCREL 1993c-2014), which are, however, not specifically geared at spoken phenomena. However, what needs to be added is that many features of spoken discourse were already absent in the transcriptions of some of these early corpora, e.g. in the BNC (Garside 1995: 172), or removed before the tagging process, as was done for SEC (Taylor & Knowles 1988). It could be suspected that this might be one of the reasons why the tagging of BNC was described as unproblematic (Garside 1996: 172) as features which are not present do not need to be considered in, and do not cause problems for, the tagging categorisation process.

Some corpora which were compiled later contain some more differentiations, though the categories remain relatively broad and tags for e.g. discoursal features, such as pauses or various noises, e.g. laughter, are often absent. In **ICE-GB**, for example, spoken markers are represented by three different tags: "INTERJEC" ("interjection") (Nelson 2005: 13), "REACT" ("reaction signal"), which includes response markers, e.g. "*no*, […] *yes*" [original emphasis], but also some words which could be classified as discourse markers, e.g. "*all right*, *fine*, *good*" [original emphasis] (Nelson 2005: 28), and "FRM" ("formulaic expression") (Nelson 2005: 13). Additionally, there is a tag called "UNTAG" (Greenbaum & Yibin 1994: 37)[32], which is used for "word partials" (ibid.), ambiguities and tokens "that comprised two grammatical words", e.g. "dunno […], innit […] wanna" (Greenbaum & Yibin 1994: 38).

Similarly, the tagset of the Swedish **GSLC**, which is an extended version of the Stockholm-Umeå Corpus, contains four different tags for spoken markers. The original tagset of the Stockholm-Umeå Corpus contains a tag for interjections (IN), which was also used for the GSLC (Nivre & Grönqvist 2001: 77), and one for "verbal particles" (PL) which was added later (Gustafson-Capková & Hartmann 2006: 19). For the GSLC, tags for "feedback" (fb) and "own communication management" (ocm), e.g. hesitations, were also added (Nivre & Grönqvist 2001: 56f.). Foreign words are tagged according to their part-of-speech rather than

[32] No explanation as to what the letters in 'UNTAG' specifically stand for could be detected.

receiving a uniform tag (Nivre & Grönqvist 2001: 57). The **Spoken Dutch Corpus** uses a small tagset with additional morphosyntactic information (van Eynde, Zavrel & Daelemans 2000) including only one tag for 'interjection'. Other spoken features (incomplete, unintelligible) or code-switching (foreign) are also marked, but not with a POS tag (van Eynde, Zavrel & Daelemans 2000). **NECTE** and **MICASE** both used the CLAWS tagger with the C8 and C8++ tagset, repectively. The tagset contains tags as "FU [for] unclassified word[s]", "FW [for] foreign word[s]" and "UH [for] interjection[s]" (UCREL 1995-2004). For both corpora, the CLAWS tagset was changed slightly. For NECTE, the lack of a tag for discourse markers was addressed, and as a solution, the category UH was broadened to include not only interjections but also discourse markers, e.g. "*wey*, *like*, *aye*, *ah*, *well*, *uhhuh*, *huh*" [original emphasis] (Beal et al. 2007). Some multi-word discourse markers, such as *you know* co-occurring with others, such as *like*, were however not always given an UH tag (ibid.). Beal at al. (2007) explain that "[i]n this case the fragment [!VVI] ('not an infinitive') blocks *you know* from being treated as a discourse marker if CLAWS has already identified *know* as an infinitive, as typically happens in the context of questions formed with *do*, e.g. *Do you know...?*"[original emphasis] This was also the case with "*I mean* and *dear me how*" [original emphasis], which were also not tagged UH but with traditional tags. The tag "FU" in NECTE referred only to incomplete ("truncated") words (Beal et al. 2007). Additionally, the CLAWS lexicon and idiom list was enhanced with words and expressions which did not occur in CLAWS but in the NECTE data. MICASE added three additional tags which sub-divide the category interjections (UH), as indicated below:

> [The following three tags are unique to MICASE:]
> UHY interjection, positive (e.g. uh-huh, yuh-huh)
> UHN interjection, negative (e.g. uh-uh, huh-uh, m-m, hm-m, nuh-uh)
> UHE interjection, exclamatory (e.g. uh-oh, oh-oh, oops)
>
> (Simpson, Lee & Leicher 2007: 39)

Although the tagsets for NECTE and MICASE have enhanced the C8 tagset slightly, the annotation does not include many other spoken features such as pauses, speaker noises, unintelligible speech or onomatopoeia. Many of these were transcribed in NECTE but ignored for the tagging process (Beal et al. 2007). MICASE also does not include a tag for pragmatic markers either.

CHILDES has similarly broad categories for spoken markers. These are, however, described in detail and considered as important to be included in the part-of-speech tagging (MacWhinney 2012a: 2376). The tagset includes a tag "CO" for "Communicator" (MacWhinney 2012b: 112), which is a collective category

> for interactive and communicative forms which fulfil a variety of functions in speech and conversation. Many of these are formulaic expressions such as **hello, good-morning, good-bye, please, thank-you**. Also included in this category are words used to express emotion, as well as imitative and onomatopeic forms, such as **ah, aw, boom, boom-boom, icky, wow, yuck,** and **yummy** [original emphasis]
>
> (MacWhinney 2012b: 113)

The tagset also includes a tag for "[f]illers" ("FIL") (MacWhinney 2012c: 178), which is used for hesitation markers such as "um, uh, er" (ibid.). An important guideline in the tagging of CHILDES was the consistent annotation of typical features of spoken language. For the tagging scheme, great effort was made to adequately represent the data. It is stressed that "it is important to include consistent representations for informal forms such as interjections, communicators, and dialect forms" (MacWhinney 2012a: 2376). For example, "monomorphemic interjections such as 'amen'", are differentiated from "compound interjections such as 'good morning'" (MacWhinney 2012a: 2376). The lexicon includes various lists of peculiarities of the data, e.g. items for fillers, vocative/regular communicators, rhymes as interjections, loan words, various types of compounds and baby-talk items, onomatopoeia and "omitted words" (MacWhinney 2012b: 163). Moreover, a special annotation scheme for different types of compounds is implemented (MacWhinney 2012a: 2376) and the tagging of "word combinations" (MacWhinney 2012a: 2376) (otherwise called "ditto-tags", or "multi-word items" in VOICE) improved tagger accuracy.

Some corpus compilation projects, most of them rather recently compiled, critique these annotation schemes for markers of spoken language with the broad categories for typical aspects of spoken language (e.g. Rehbein & Schalowski 2012: 239), and introduce very detailed annotation schemes for markers of spoken language. In this context, the Scandinavian corpus **NoTa**, the German corpus **KidKo,** the **Corpus of spoken Irish** and **CHRISTINE** are especially noteworthy. **NoTa** includes a detailed transcription of many spoken features, which can all be searched for in the online interface of the Nordic dialect corpus, as Johanessen (2007: 30) writes, "[a]ll transcriptions of the speech occurring in the corpus are searchable, as are the specially annotated events such as laughter and coughing, plus a variety of interjections and exclamations, extralinguistic noises etc.". The part-of-speech tags for spoken language features include: "interj" for "interjection[s]","pause" and "pause2" for two different types of pauses, "clb" for "clause boundar[ies]", "anf" for "quote mark[s]", "ukjent" for "unknown word[s]", tags such as "engelsk" and "latin" for code-switches (Jørgensen 2007: 3), as well as "meaningful sounds" (Johannessen et al. 2007: 32), including e.g. yawns, laughing and breath which are "annotated in the corpus as tags" (ibid.). In the Nordic dialect online interface, one can see that these speaker noises, e.g. breathing, laughter, coughing,

onomatopoeia are not tagged with POS-tags but can be found and searched for under the 'nlex'-category.

The tagset of **KiDKo** is an extension of the STTS, which already included a tag for interjections (cf. Schiller, Teufel & Stöckert 1999). Rehbein, Schalowski and Wiese (2012a: 32) write that 8 tags were added for the POS tagging of KidKo. They write, "[o]ur extended version comprises additional tags for filled pauses, question particles, backchannel signals, onomatopoeia, unspecific particles, breaks, uninterpretable material, and a new punctuation tag to mark unfinished sentences" (Rehbein, Schalowski & Wiese 2012a: 32) and mention that their "part-of-speech annotation of interjections, discourse markers and fillers is also more fine-grained than the one in the SWITCHBOARD corpus which combines them all into one tag (INTJ)". In their article from 2013, they report that, in fact, 11 new tags were added (see Table 1). The authors thus indicate that the detailed representation of spoken markers was prioritised in the tagging of KidKo. The extension of the KiDKo tagset with various tags has been discussed in Rehbein and Schalowski (2013).

Table 1: Additional POS tags for spoken language data in KiDKO. Source: Rehbein & Schalowski (2013: 202). Reproduced with permission.

POS	description	example	literal translation
PAUSE	*pause, silent*	so ein **(-)** Rapper	such a (-) rapper
PTKFILL	*particle, filler*	ich **äh** ich komme auch .	I er I come too
PTKINI	*particle in utterance-initial position*	**ja** kommst du denn auch ?	PART come you then too
PTKRZ	*backchannel signal*	A: ich komme auch . B: **hm-hm** .	A: I come too B: uh-huh
PTKQU	*question particle*	du kommst auch . **Ne** ?	you come too . no ?
PTKONO	*onomatopoeia*	das Lied ging so **lalala** .	the song went so lalala
PTKPH	*placeholder particle*	er hat **dings** hier .	he has thingy here
VVNI	*uninflected verb*	seufz	sigh
XYB	*unfinished word, interruption*	ich **ko** #	I co #
XYU	*uninterpretable*	**(unverständlich)** #	(uninterpretable) #
$#	*unfinished utterance*	ich ko #	I co #

The success of the part-of-speech tagging of the **corpus of spoken Irish** has not yet been determined. However, Uí Dhonnchadha , Frenda and Vaughan (2012: 4f.) describe that markers of spoken discourse were considered in the tagging procedure, e.g. "filled pauses (em, eheh etc, [sic]) […] codes for non-verbal events (coughs, laughs, sneezes etc.), phonetic fragments (*b- b- bosca* 'b- b- box') and indecipherable material (xxx)" as well as "[d]ialectal variants". Regarding multiword items they write that "fixed phrases", such as "you know, so, just etc." (Uí Dhonnchadha, Frenda & Vaughan 2012: 3) were standardised for the transcription process. Although nothing is said about how they are going to be dealt with in

tagging, it seems reasonable to assume that the transcription conventions of such multi-word items will also be incorporated in the tagging process. Strategies for considering code-switching from Irish into English might also be considered at a later stage (Uí Dhonnchadha, Frenda & Vaughan 2012: 5), as "[s]poken transcripts contain more English words than would be found in written Irish". So even though the part-of-speech tagging of the Irish corpus is still work-in-progress, the detailed consideration of many characteristics of spoken language is promising.

Finally, the tagging scheme for **CHRISTINE** should also be mentioned. This was published a number of years before the tagging descriptions for the other three non-English corpora described above. Is quite different in nature from most papers on tagging published around the same time, as it seeks to "examine a number of conceptual problems in defining rigorous annotation standards for spontaneous speech" (Rahman & Sampson 2000: 300). As language is "one of the most complex phenomena dealt with by any branch of IT" they consider "consciousness-raising about problematic aspects of the subject-matter […] a high priority" (Rahman & Sampson 2000: 301). The detail of the tagging description mirrors this theoretical position. As mentioned in section 3.1.2, CHRISTINE is an extension of the tagging scheme developed for the written data in SUSANNE. In contrast to SUSANNE, CHRISTINE includes tags for swearing, "anonymized name[s]", "hail and farewell", "non-linguistic sounds: e.g. *ding ding dee dee*" [original emphasis], "silent pauses", "nasal and vocalic filled pauses", "non-linguistic vocalizations", "vocal shifts", "inaudible wording" etc. (Sampson 2000). Furthermore, "slashtags" are used as a strategy to deal with incomplete words for which the presumed intended word can be identified, and a detailed annotation scheme for speech repairs is proposed, by which e.g. the beginnings and ends of speech repairs are identified (Rahman & Sampson 2000: 302ff.). Discourse markers are given special prominence and are divided into "Expletives (e.g. *gosh*), Response (e.g. *ah*), and Imitated Noise (e.g. *glug glug*)" (Rahman & Sampson 2000: 302), reporting however, that it was not always found to be straightforward to distinguish the three in categorisation, as "discourse items are not normally syntactically integrated into wider structures" (ibid.). However, discourse markers with homonyms in other word class categories, such as *well*, and their classification are not mentioned, even though these are clearly not contained in the three sub-specifications. Many other issues, such as how to deal with non-standard language and how to determine whether an utterance is dialectal or a "performance error" (Rahman & Sampson 2000: 307) are also discussed for the part-of-speech annotation of CHRISTINE, and we will come back to this in subsequent chapters of this thesis.

In general, it can be said that this annotation scheme, though fairly complex in parts to the point where it might be impractical (such as the guidelines for annotating speech annotation repairs), can not only point researchers designing an annotation scheme for a spoken corpus to issues they will almost certainly encounter, but it also gives solutions to these problems. These can certainly be used as a starting point for the annotation of spoken corpora.

As a final issue, it is worth noting that in tagging procedures, the removal of certain characteristics of spoken language data which add to the structuring of spoken discourse is not necessarily beneficial to the tagging result. For the tagging of NoTa, Nøklestad and Søfteland (2007) found that the removal of pauses and/or hesitations decreased tagging accuracy, rather than adding to it. The authors conclude that "[t]he indication that pauses in particular may provide important grammatical information supports the finding by Strangert (1993) that pauses tend to occur at positions that are relevant to the underlying message, including syntactic boundaries" (2007: 247). Rehbein, Schalowski and Wiese (2012a: 35) also find it worthwhile to consider spoken features in the tagging procedure, even if this involves a greater workload. The authors (ibid.) "strongly argue for including disfluencies (which are often, but not exclusively caused by self-repairs) in the corpus, even if this will result in more manual effort for correcting automatically predicted POS tags during preprocessing."

The review of the tagging literature suggests several things with regard to the dealing with spoken features. Firstly, there seems to be a tendency that spoken corpora which were published earlier, have broader categories for dealing with spoken phenomena and that more recent corpora, in turn, introduce more detailed distinctions. While the BNC, SWITCHBOARD, and SEC only contain one or two tags for all spoken features, later corpora such as ICE-GB, GSLC, The Spoken Dutch Corpus, NECTE, MICASE and CHILDES use slightly more fine-grained categories. However, in the majority of these tagging descriptions, important aspects of spoken language, such as pragmatic markers (an exception being NECTE) or non-verbal markers are still not addressed. In the most recent corpora discussed, i.e. NoTa, KiDKo and the Corpus of spoken Irish, as well as the earlier corpus CHRISTINE, markers of spoken language are given considerably more prominence, with up to eleven individual tags (KidKo). This tendency mirrors what Gilquin and De Cock (2011: 152) describe in terms of a change in linguistics, arguing that spoken features are increasingly represented in dictionaries and grammars and are increasingly investigated by researchers. As spoken features are increasingly recognised as legitimate subjects of research, they are also better represented and described in tagging systems, as the recent corpus projects demonstrate. This is also encouraging for linguists working with spoken language, as it is

rather problematic, if not impossible, to investigate features of spoken language in corpora which contain large 'dumping' categories for these or for which many spoken features have not even been included in the transcriptions. Moreover, an inclusion of features of spoken discourse which fulfil the task of structuring the discourse can be beneficial in the tagging process.

However, there is still much room for improvement. Some features of spoken language still do not receive much attention, such as discourse markers with homonyms in other word-class categories, such as *like, so, well, you know* etc. Moreover, the tagging of spoken features is rather heterogeneous in the different tagging systems, and the detail of the tagging is often still not described to an extent which makes it a useful resource for new corpus projects which seek to annotate their data with parts of speech. A more detailed documentation would mean that common tagging schemata (such as has been attempted in, e.g. the EAGLES guidelines) could be developed for spoken, multilingual corpora and thus save much time and effort in the tagging of spoken corpora and, in addition, create more consistency and comparability for the annotation of spoken phenomena in tagging procedures.

The remaining sections of this chapter will discuss three other aspects which were found to be a challenge in the tagging of VOICE, namely uncertain tag assignment, non-standard language use, variation in spoken language and the theoretical considerations arising from the process of annotating transcriptions of spoken discourse. The aim is to review how other spoken corpora dealt with these aspects. In many cases, this will, however, be difficult, if not impossible, as there are merely a few papers in which these aspects are treated.

3.2.3.3 Uncertain tag assignment and tagset size

The first of these issues, uncertainties in tag assignment, has also been a point of discussion in some of the literature. It is closely related to the issue of the extension of tagsets with tags for spoken data, as ambiguities in the data can be highlighted or reduced depending on the size of the tagset. Although much more could be said about issues concerning the selection and size of tagsets, only a few points relevant to this thesis are singled out below.

The size of tagsets in the spoken corpora discussed in this thesis varies from rather small, as in, e.g. the Penn Treebank, to rather large tagsets, containing a few hundred tags, as e.g. the Spoken Dutch Corpus or the Corpus of Spoken Icelandic. The number of categories used for a tagset of spoken language can simplify or complicate tagging decisions, and is often related to the desired tagging outcome. For example, in the tagging of the Spoken Dutch Corpus, which uses a total of 300 tags (van Eynde, Zavrel & Daelemans 2000), tagging challenges were

attributed to the fine-grainedness of the tagset. The authors suggest that "[t]agsets with a high degree of granularity, […] such as the one of Spoken Dutch Corpus, are much more demanding, both conceptually and computationally" (ibid.). For both written and spoken parts of the BNC, the tagset was reduced from an earlier version in order to facilitate the tagging process (Leech, Garside & Bryant 1994b: 51). This was done to reduce the number of possible tag ambiguities and thus to "improve the success rate of the automatic tagging" (Leech, Garside & Bryant 1994b: 53).

It is also fairly common to carry out tagging experiments of spoken corpora with two different tagsets, one smaller and one larger, in order to determine the effect on the tagging outcome, as for example Hamaker (1999) did for SWITCHBOARD, and as it was done for the BNC, KidKo, GSLC and NoTa. Hamaker (1999) found, "that a smaller tag set seems to be the avenue to explore in further research since this allows for the irregularities common in spontaneous speech", and achieved up to 90.18% accuracy when tagging a part of SWICHBOARD. Similarly, for the GSLC, a tagset with the 23 part-of-speech tags from the Stockholm-Umeå Corpus was used, including two new tags (Nivre & Grönqvist 2001: 56) (cf. section 3.2.3.2) but also a more coarse-grained tagset of 11 tags for the application of creating a dictionary, "corresponding to the traditional parts-of-speech found in grammar books, extended with fb and ocm" (Nivre & Grönqvist 2001: 56). They found that this smaller tagset yielded better results than the larger tagset, i.e. " 95.29% (±0.43%) (large tagset) and 97.44% (±0.32%) (small tagset)." (Nivre & Grönqvist 2001: 67). Finally, Jørgensen (2007: 2) also reports that the experiments described in his 2007 paper were carried out with a larger tagset of 302, as well as with a smaller tagset without "morphological information" (2007: 2) and slightly better tagging accuracy was achieved with the smaller tagset (2007: 4). In particular, the broader categories necessarily used in more reduced tagsets have been found to be beneficial to the tagging accuracy for spoken data. Nivre and Grönqvist (Nivre & Grönqvist 2001: 72) subsume determiners and pronouns ('pron') under one pronoun tag and wh-adverbs in one category with other adverbs ('adv'), and report to "simply observe that the generally higher recall/precision figures (as compared to the large tagset) are due to the fact that several of the difficult tag pairs discussed above are simply merged in the smaller tagset" (ibid.). Thus, smaller tagsets constitute a possibility for dealing with the uncertainty of tag assignment that arises from the character and variability of spoken language in use.

In general, there are three other options to deal with the high number of tags which cannot be assigned securely in spoken, interactive language data, which are also related to tagset size

and will be discussed below. These are the use of an all-incorporating **'dumping' tag**, as it may be called, as well as that of **generic** or **ambiguous tags**.

Firstly, many corpora have a **'dumping' tag** for categories which cannot be easily assigned, such as the UNTAG or '?' in ICE-GB (Greenbaum & Yibin 1994: 37f., Greenbaum & Nelson 1996: 34), or a tag labelled 'unclassified' in BNC (Leech & Smith 2000, subpage: Guidelines to word class tagging). Rehbein, Schalowski and Wiese (2012b) introduced the tag "XYU" for items which were unintelligible due to e.g. the quality of the audio material, or due to other languages being used ('code-switching'), as well as for "Nichtworte" [transl. 'non-words'], e.g. "Ndrrisch!". Another option, which needs to be distinguished from the broad categories described above, is the use of **generic tags**, or, as Rahman and Sampson (2000: 309) phrase it, "a neutral fallback annotation", as is suggested in the CHRISTINE annotation scheme. However, Rahman and Sampson (2000: 309) point out that this might not be as easily applicable to spoken language as it is to written, as in spoken language it is not often clear what the appropriate annotation is: "[I]t is often very difficult to define fallback notations which enable the annotator to avoid attributing properties for which there is no evidence, while allowing what can safely be said to be expressed" (ibid.). Finally, for the NoTa corpus, **ambiguous tags** were used to deal with uncertain tag assignment, i.e. cases which were "undecided by the human annotator" (Jørgensen 2007: 2). They introduced 26 different tags which are ambiguous between two or three word-class categories, e.g. "konj/prep/adv, subst/adj" (Jørgensen 2007: 3) and are regular components of the tagset (Jørgensen 2007: 4).

However, small tagsets, generic and ambiguous tags are not the only possibilities for dealing with the uncertainties of categorising spoken data, as some corpus projects demonstrate that good tagging results can also be achieved by using more fine-grained tagsets. A test of tagger performance of TreeTagger on KiDKo data with a tagset of 65 components (extended with the tags for spoken features) gained accuracies of around 90% (Rehbein, Schalowski & Wiese 2012b). Also for CHILDES, a slightly larger tagset was used than for e.g. GSLC, consisting of 31 tags (Sagae et al. 2007: 26), gaining "98% accuracy on adult corpora and 97% accuracy on child language corpora" with MOR (MacWhinney 2012a: 2375). The most striking example, however, is probably the Corpus of Spoken Icelandic, for which a very fine-grained tagset was used "[d]ue to the inflectional character of Icelandic […] containing 639 different tags in all." (Rögnvaldsson 2005). A very satisfactory results for tagging the corpus with this tagset is reported:

> The results of the tagging were a pleasant surprise. On the first pass, prior to any simplification of tags etc., the accuracy for the spoken language corpus was around 92.5%, which means that it was considerably higher than for the written texts that the tagger was trained on.
>
> (Rögnvaldsson 2005)

However, as already reported, the data was normalised before the tagging process by removing typical markers of spoken language, which may have influenced the tagging accuracy. For NoTa, it is remarkable that the much larger tagset only resulted in slightly lower tagging accuracies and that, indeed, the removal of spoken markers decreased tagging accuracy for both the larger and the smaller tagset (cf. Jørgensen 2007: 4, Table 4: Results after Filtering). Finally, the Spoken Dutch Corpus, which uses a rather large tagset comprising 300 categories in total (Oostdijk 2000)[33] achieved "a final accuracy of 94.3%" when using a combination of different taggers (van Eynde, Zavrel & Daelemans 2000)[34].

What can be said in conclusion is that there are different strategies which have been attempted in order to deal with the uncertainty of tag assignment which spoken, variable language sometimes entails. These tactics render the tagging faster, and in some cases, more feasible to carry out, however, in general also come with the inherent risk of simplifying the data at hand by not displaying a large proportion of characteristics of natural, spoken language. For example, smaller tagsets have often shown slightly better results than larger tagsets. However, it has also been demonstrated that the inclusion of spoken markers in the tagging process can be beneficial. Moreover, the conflicting results indicate that the tagging accuracy is dependent on many aspects other than the tagset (such as the different taggers used, the different features represented in the original transcription), and that these are also factors to be taken into account when evaluating tagging accuracies of spoken language.

3.2.3.4 Non-standard language use

Under this heading, two aspects of non-standard language use will be elaborated on: Firstly, the tagging of non-standard forms and secondly, the tagging of standard forms with non-standard functions. Both aspects were major issues of discussion in the tagging of VOICE, as they were encountered frequently.

In the tagging literature, what is classified as non-standard or non-canonical often does not refer to what is conventionally understood by the term, e.g. what is acceptable to a group of speakers or what can be found in a dictionary of a language. It sometimes refers to items which are not known by the tagger or to "word forms not occurring in the training corpus"

[33] No page numbers in pdf.
[34] No page numbers in pdf.

(Nivre & Grönqvist 2001: 59) which are sometimes referred as "unknown words" (ibid.). Rehbein and Schalowski (2012: 238) mention "non-canonical" forms, which in corpus annotation often is meant to describe differences of "all kinds of language data which deviate from standard written text", hence also including phenomena in the term which may be accepted and even standardised for spoken language. The idea that the points of reference for spoken language may often differ to these of written language is in line with Carter and McCarthy (2006a: 168), who argue that "[w]hat may be considered 'non-standard' in writing may well be 'standard' in speech". By the term 'standard', it needs to be noted, the authors mean "a description of the recurrent spoken usage of adult native speakers" (ibid.), i.e. something which is conventionally acceptable or canonical. A so-called data-driven view is proposed by Hirschmann, Doolittle and Lüdeling (2007: 1) who state that,

> '[n]on-canonical' in this paper refers to structures that cannot be described or generated by a given linguistic framework – canonicity can only be defined with respect to that framework. A structure may be non-canonical because it is ungrammatical, or it may be non-canonical because the given framework is not able to analyse it.

This definition of canonicity is therefore only valid for a particular framework or dataset. However, Rahman and Sampson (2000: 307) observe that sometimes, it is not even a straightforward enterprise to decide whether an expression is "well-formed with respect to the speaker's nonstandard dialect" or whether the speaker is omitting something, e.g. "*she was shouting him", "I shout you for tea" or "I was shouted*" [original emphasis] (Rahman & Sampson 2000: 307). Even though the authors regard it as crucial to keep these two issues apart in annotation, "analysts will often in practice have no basis for applying the distinction to particular examples." (Rahman & Sampson 2000: 308). This is something which is rarely discussed in descriptions of other corpora, but is certainly an issue in ELF corpora, where it is often unclear what an appropriate point of reference is. For CHRISTINE, the solution to the problem was that once similar occurrences by other speakers in the corpus were detected, the conclusion was drawn that this was a legitimate use of the items, i.e. in the above case for "shout to have a regular transitive use in nonstandard English" (Rahman & Sampson 2000: 307). As discussed in chapter 3.1.5, the way the nonstandard forms are dealt with in CHRISTINE is different to most L2 corpora, where deviations from a (usually written) standard are generally described in terms of over- and under-usage or errors. Hirschmann, Doolittle and Lüdeling (2007: 14) describe these different approaches rather explicitly,

> [t]he same general scheme can be used for different varieties. The interpretation of the deviations from the canonical structure is a further step that depends on the variety at

> hand and on the research question. In learner language, a deviation might be analysed as an error, in other varieties it might be analysed as a feature.

This quote demonstrates that NNS use of non-canonical structures are often equated with erroneous language use, while for NS use, this is accepted as "a feature", as Hirschmann, Doolittle and Lüdeling (ibid.) argue. However, the important theoretical question as to whether this is a justified way of approaching the issue is not taken up by the authors, and indeed is generally not discussed in the tagging literature.

Let us now turn to the strategies for handling non-canonical forms and non-canonical form-function relationships as described in some of the tagging literature. For non-canonical forms, a strategy commonly applied is to add them to the tagger's lexicon, as was done in NECTE for dialectal lexis (Beal et al. 2007), and in the BNC for "interjections, slang and taboo words, and well as contractions". Although the dealing with non-standard words is not explicitly discussed for the tagging of CHILDES, there are various files in the lexicon folder which contain possible non-standard items, such as lexical items, e.g. verbs (e.g. "wee"), adjectives (e.g. "bunchy"), double forms (e.g. "nice-nice"), and other forms (e.g. "booboo") typically used by infants, as well as "terms local to the UK" (e.g. "fave"), some bilingual forms (e.g. "wai"), and loan words (e.g. "smuck") (MacWhinney 2012c: 163f.). Such forms also have to be added to a tagger's lexicon as many words occurring in spoken language are not available in tagger lexicons developed for written language. When added to the dictionary, the words also need to be assigned a word-class category. However, not much information is available on this issue. An example discussed in Greenbaum and Yibin (1994) for the ICE-GB demonstrates that this is in some cases done according to the token's function in the respective environment. Although Greenbaum and Yibin (1994: 42) do not discuss the problem of non-standard forms in detail, they give the example of the token "adither" [original emphasis], which is "not an established word" in the (written part of) ICE-GB. They decided to tag this word according to its syntactic characteristics, as it e.g. "can be intensified by *very*" [original emphasis] (ibid.). They write that many word classes of lexical items are not captured in dictionaries, even "of words that are well-established" (ibid.).

A second, and possibly more challenging, issue are cases in which the paradigmatic form and syntagmatic form of a token diverge in some way. When talking about forms and functions, this is used to refer to the terms 'form' and 'function' on a morphosyntactic level, referring to syntagmatic functions and paradigmatic forms of individual lexical entries (Atwell 2008: 507). The term 'paradigmatic form' in this context refers to, e.g. nouns, which can be inflected to refer to singular or plural whereas e.g. adjectives cannot. 'Syntagmatic function'

refers to “specific sentence-slots” where forms can be placed, e.g. heads of an NP (ibid.). Atwell (ibid.) writes that, “[u]sually paradigmatic and syntagmatic criteria coincide, but there are some exceptional cases, and different English corpus tag sets may handle these borderline cases differently.” Examples Atwell (ibid.) gives are words ending in *–ing*, which are “inflected forms of verbs” but “can also function as an adjective or noun”. In cases where such forms and functions diverge, the corpus compilers have to make decisions as to which tags should be assigned, e.g. whether form or function should be given preference. This is an issue for monolingual as well as for multi-lingual corpora, for those dealing with standard, as well as those dealing with non-standard language, e.g. dialectal, data. Surprisingly this issue, which requires difficult and consistent decisions in the tagging process, is seldom addressed in the tagging literature on spoken corpora. It is unlikely that this is due to a low occurrence of such cases. Rahman and Sampson (2000: 305), for example, write that “spontaneous speech […] contains a high incidence of such forms”, and as a consequence, devote a whole section of their article to this problem.

A review of the available literature reveals that the issue of assigning functions to certain forms is not particular to ELF data. In fact, even in corpora dealing with standard or written language, decisions have to be made whether to tag certain forms with regard to form or function of the verb, an example being collective nouns. Cases such as “een groep toeristen” (‘a group of tourists’) in the Spoken Dutch Corpus, make evident a discrepancy between singular (“een groep”) and plural (“toeristen”) (van Eynde, Zavrel & Daelemans 2000)[35], as “the NPs just after the finite verb denote an aggregate of […] tourists”. The decision which has to be taken here is whether to give *group* a singular or a plural tag. In Spoken Dutch Corpus the guideline was to tag according to form, i.e. give *group* a singular noun tag despite its “semantic plurality” because the word “lack[s] the affi[x] which [is] typical of plural nouns” (van Eynde, Zavrel & Daelemans 2000). In the Penn Guidelines, collective nouns receive a tag according to the verb which follows, e.g. if the verb is singular, the collective noun is tagged singular too (tag: NN), if the verb is plural, the collective noun is tagged plural (tag: NNS) (Santorini 1991: 18). In the LOB tagset, words with the suffix *–ing*, are tagged according to syntagmatic function, as verb, adjective or noun (Atwell 2008: 507).

In NECTE, form-function differences were encountered in terms of forms which are present in the dialectal data and also occur in Standard English but have a different function there, i.e. “a different grammatical reading according to context” (Beal et al. 2007). Examples given are past forms having the function of past-participles, e.g. went in “if I’d *went*”, simple present

[35] No page numbers in pdf.

forms in function of past forms, e.g. in "then well they *give* us the flat", or "the normally subjective form *we* is used in object function after *put*" in "my mam both put *we* into swimming when we were younger" (Beal et al. 2007). They write that cases such as *went* in "if I'd *went*" were tagged with a past participle tag, so according to the syntagmatic function. However, Beal et al. (2007) write that the tagging of such form-function differences cannot always be done automatically as "few examples in the corpus provide such explicit clues, and even then the clues may not be wholly reliable" (ibid.).

For CHRISTINE, "words used in nonstandard grammatical functions are given the same wordtags that the relevant wordforms are given in their standard uses" (Sampson 2000), as the tagging according to function was found to be unfeasible for spoken language, where the function cannot always be clearly assigned (Sampson 2000). This means that the tokens are tagged according to form, e.g. *awful* in *awful quiet* is tagged adjective (JJ) according to its form, not adverb, according to its function (Sampson 2000). Other examples which Rahman and Sampson (2000: 306) mention are object pronoun forms, e.g. *me* in *"me day"*[36], in the function of possessive pronouns or base verbs which function as verbs referring to the past, e.g. "*a man bought a horse and give it to her*". However, the authors write that the tag assignment in such cases is often not clear, as the form could represent *given* or *gave*, as in "[*w*]*hat I done*". The reason for this is that in nonstandard dialects, the auxiliary is sometimes not present (Rahman & Sampson 2000: 306). For cases such as "[*w*]*hat I done*" the authors raise the question whether this can be explained in terms of the past participle replacing both past tense forms, as has been suggested by Eisikovits (Eisikovits 1987: 134 [page refers to a reprint of the article in Trudgill & Chambers 1991]), but Rahman and Sampson (Rahman & Sampson 2000: 306) argue that this explanation disregards cases such as *got*, "where got clearly corresponds to standard *have got*, meaning "have", and not to a past tense". The solution to this is "that any verb form in a nonstandard structure with past reference will be classified as past participle" (Rahman & Sampson 2000: 306).

The above examples demonstrate that differences between form and function are present in L1 corpora, whether they are concerned with dialect or standard data. It is unfortunate that their handling has been rarely discussed in detail and is limited to a small number of individual cases. For the L2 corpora,with second and foreign language data which were discussed in 3.1.4, corpus annotators are bound to encounter a vast number of such cases. However, these issues are not discussed there either, neither in the ICE- nor the tagged LINDSEI sub-corpora. However, Meurers and Wunsch (2010) and Rastelli (2009), all of

[36] Original emphasis for all examples by Rahman &Sampson (2000).

whom are in favour of investigating L2 language "in its own right" (Meurers & Wunsch 2010: 1), as opposed to merely in comparison to the L1 target language, address this issue with regard to written learner language and propose practical solutions for the tagging process which can also be useful for dealing with spoken multilingual corpora. Rastelli (2009) used TreeTagger (cf. Schmid 1994), for the purpose of part-of-speech tagging a corpus of Italian L2 speakers, instead of error-tagging non-codified items. TreeTagger is a probabilistic, decision tree based L1 tagger which assigns tags according to lexical root, morphology and context, using a decision tree. Moreover, TreeTagger assigns confidence rates to tags, e.g. a target word is assigned a 70% likelihood for it to be a noun and a 10% likelihood for an adjective (Figure 3).

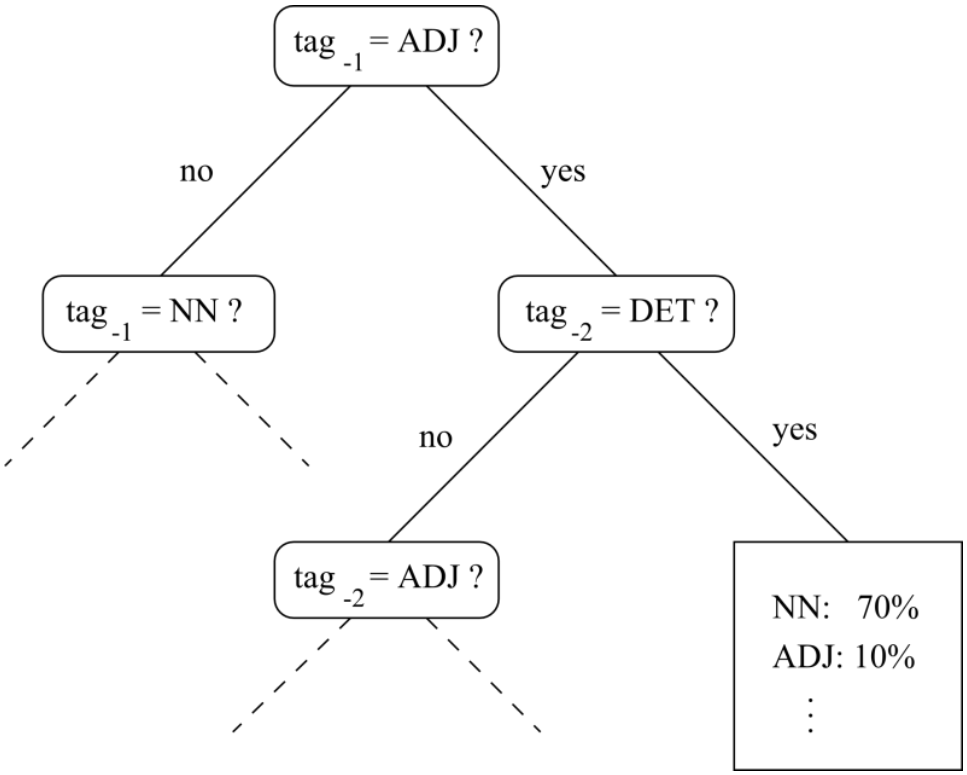

Figure 3: A sample decision tree. Source: Schmid (1994). Reproduced with permission.

The researcher can decide whether to output high or low confidence rates with analytical corpus tools. Rastelli (2009: 60) views the POS tags which TreeTagger assigns as "virtual" categories, rather than "psychological realia in a learner's mind" (2009: 59). He suggests that, in order to better investigate the language learners' language systems, the matching of forms and TL function should be done at the later, analytical, stage. The method he used in his study was to use low confidence rates of tags as indicators for form-function mismatches and as a starting point for analysing the interlanguage system.

Meurers and Wunsch (2010: 2) also raise this issue, arguing that for written learner language from the NOCE corpus, "lexical evidence, morphological evidence, and evidence drawn from the distribution" often diverge and cannot be combined to just one part-of-speech tag[37]. Examples they give are "*to be **choiced** for a job*", or "*one of the favourite places to visit for many **foreigns***" [original emphasis] (Meurers & Wunsch 2010: 2). For the latter, they summarise that "the token *foreigns* […] is classified as an adjective according to its lexical stem, but as a noun according to its distribution and morphology" [original emphasis] (ibid.). They propose "an automatic tagging approach which separately annotates the three sources of evidence" in order "to encode conflicting information" (Meurers & Wunsch 2010: 3). Meurers and Wunsch (2010: 2) mention that for L1 data, these three aspects "are assumed to be compatible sources of evidence jointly contributing to a single classification". However, as the examples from NECTE and CHRISTINE demonstrate, this is an assumption which oversimplifies the nature of native speaker data, where form-function differences occur just as they do in non-native data.

Both approaches described by Meurers and Wunsch (2010) and Rastelli (2009) differ from the procedures used in L1 corpora in that they advocate a more differentiated handling of form-function differences, rather than collectively tagging all such cases according to form or function. It becomes clear that when investigating the categorisation of L2 use in detail, it is hard to avoid careful examination of the issue of form-function differences and the crossing of word class boundaries, as it has regrettably been done for L1 corpora. Also, both approaches, although they were designed for written learner data, which is in many aspects different from the multilingual, spoken ELF data, can serve as useful starting points and suggestions for tagging VOICE. In a multilingual corpus such as VOICE, which contains much less uniform data, we came across numerous cases which overlap with those described for NECTE and CHRISTINE, as well as those described by Rastelli (2009) and Meurers and Wunsch (2010). However, in the case of ELF data, the matter is even more complicated: While in native speaker varieties there is a clear standard variety to refer to, there are various varieties to choose from in an ELF corpus and it is questionable which, if any, of these should be chosen as a point of reference, as ELF is predominantly used among NNSs of English. In addition, the number of such form-function differences is so great that in order to produce a meaningful POS tagging scheme, these cannot be dismissed as isolated cases but need to be analysed and described in detail.

[37] Meurers and Wunsch (2010: 2) also argue that this is the case for L1 corpora, where "the three [sources of evidence] are assumed to be compatible sources of evidence jointly contributing to a single classification". However, as the examples from NECTE and CHRISTINE demonstrate, this is an assumption which oversimplifies the nature of native speaker data, where form-function differences also occur.

3.2.3.5 Variability of spoken language

Another issue which was encountered in the tagging of VOICE is that of variability and fluidity of spoken ELF. As Seidlhofer (2011b: x) observes, language is by its nature "dynamic and variable" (ibid.) and this is the pre-requisite for it being used as a tool for interaction. According to Seidlhofer (2011b: 110), variation is defined as "some virtual capacity for exploitation, inherent in the encoded language itself" and this "constitutes the basis of its creativity" (ibid.). This is especially evident in ELF communication, where English "continually gets appropriated and re-fashioned by its speakers" (Seidlhofer 2011b: 111). The argument continues, following Widdowson (1997, 2003), that such a hybrid language use as ELF can be explained by an underlying virtual language, i.e. the "encoding properties" of a language (2011b: 114). In ELF, the variability originates mainly from the different language resources speakers draw on, and their adaptation of these for the communicative situation at hand. Given the large number of different lingua-cultural backgrounds of the speakers in ELF contexts, it is therefore not surprising that reference works used in POS tagging procedures, such as dictionaries and grammars, as well as other corpus data, are too restricted to capture the nature of the data. Even though this will also be true, according to Seidlhofer, of any language use other than ELF, this issue is very rarely addressed in the tagging literature.

Admittedly, there is not likely to be one single solution for dealing with variation in part-of-speech tagging. The issue of variation is mentioned in Rahman and Sampson (2000: 307), in reference to the corpus CHRISTINE, where a number of different dialects occur. It is argued there that "grammatical dialect variation" is much more an issue in spoken, as opposed to written data. In spoken data, the authors (2000: 307) continue, "it cannot be ignored, but the exercise of specifying annotation standards for unpredictably varying structures seems conceptually confused", and that it is difficult to account for this variation in terms of "consistent rules" (Rahman & Sampson 2000: 306). In VOICE, individual solutions were developed for the variation encountered, e.g. the variable use of paradigmatic forms (discussed in section 4.3.1.1), form-function relationships (section 4.3.1.2) or the flexible use of collective nouns, and tokens ending in the suffix *–ing* (addressed in section 4.3.3). The discussion of the flexible use of language also gives rise to theoretical questions about the degree to which spoken ELF can or cannot be categorised and described.

3.2.3.6 Other theoretical considerations in the tagging process

The many challenges which arose in the tagging process of VOICE, the numerous discussions the project team had on these and the difficulties we had with making decisions gave rise to the consideration of a number of theoretical questions. As has been discussed, the data

encountered in VOICE is unlike other data in terms of heightened variability and complexity. At the same time, however, it has been shown in the previous sections of this chapter that other corpus data gives rise to similar issues to those encountered in VOICE. It is therefore surprising that, when consulting the tagging literature, only very little is said about theoretical considerations which emerge from tagging spoken language in use, e.g. fundamental questions regarding categorisation of language. Rather, the problems which are mentioned are often restricted to practical issues of language categorisation within the tagging procedures.

Occasionally, however, there have been calls for a theoretical discussion of the annotation of spoken data with parts of speech. Rahman and Sampson (2000: 300) call for such a theoretical engagement with the topic, saying that discussing "conceptual problems that arise in defining rigorous annotation standards for spontaneous speech" is highly important, despite being often not considered as "advancing the discipline [of computational linguistics]" (ibid.), and that "[a]t the current juncture in computational linguistics, conscious-raising about problematic aspects of the subject-matter is a high priority" (2000: 301) and do indeed raise theoretical issues. MacWhinney (2009: 54) argues along similar lines for the data in CHILDES, writing that a theoretical discussion can emerge out of morphosyntactically annotating spoken data:

> Much of the material discussed in this chapter has involved issues that could be dismissed as "messy" coding details. However, in reality, making systematic decisions about the treatment of interjections, adverbs, or specific grammatical relations involve fundamental theoretical thinking about the shape of human language and the ways in which language presents itself to the child. (ibid.)

One specific theoretical issue which arose by tagging KiDKo is discussed by Rehbein, Schalowski and Wiese (2012a). They argue that in the tagging process, it is crucial to consider the way the data is approached, i.e. a "theory-driven" approach, where a theory is the point of departure and compared to the data, or a "data-driven" approach, where the data is investigated in its own right (2012a: 31). While an approach which is "theory-driven" helps to enhance a uniform morphosyntactic annotation, it can also lead either to referring the new features to parts of the theory which do not match or to ignoring phenomena which do not occur in the theory (Rehbein, Schalowski & Wiese 2012a: 31). They (ibid.) write, "[w]e argue that the theory-driven approach runs the risk of ignoring phenomena which are not licensed by the theory". When looking at the majority of the tagging literature, where such problems are not discussed in the tagging process and e.g. spoken features of the data are fitted into sometimes very broad categories, it would appear that what Rehbein, Schalowski and Wiese (2012a) argue has been happening. However, adopting a strongly data-driven approach is,

although preferable, not without problems either. When working with ELF data, even basic, generally accepted constituents of linguistics, such as word-class categories, are put into question, as these are often transcended by the speakers and there are many cases where no word-class category can be assigned unambiguously. However, while reaching a practical outcome, i.e. a tagged corpus, without having any theoretical presumptions such as word class categories is of course unrealistic. This dilemma makes it, as Rehbein, Schalowski and Wiese (2012a) point out, crucial to critically examine one's approach before undertaking the categorisation of data in part-of-speech tagging a corpus.[38]

In conclusion, it emerges from reviewing the relevant literature that there is still much to be said about what the assignment of word class categories can tell us about the nature of language and language use, and many theoretical questions that arise from the part-of-speech tagging process remain to be explored, as Rahman and Sampson (2000) pointed out over a decade ago. It is hoped that this thesis can contribute to both of these aspects.

3.3 Conclusion: Relevance of the findings for tagging VOICE[39]

In this chapter, it was found that even though there is a bias towards written language, there exist a number of spoken corpora, including L1 English, L1 other than English and L2 corpora. It is, however, not always straightforward to categorise corpora as belonging to only one category, as naturally occurring data, e.g. KiDKo and the corpus of spoken Irish may include speech of native, as well as non-native speakers of a language. Corpora vary in the kind of spoken language data to which tagging procedures have been applied and can in some cases be monologic (as in SEC) whilst in others it is interactive (as, e.g., parts of MICASE). Furthermore, some data is elicited (as in SWITCHBOARD) and some naturally occurring (as in the Gothenburg Spoken Language Corpus), either spontaneously (e.g. in the Nordic Dialect Corpus or Kiezdeutsch-Korpus) or to some degree pre-scripted (corpus of spoken Irish, components of LINDSEI). There are also tagged corpora of standard (e.g. Spoken Dutch Corpus, ICE corpora) and dialectal usage (NECTE) and in specific domains, e.g. television (COCA), teenage talk (COLT) or adult-child conversations (CHILDES). Not only corpora of L1 data, such as all corpora mentioned above, have been annotated with parts of speech, but also a number of the ICE corpora, as well as parts of two sub-corpora of LINDSEI.

[38]In this article, the non-canonical variation is only discussed on a syntactic level (i.e. how many constituents standard German allows for the prefield, cf. Rehbein, Schalowski and Wiese (2012a: 31), not with regard to part-of-speech tagging in particular. It remains an open question whether solutions as straightforward as this could be implemented for POS tagging.

[39] The first two paragraphs of this section will be published in Osimk-Teasdale and Dorn (forthc.).

In comparison, all the data in VOICE is spontaneous, naturally occurring and interactive (VOICE Project 2013d). However, it is distinctive in that it is not of a particular preconceived variety: it is not standard or dialectal and is not confined to a particular domain or even to one language. Because VOICE data was collected in entirely naturally occurring and highly interactive settings, neither scripted speech, nor monologues were used for the corpus (Breiteneder et al. 2006: 164) and a full range of features of spoken contexts (such as discourse markers, filled and unfilled pauses, response markers, truncated words, repetitions, speaker interruption, onomatopoeia and formulaic expressions) is present. In addition, what differentiates VOICE from other corpus data is that it is a plurilingual corpus, consisting largely of the use of an additional language (mostly English) but also that of other additional languages, as well as native language use (of English and other first languages). The range of first languages is by far larger than in any L1 corpus discussed elsewhere, e.g. the corpus of spoken Irish described in Uí Dhonnchadh, Frenda and Vaughan (2012: 5) or the KiDKo (cf. (Rehbein, Schalowski & Wiese 2012a: 29), and the speakers largely do not share a lingua-cultural background in the same way that e.g. the speakers in KiDKo most likely do. Due to these highly diverse cultural and linguistic background of the speakers, as well as the natural occurrence of the speech situations which were recorded, and which can of course be much less 'controlled' than e.g. scripted or elicited speech, the variation encountered is therefore likely to be much greater than in any of the corpora discussed.

In general, it can be noticed that unfortunately, the procedures used in the tagging of spoken language are rarely described in much detail. This is especially true for the few tagged L2 corpora, in particular the learner corpora which have been mentioned. If there is a description of the problems and the proposed solutions, it often does not contain enough detail to be sufficiently helpful for the tagging of new data such as in VOICE. This is what we noticed when trying to apply existing procedures to the tagging of VOICE. Some of the newer publications which discuss more detailed tagging of spoken features, e.g. that of KiDKo described in Rehbein, Schalowski and Wiese (2012a) or Rehbein and Schalowski (2013), were not available at the time when the most important decisions were made as regards tagging in the VOICE project. The general lack of discussion of the problems also poses the question of why this should be the case. Either there are no problems or they are not seen as worth discussing? The fact that a small number of corpus compilers do discuss the issues very explicitly indicates that the latter is the case and that the problems encountered were often simply ignored.

However, what little adaptation can be drawn from the literature as a basis for the tagging of VOICE indicates that the use of written taggers for spoken language is viable, and that good results have been achieved with this method. Furthermore, that the representation of spoken features is not necessarily a disadvantage, but can be beneficial to the POS tagging procedure and that there are various methods of dealing with uncertainties and ambiguities in the data, such as generic tags or smaller tagsets. Also, the dealing with non-codified language use in other corpora can serve as a starting point, in that the adding of non-standard lexis to the tagger's lexicon is one way of dealing with these issues. Some issues however, such as the handling of non-codified form-function relationships, as well as the variability of spoken data, leave much to be desired as does the discussion of theoretical issues arising from the part-of-speech tagging processes. The discussion of these issues can merely serve as a first step for tagging VOICE.

Therefore, with a number of issues relevant to ELF data not addressed, the approaches adopted to tagging L2 data being unsuitable, the documentation of categorisation not containing enough detail for developing tagging procedures and, largely, no theoretical questions being raised, new ways had to be found for tagging VOICE data. Moreover, with regard to the attitude towards the spoken nature of the data at hand, the approach adopted in VOICE is similar to those papers which adopt a 'data-driven' approach, as described for KidKo. We decided to have an explicitly positive and open-minded approach towards the nature of our data, which naturally resulted in the representation of spoken features and features typical of multilingual language use in the tagset. Within the framework of ELF research, the variation it features has always been embraced, and hence the approach in tagging VOICE was to seek to display this variation within the part-of-speech tagging process. It was this positive attitude towards spoken language, but also the occurrence of non-codified forms and all other variations that required us to discuss the issues emerging through the tagging process in more detail. Part-of-speech tagging in VOICE was not merely seen as a means to an end but first and foremost as a journey of discovery which can uncover interesting aspects of the data, rather than merely presenting us with problems of how to tag this variable type of data. In the next chapter, the theoretical issues in establishing a methodology for assigning part-of-speech categories to the ELF data in VOICE will be elaborated on.

4 THE POS TAGGING OF VOICE

4.1 Introduction: Issues and challenges in the categorisation for part-of-speech tagging VOICE and ELF data

When embarking on the challenge of assigning word class categories to the data in VOICE, a number of theoretical and practical challenges emerged. These were:

1. the **feasibility of categorising** ELF data,
2. the usefulness of **external points of reference**,
3. **non-canonical language use**, especially that of non-codified forms and non-codified form-function relationships
4. **highly interactive, spoken** nature of the data resulting in a vast number of possible tags,
5. **the fluid, creative** nature of the data, resulting in a high variability of forms and
6. other issues arising from **tagging transcribed data** in VOICE.

While the first two points concern general issues in the tagging process (discussed in section 4.2), the remaining points 3 to 6 refer to the practical dealing with individual aspects of the data (see section 4.3). From the discussion of the individual issues listed above during project meetings, certain principles which guided the tagging process of the VOICE data were developed (see section 4.4). Without these, reaching a certain consistency in the tag assignment would have been impossible. The present chapter of this thesis will elaborate on how these theoretical, as well as the more practical, issues were dealt with in the tagging process. While the technical implementation of the tagging had its own numerous challenges, these sections focus on the challenges in tag categorisation and the procedures developed for the large amount of data which had to be annotated manually. Details in the technical aspects of tagging VOICE, e.g. the taggers and tagging procedures used, can be found in the VOICE Part-of-Speech Tagging and Lemmatization Manual (VOICE Project 2014a: 6ff.) and are subject of Radeka (in prep.).

4.2 Issues of general principle

4.2.1 The desirability and feasibility of categorising VOICE data[40]

The two questions which were foregrounded in the beginning of the endeavour of attaching part-of-speech categories to the data in VOICE were firstly, is it **desirable** to tag ELF data, and, secondly, is it at all **possible** to categorise ELF data in this way?

[40] Parts of this chapter have been published in Osimk-Teasdale (2013).

The first question, concerning the **desirability** of tagging VOICE, was concerned with the issue of whether putting ELF data 'in a box' of POS-categories was something that was worth aiming for within our approach of ELF, where the aim was to do justice to the flexible and variable character of the data. After all, the concept of word class categories as stable entities has been critiqued, and they have been described as something 'fuzzy' and 'gradient' (see section 2.2). This is something which is even more true of ELF data, where we encountered numerous cases which did not easily fit into word class categories that had been defined for standard ENL. Moreover, part-of-speech tagging also means having to refer to certain standards, usually ENL based, which is potentially problematic for investigating ELF in 'its own right'.

However, some good reasons for attempting to annotate VOICE with parts of speech are that first, it is standard practice in corpus linguistics, which, amongst other advantages (as discussed in section 2.3), enhances the usefulness of a corpus for researchers. Since one of the main goals in the second project phase of VOICE was to enhance the usability of VOICE, part-of-speech tagging seemed one worthwhile possibility to do this, as the disambiguation done through the part-of-speech tagging facilitates many searches in the corpus for its users. Secondly, the process of trying to categorise the data and the encountering of problems are in themselves valuable processes because they point to issues which could otherwise not easily be investigated. For example, research into word class shifts, as described in chapter 5 of this thesis, is something which is extremely difficult to conduct with a corpus which is not part-of-speech tagged, as such information is usually not captured in the transcripts.

However, having noted the advantages of part-of-speech tagging VOICE, there are also certain considerations, to some extent 'drawbacks', if one will, which need to be taken into account. Firstly, the part-of-speech tagging of spoken ELF means the addition of a second layer of interpretation on the already transcribed interactions. Hence, the tagged data is even one step further removed from the original conversation than the transcribed data. This is something that researchers working with the tagged version of a corpus need to be aware of. Secondly, the process of assigning part-of-speech tags means that we are per definition referring to an idealised point of reference, namely word class categories that are closely connected to the (idealised) usage of Standard English. This means that language items are assigned certain characteristics and not others, and this categorisation might not always be entirely suitable to the data found in VOICE.

The second question which addresses whether part-of-speech tagging VOICE is **practicable** arose from the close engagement with the data in previous research, which made it evident

that large proportions of the data would be difficult to categorise. This is due to the highly interactive, spoken nature of the data, the nature of ELF and the multilingual language material, as well as the reliance on the transcripts, which structured the often overlapping speech into individual utterances and which also included an extensive amount of verbal as well as non-verbal mark-up. Also, although the audio data had been carefully transcribed and later thoroughly checked by project researchers, inevitably, some formal typing errors as well as a number of transcription errors were contained in the transcripts, which would cause some problems in the part-of-speech tagging process. In order to determine the feasibility of tagging VOICE, first, a small sample consisting of 543 tokens from the corpus (taken from the speech events LEcon565, LEcon566, LEcon573 and LEcon575) was tagged with parts of speech using TreeTagger.[41] This was then compared to a manual annotation of the test sample. In total, 459 of the 543 tokens assigned by TreeTagger were in agreement with the manual tags, reflecting an accuracy of 84.5%. A number of areas could be identified in which TreeTagger assigned a different tag than was assigned by manual annotation. An example includes the non-capitalised 'i', which was tagged as proper noun (NP) rather than a personal pronoun (PP). This was something that we considered for the further tagging process, as capitalization in VOICE transcripts is used exclusively for signalling stress. Secondly, base form verbs (VV) in questions were tagged as present tense verbs (VVP) or present tense verb 'have' (VHP) rather than base form verbs (VV) or base form have (VH), e.g. (3) and (4).

(3) S2: how do you say_VVP anachrom (LEcon565:23)

(4) S1: do you mind (if i have_VHP) it a bit more (LEcon566:233)

An accuracy of 84.5%, although relatively low compared to the gold standard, was a reasonable starting point for the further tagging process, especially considering that TreeTagger (as numerous other POS taggers, cf. Schmid 2008: 540) was trained on a set of data from the Penn Treebank, which is in many respects very different to the spoken and highly interactive character of the data and the nature of ELF.

Moreover, the sample selected contained a particularly large number of items which were presumably non-codified, i.e. not listed in OALD7, the reference dictionary used for VOICE, and also presumably not listed the tagger's lexicon. These were included as it was of interest how the tagger would deal with these cases. These items are referred to as pronunciation variations and coinages (<pvc>s) in VOICE.[42] Out of a total of 21 <pvc>s contained in the

[41] For the reasons why TreeTagger presented a good option for tagging VOICE, cf. Osimk-Teasdale (2013).
[42] See chapter 4.3.1.1. of this thesis for a more detailed explanation of <pvc>s in VOICE.

test sample, the manual tag and TreeTagger agreed for 12 (white cells) and disagreed for 9 (red cells).

Table 2. The 21 extracted <pvc> items

[Explanation] <pvc>s highlighted in bold with immediate co-text. Cases in which manually assigned tag and tag assigned by TreeTagger disagreed are highlighted in red. Numbers refer to the estimated probabilies with which TreeTagger assigns a particular tag.

Pvc	Word with immediate co-text	Manual tag	TreeTagger
1	is it **spanishy** or	JJ	JJ 0.929969
2	or **portuguesey** whatever shop	JJ	NN 0.730337
3	you say **anachrom**	NN	NN 0.532445
4	just like **putted** things in	VVD	VVN 0.989772
5	is slightly **liquidy** but	JJ	JJ 0.794834
6	is more **liquidy** yeah	JJ	NN 0.889958
7	it isn't **liquidy** i think	JJ	JJ 0.914492
8	never then **chinesey** ones	JJ	JJ 0.905536
9	have something **liquidy** then	JJ	NN 0.975937
10	not really **softish** huh	JJ	JJ 1.000000
11	just like **claustrophobicy** get	JJ	NN 0.987108
12	they look all **frenchers**	NNS	NNS 0.918265
13	but **slutty**	JJ	JJ 1.000000
14	and those **slutty** forty year	JJ	JJ 1.000000
15	a surreal **inscenation** or something	NN	NN 0.996255
16	grey zone **anyways**	RB	RB 1.000000
17	are not **dimensioned** we can't	JJ	VVD 0.501130
18	did not **re-enrol** he probably	VV	NN 0.682976
19	students don't **re-enrol** for	VV	NN 0.809839
20	you didn't **re-enrol**	VV	NN 0.510386
21	create the **sotteck** socket	NN	NN 0.720744

This showed that although the tagger could deal with some of the non-codified items by searching the fullform and the suffix lexicon (cf. Schmid 1994), this information was not sufficient for the tagging of these forms to be conducted automatically (cf. Osimk-Teasdale 2013 for a more detailed description of the experiment).

We also analysed to what extent the general tagging performance of TreeTagger could be improved when some problematic areas (which could be added to the lexicon in the further tagging process) were removed. In order to do this, the sample was tested while ignoring 11 VOICE-specific items (IPAs, incomplete words, proper nouns, significant non-verbal use of breathing, unintelligible speech). We also ignored the 21 <pvc>s, as their occurrence in the retrieved utterances (3.9%) is much higher in the chosen sample and thus not representative of the occurrence in the overall corpus (0.2%). All of the ignored items could be manually added to the lexicon in the further tagging process. Without the 21 <pvc>s and 11 VOICE-specific items, the accuracy increased slightly to 87.5 %. This result was considered a reasonable starting point for the further tagging process, but it also meant that it was necessary to expand the manually annotated material before a more thorough analysis of the error types made by TreeTagger on VOICE data could be conducted.

The extension of manually annotated material was done in the next step, the second manual tagging stage. In order to train the tagger, a manual annotation of 10000 words was carried out by my colleague Michael Radeka and myself between November 2010 and September 2011. This served to gain further insights into the issues of categorisation of the tokens, the suitability of the tagset used for the data at hand and to a better understanding of which items the tagger would be able to deal with automatically and which would have to be tagged manually. Furthermore, it served to train the automatic taggers that were used (see VOICE Project 2014a). And indeed, this manual tagging process pointed towards a number of cases where it was unclear which tag to assign. These will be discussed in the subsequent sections of this chapter.

This second manual tagging stage was carried out in three steps. First, for the so-called ‘GOLD 2500’, a sub-corpus of 2511 words in 252 utterances was annotated in late 2010. These were taken from the same speech events as the data used for the experiment of the initial 543 tokens with TreeTagger. In the second step, ‘GOLD 5000’, another 2685 tokens were added to GOLD 2500. GOLD5000 contained utterances with syntactic structures and features different to the ones already annotated in the first step or which contained features which were noticed to be problematic and where a decision on the handling was needed. Features that were included in the second step were past tense forms, subordinated clauses, unfinished clauses, hesitations, reformulations, interruptions, compounds, answer particles, discourse markers, and tag questions. In addition to the already used speech events from the leisure domain, data from PBmtg27 and POmtg404 was used. In a third step, another 5000 words were tagged manually, resulting in roughly 10000 manually tagged tokens in total,

hence called 'GOLD 10000'. This time, we mainly included speech event types (SPETs) and domains from VOICE which had not been present in the already annotated 5515 tokens, i.e. seminar discussion, service encounter, question-answer session, working group discussion workshop discussion, interview and panel. The speech events included were EDsed251, EDsve451, PBqas412, POwgd524, POwsd266, PRint30, PRint603, PRint595, PRpan13, PRpan294 and PRqas18. This served to obtain a better distribution of the training material regarding longer and shorter speaker turns, interactive and monological speech and different syntactic structures in the data, as well as a good bandwidth of lexis through spreading the training data over all five domains of VOICE.

In this manual tagging procedure, a number of issues of particular practice emerged which needed particular attention. These are discussed in subsequent sections of this chapter, in section 4.3. However, before continuing with these individual practical issues attention will be drawn to the use of external points of reference for ELF data. This has been an issue in ELF research in general and was also of great practical relevance throughout the process of tagging of VOICE.

4.2.2 External points of reference and ELF data[43]

4.2.2.1 Use of external points of reference in ELF

As was mentioned in the previous section, the decision to part-of-speech tag VOICE data necessarily entails referring to external codifications of language, which are not always entirely applicable to ELF data. The aim of this section is to explore the conceptual as well as practical problems which are nevertheless involved in assuming points of reference when working with ELF data, or any kind of data for this matter for which no standards are available, in short, any spoken language in use. So the dilemma we were facing is that while it is difficult not to use external sources which serve as common points of reference, there are many arguments against direct comparison of L2 or multilingual data to other sources, especially those which are based on descriptions of L1s. While it is apparently self-evident for other corpora, dealing with L1 and L2 language alike, to refer to standards, to a point where this is not even questioned in the literature, the issue of the role of standards and so-called 'native speaker English' have been a major issue in ELF research (e.g. Seidlhofer 2011b), as well as in some areas of SLA, and it has been so for many good reasons.

In ELF research, a direct comparison with ENL standards has been criticised as it has been shown that ELF speakers communicate effectively while making frequent use of forms which

[43] Parts of this chapter will be published in Osimk-Teasdale and Dorn (forthc.).

are neither "'correct' or idiomatic by ENL standards" (Seidlhofer 2001: 148). This has been demonstrated for different linguistic levels, e.g. lexico-grammar (cf. e.g. Breiteneder 2009, Hülmbauer 2009) and for pronunciation (cf. e.g. Jenkins 2000, Osimk 2009). In fact, Dewey (2009: 73) argues that it has been shown in ELF research that ELF users might be more successful precisely because they deviate from norms and, instead, adapt to a given communicative context and the other speakers. And, as mentioned in section 3.1.5, Smit (2009: 202) points out that that too much comparison to native speaker varieties can be a disadvantage in that it may lead to missing the defining features of ELF discourse. All of these aspects call the usefulness of L1 standards into question.

In SLA research, scholars such as Cook (2002a), Meurers (2010), Ortega (2010) or Rastelli (2009) point out that direct comparisons with monolingual L1 data have serious disadvantages and are highly problematic[44]. Cook (2002a: 8ff.), for example, argues that an L2 user's language system is complete at the time when it is used and that, hence, it is insignificant in how far it can be compared to an L1 user's system:

> L2 grammar is neither a defective version of the monolingual native grammar nor a partial transfer from the second, even if there are elements of both within it, but a creation of its own. In L1 acquisition the independent grammars assumption meant decoupling the child's grammar from the adult's. In L2 acquisition it means decoupling the L2 user's grammar from that of the monolingual native speaker.
>
> (Cook 2002a: 9)

For tagging ELF, Cook's approach of non-comparison to "monolingual native speaker[s]" is in line with the criticism of direct comparison with L1 data which has been proposed for ELF research. And it is, of course, also relevant for using points of reference based on descriptions of native language output in tagging ELF data, as it means that the language produced by ELF speakers will to some degree be fitted into categories described for a standardised L1 output, which has been, additionally, usually mainly based on written language.

However, before we continue, it is necessary to look at ELF research and note that it is often not straightforward to find a suitable way of referring to L1 standards there. For example, it can be noticed that the terms 'English as a native language' and 'Standard English' have often been confused. Seidlhofer (2011b: 46) argues that in general, these two terms have often not been distinguished sufficiently, although this would be necessary:

> […] StE is only *one* version of English as a native language (ENL). In spite of the frequent practice of referring to the two as if they were synonymous, StE does not

[44] And while it has also been noted that this has not reached main-stream SLA research (Cook 2002a; Ortega 2010) these are important developments in SLA which are in line with ELF research.

> refer to the language of native speakers as such but only a relatively small subset of them. Most native speakers of English do not, in fact, conform to StE norms, especially in their spoken usage. This is why when people who learnt English in their own countries first travel to Britain, the US, etc., it often comes as a nasty surprise to them that they very rarely encounter, particularly in spoken interactions, the Standard English they were expecting to hear [original emphasis].

And while Seidlhofer does not specifically refer to ELF research here, we can note that this is not only a problem in research dealing with different varieties of English, native and non-native, but can also be found in ELF research.

For example, Björkman (2009: 229) analyses non-standard morphosyntax in the language output of teachers and students in an academic ELF setting, reporting that, "the material was checked for non- standard and non-prescriptive features". Björkman (ibid.) grants that native speaker varieties may differ in what is "acceptable". As an example of this, the expression *travels slower* is given (Björkman 2009: 230):

> i.e. using 'slower' instead of the standard adverb form 'more slowly'. 'Travels slower' occurs in *standard usage* and *native speaker discourse* although the *prescriptive form* is 'travels more slowly', where an adverb is used to describe how the action is performed. [my emphasis]

In these extract, three types of language use are mentioned: *standard usage*, *native speaker usage* and *prescriptive language*. However, there is no indication of how these are defined, and where Björkman sees for example the difference between the standard and the prescribed use. What defines *non-standard […] features* compared to Standard English features? There is also no indication of what reference material *native speaker discourse* refers to. It seems that in the absence of saying what exact reference tools have been used (e.g. reference dictionaries, grammars, corpora), the terms 'standard usage', 'native speaker usage' and 'prescriptive language' stay abstract and subject to intuition. And indeed, Björkman (2009: 230) reports that way the non-standard features were detected was by 2 researchers who manually checked the transcripts and compared their findings, and hence it appears that the main source of reference may have been the researchers' intuitions.

The lack of references given also results in the concepts of Standard English and 'native speaker English' to be displayed in an over-simplified manner in ELF research, which is the second issue that arises in this context and which needs some discussion. When the term 'Standard English' is used without any indication of what the authors mean by this in detail, it could be argued, what is lost is the fact that Standard English is one of many English dialects, as Trudgill (2011 (revised version from 1999)) argues, which has merely 12-15% native users in Great Britain. Moreover, what might be discarded that there is not only one singular dialect

of Standard English, but that we might also refer to "Scottish Standard English, or American Standard English, or English Standard English" (ibid.)[45], and of course, the Standards of other Englishes which Trudgill does not mention. It needs to be stressed therefore, that Standard English, if anything, represents only a very minor part of the reality of English in use. It is, however, the kind of English which is presumably described in grammars and dictionaries, and hence inevitably represents an idealised abstraction from the much broader reality of English usage.

With regard to the way terms related to 'native speaker' are referred to in ELF studies, it can be noticed that even though the variation and different varieties of English are generally stressed, native varieties are often displayed as one entity. We come across terms such as "NS speech" or "NS baseline data" (Ranta 2009: 88). And while in this paper, the reference used is indicated, namely MICASE (Ranta 2009: 93), no further specification of what kind of native speaker data it is that is being referred to, i.e. American native English, implying perhaps, that such a differentiation is not necessary. This is, however, not the only example of a very loose application of 'native speaker (English)'. As conclusion of her investigation of humour in BELF (Business English as a Lingua Franca) Pullin Stark (2009: 171) writes,

> [a]nother area that would seem to be of interest for future research is with regard to the use of humour between *native and non-native speakers*. There is some evidence to suggest that whereas ELF speakers manage very well amongst themselves, despite linguistic limitations, contact with *native speakers* can prove problematic (Rogerson-Revell 2007), perhaps because humour becomes more culturally-based and its interpretation is dependent on "insider" knowledge [my emphasis].

Here, the concept of the *native* relates to native speakers in general, indicating that there is a culturally shared understanding of humour between all native speakers of English from e.g. the UK, the US and India and resulting in a simplification of the notion of the native speaker.

Moreover, even though this has been critiqued, ELF research has used direct comparison to native speaker standards, without calling this into question. Terms for example found in one article are, e.g. "overuse" (Björkman 2009: 234) or "lack of" (Björkman 2009: 233). And even though in other ELF research, efforts have been made to use more neutral terms and to specify the points of reference used, it can be said that in general, the use of terms and points of reference is not straight forward to handle in ELF research, where this has been topic of much discussion.

[45] No page numbers in pdf.

The issues we faced in part-of-speech tagging VOICE were of course not dissimilar to the difficulties which emerge when looking at the notions of standards, native speaker Englishes and prescriptive language use in relation to ELF research. It was the topic of many discussions if external references which are based on prescription or an idealised construct of language use by native speakers could and should be used, as well as which role these external references should have for ELF and the tagging of VOICE. One of the main topics was that points of reference, such as word class categories, are closely connected to the (idealised) usage of Standard English. This means that categories are attributed certain characteristics and not others, and this categorisation might not always be entirely suitable to the data found in VOICE.

Moreover, an analysis of the ELF data as a complete system entirely in its own right, as Cook (2002a: 8ff.) suggests for L2 speech, is also difficult to follow through in corpus annotation, although we would generally support the viewpoint put forward. For example, Cook (2002a: 10) proposes that the word *interesting* uttered by L2 users in a context where L1 users would typically produce *interested* is not an "incomplete" (ibid.) semantic use of the word, but "has a meaning of its own" (ibid.) and that "[w]hat is complete is the system at a given moment for a given user." (ibid.). We recognise for many of the instances in VOICE, where the speakers use language "in and on their own terms" (Seidlhofer 2011b: 23), that this might not be comparable to the way English is used in different inner and outer circle varieties. For example, in the case of the utterance *he speak*, this is not regarded as erroneous compared to a certain L1 dialect (e.g. Standard English), but as a legitimate manifestation of the speaker's multilingual language system, which has its meaning within this system. It might be that in this system, the 3rd person is marked in the pronoun 'he', or that the base form of the verb 'speak' also functions as 3rd person singular. However, while we recognise that either of these explanations (or, in fact, an entirely different one) might be viable, we also need to acknowledge that, in reality, it is often impossible to say which of these explanations may be most likely. This is especially true for a corpus of ELF data as VOICE, which includes 753 individuals with a total of 49 different first languages (not yet factoring in other languages, which, in addition to English, constitute the speakers' language repertoires), and various other factors which influence the way language functions, such as e.g. the pressure of online production and processing, or interactional factors such as accommodating to interlocutors. So even though the categories used by the speakers in VOICE may not always correspond to the categories used by monolingual English speakers of various Englishes, the multitude of factors involved renders it impossible to determine how the categories function in an individual users system and hence will often a result in a mere guess in the tagging process.

Moreover, the kind of analysis involved in attempting to investigate the speakers' individual language systems and categorisation of word classes in the way described would be far beyond the scope of part-of-speech tagging.

Thus, to throw external references, especially the part-of-speech categories, overboard because they often do not describe ELF fully appropriately, would have resulted in either no part-of-speech tagged VOICE corpus or in one that was most likely not intuitively accessible by other researchers working with the corpus. While the drawbacks have been discussed, it is also clear that there are certain **reasons for using points of reference for tagging ELF data** (see Figure 4).

First, commonly established guidelines and external points of reference enable the **usability** of corpus annotation as they constitute an assumed common point of reference for corpus users. Without this, an annotated corpus would be less approachable to corpus users, because familiarising themselves with newly established, internally defined categories and guidelines would be a tedious business. One aspect of this usability is also the enhancement of comparability with other corpora, as Rehbein and Schalowski (2012: 238) have argued. In our case, the Penn Guidelines served as a starting point for the VOICE Tagset. The tags used there are similar to those used in many other corpora, e.g. those which have been tagged with CLAWS. This facilitates the contrasting of different types of data and enables researchers to familiarise themselves more quickly with the annotation scheme used in a newly annotated corpus.

Secondly, the reference to established guidelines renders the tagging process **feasible** in the first place. It saves time because researchers can refer to references which have been agreed on instead of having to engage in lengthy discussion of decisions of how to tag a particular case. Moreover, it ensures a level of consistency which is difficult to achieve if individual tag decisions, which are often highly arbitrary, are carried out merely based on the manual tagger's intuition.

A third reason for using external points of reference and comparing the data to these, is that this type of **comparison can be helpful in describing a new type of data** by determining in what way one type of data is similar or dissimilar to another type of data. This can also be useful in corpus annotation, as Pitzl, Breiteneder and Klimpfinger (2008: 25) point out regarding the procedure of determining what was marked as a <pvc> item in VOICE:

> Basically, if something varies, there must be something it varies from. This means that one needs to know what is 'normal' or 'established' in order to judge what is

> ‘different’ or ‘new’. Of course, one might choose to make this distinction on the basis of intuition, but it seems prudent to look for a more reliable point of reference on which to base this distinction, i.e. a dictionary or corpus. Dictionaries and corpora provide such a stable reference point in that they record and capture language as it is used at a specific point in time. In so doing they create abstractions, recording a snapshot of the synchronic state of the continuously evolving and changing lexicon of a language. Lexicographers then take upon themselves the additional task of categorizing the lexicon of a language into lemmas and, depending on the size, scope, and purpose of the dictionary, define what is ‘normal’ usage and what is not.

In line with Pitzl, Breiteneder and Klimpfinger (ibid.), I would argue that the points of reference in themselves are not necessarily problematic within an ELF framework. This is especially true if, as has been proposed, the terms ‘Standard’ and ‘Native speaker English’ are understood as somewhat artificial notions, which cannot fully resemble the variable reality of language in use. If what they refer to is specified, such notions can fulfil a useful tool function. They are only problematic when it is either not specified what they refer to, or when ELF language is compared to native language with the underlying assumption that this is that multilingual L2 users have to live up to monolingual speakers’ standards (cf. Cook 2002b; Ortega 2010). And while Seidlhofer (2002: 271) argues that, apart from “the nature of the language itself as an international means of communication”, the exploration of the dissimilarities between ELF and ENL should be at “the very heart of the matter” (ibid.), she also assumes that research in ELF should adopt “a difference perspective with an acknowledgement of plurality”, rather than a “deficit perspective in which variation is perceived as deviation from ENL” (Seidlhofer 2009a: 238). This underlines the point mentioned above that points of reference are not problematic per se, but only when they are proposed with the assumption that what is compared is inferior to the reference used.

A last reason for using external points of reference such as ‘Standard English’, which is generally described by dictionaries and grammars, I would argue that it is safe to assume some notion of ‘Standard English’ serves as a common **point of reference for the ELF speakers** in VOICE. The majority of ELF users probably learned English in some formal setting (e.g. as a foreign language at school), and we can assume that they were taught a sort of ‘Standard English’. In these formal settings, as Widdowson (2012: 22) argues, their English “has been abstracted from the actual language performance they have been presented with and practised in”. It is thus a common point of reference that most ELF speakers share and has and probably still shapes their categorisation of language. Interestingly, extracts from the corpus itself in which the speakers discuss language use show this too. The three extracts from VOICE demonstrate that the view that some Standard English, even though what exactly this refers to may vary, is perceived as a common, although artificial point of reference (*that*

english that really NO one no one speaks, example (5)) which they speaker have been taught (*you learn a STANDARD english*, example (5)) and which they refer to when communicating (*we're all speaking standard english may have a different accent but* […] *you're still standard*, example (6))[46].

(5) S2: yeah you have to correct me there if (.) if it's not right but that's what i'm learning.
S1: yah yah <1> you are what what what </1> you're learning what you're learning it's the same. (.) well you are learning english
S2: <1> that what THEY tell me at university.</1>
S2: yeah?
S1: **you learn a STANDARD english**
S2: <fast> yeah yeah </fast>
S1: you kind of go (.) proper proper <2> e:r english.</2> is that **english** <un> xx </un> **that really NO one no one speaks.** (.)
S2: <2> yeah you judge about <@> english </@> @@ </2>
S1: like (.) just like the queen of england (1)
S2: <3> no:</3> she neither (.)
S1: <3> and </3>
S1: er a:nd er like someone who works on the <spel> b b c </spel> on television [bold emphasis added]

(LEcon352:151-162, S1: L1= spa-ES; S2: L1= ger-AT)

(6) S13: ENglish people always want a kind of give it erm:<smacks lips> (1) a place of where it comes from so ev- some people might think here we're speaking (.) american english or everyone learns eng- eng- lingua franca (.) being american english. or british english. but it's NOT when people use a LANguage to communicate to people from DIverse country and diverse language using a standardized version (.) it's like (.) arabic. arabic is spoken by SO many people there's egyptian arabic (.) jordanian arabic iraqi <1> arabic </1> but they speak STANDARD arabic (.)
S1: <1> mhm </1>
S13: <2> when they </2> come togeth<3>er </3>
SX-12: <2> mhm </2>
S18: <3> m</3>hm
S13: which is like what **we're all speaking standard english may have a different accent but**
S1: mhm (.)
S13: **you're still standard**
S1: it's a good example.
SX-8: <2> yeah.</2>
SX-m: <2> mhm </2> [bold emphasis added]

(EDwsd303:714-722, S13: L1=eng-GB, spa-ES, S1: L1=dut-NL, SX-12: L1=rum-RO, S18: L1=ger-AT, SX-8: L1= mac-MK, SX-m: L1=unidentified)

This does, of course, not mean that the speakers' language does not differ from a s*tandard* or that they sometimes choose to differ from it, nor does it mean that this *standard* is the **only** point of reference available to the speakers. Quite the opposite, the individual speaker's

[46] VOICE style.

lingua-cultural background(s) are legitimate constituents of ELF, as S22 says in extract (7), marked in bold:

(7) S22: as i (.) see it what makes american (.) as a language interesting is that is is in fact a lingua franca that has BECOME a mother tongue for people. because (.) english WAS (.) the language for e:r people of MAny different nationalities and mother tongues who came to america that they communicated IN (.) whereas it has (.) deVEloped into e:r (.) a mother tongue. (.) with (.) a certain (1) culture o:r e:r (.) background which is (1) a very mixed (.) background. (.) and **i think you have to: to accept that a lingua franca is not (.) any standard language it's (.) just a commun- communication tool but a communication tool DEVELOPing by the people using it we each (.) put SOMEthing (.) in our english from our own language =**
(EDwsd303:774, L1=dan-DK)

Figure 4: Points of reference in tagging VOICE and advantages of using these.

4.2.2.2 Guidelines for using external points of reference in VOICE

Having considered the difficulties as well as the advantages of using points of reference in tagging ELF data, let us now turn to the guidelines which evolved with regard of how the points of reference should be utilised for the part-of-speech annotation of VOICE:

Firstly, throughout the tagging process and in the descriptions thereof we generally **avoid referring to abstract and simplified notions** such as 'Standard English', 'prescriptive English' or 'native speaker English' and rather refer to 'points of reference', as has been done in this section. Secondly and more importantly, these points of reference are clearly defined, i.e. it is **acknowledged which exact points of reference were used** in the tagging descriptions, e.g. the VOICE Tagging and Lemmatization Manual (VOICE Project 2014a). For example, we clearly denote when we refer to tagging guidelines, grammars and dictionaries, as well as when we refer to other English corpora, e.g. e.g. BNC for multi-word items and ICE for formulaic items. In doing so it becomes transparent and traceable to other researchers why we assigned certain categories to certain tokens and not others. The main external points of reference which were used as starting points for VOICE were:

1) Commonly acknowledged word class categories

2) The reference dictionary (OALD7)

3) Other corpora, esp. Penn Treebank Tagging Guidelines

4) Grammars of spoken English

Thirdly, by naming and listing the points of reference that we used, we recognise their **potential but also their limits**. We acknowledge that the notions 'Standard English' and 'prescriptive use' are, to some **extent, artificial, idealised constructs**. Grammars and dictionaries, as well as other corpus data can never fully describe any language use, and especially not such varied and dynamic language use as we encounter in VOICE data. While the limited comparability to real language in use, native or non-native, is clear, such points of reference are useful in their function as *tools*, as they enhance the usability of the tagged corpus and, in fact, make the tagging feasible in the first place.

Before continuing with a more detailed description of the types of references used in tagging VOICE, it seems necessary to make two points:

First, it is obvious that the manual tagging could not be carried out without any type of 'language intuition' about what we presumed 'Standard English' or 'native speaker English' to be like. In manual tagging the data and contemplating how individual tokens or sequences should be tagged, our intuition determined which items were checked against the defined points of reference, and which were not. However, we tried to be very aware of the limitations of our own intuition and, where possible, second opinions were sought. In the end, the use of points of reference allowed the tagging process to be more traceable for other researchers than

using solely intuition and made it easier to be consistent. This is something which is sometimes impossible when relying simply on intuition, as the decisions taken tend to be arbitrary, and it is difficult to keep the overview of which decisions were taken for all of the cases encountered.

As second pre-emptive remark, points of reference can only serve as starting points in the annotation of new data. Although, therefore, we recognised that tagging VOICE would only be feasible when using certain points of reference, they had their limits when attempting to display the variability and fluidity of the data at hand. External points of reference, such as the Penn Guidelines (cf. Santorini 1991), were not designed to account for the spoken and 'ELFish' character of the data. For example, individual words, such as discourse markers (*so*, *like* etc.) and other tokens which serve different functions depending on whether they occur in spoken or written data are often not discussed the Penn Guidelines, and hence needed to be accounted for. Other tokens occurring in VOICE were not sufficiently dealt with in the Penn Guidelines, e.g. *used to*. We dealt with these cases by, for example, extending the tagset as well as tag formats proposed in the Penn Guidelines. Moreover, where it was necessary, we introduced individual solutions for features which did not occur in dictionaries but had a high number of occurrences in the data, such as various types of compounds or proper nouns and names. We will now turn to a more detailed account of the points of reference which were used for the tagging of VOICE:

Though using **established word-class categories** in part-of-speech tagging a corpus is rarely questioned, in the tagging of VOICE this was not a given. As has been elaborated on in section 2.2, linguistic categories such as word classes have been described as anything but fixed entities, and can be regarded even less stable when applied to such variable language use as ELF. As discussed above, we were therefore faced with the question of how to deal with balancing the use of pre-defined categories for ELF data, with the variable nature of this data. It was decided that in general, we would rely on pre-defined categories, because of its advantages to the corpus users, as pre-defined word classes are generally known linguistic entities which the users can more easily work with. However, we would also allow for the possibility that accounting for a different kind of data might require a reconsideration of pre-defined categories, which have after all been based on other data.

For the transcription of VOICE data, the **7th edition of the Oxford Advanced Learner's Dictionary (OALD7)** (Hornby et al. 2005) was used. The choice to use a reference dictionary was made for a number of reasons, primarily, to ensure annotation and spelling consistency within the transcription team (cf. Breiteneder et al. 2006: 179ff. and Pitzl, Breiteneder &

Klimpfinger 2008: 25). Breiteneder et al. (2006: 180) describe it as “a stable and shared point of reference for practical reasons”, meaning that it was considered up-to-date at the time when VOICE was being compiled, and a CD-ROM of the OALD7 could be made available to all transcribers, ensuring spelling consistency (ibid.). Moreover, OALD7 was chosen as its range of lexical entries was expected to correspond to the language usage of the speakers in VOICE (Breiteneder et al. 2006: 180). OALD7 also determined which items were especially marked in the transcription, i.e. those not listed in OALD7 were annotated with a <pvc>-tag in the transcription (Pitzl, Breiteneder & Klimpfinger 2008: 22). Moreover, the reference dictionary was used as a reference for spelling, other than those for which specific guidelines for VOICE were determined in the VOICE Spelling Conventions (VOICE Project 2007a), such as American spelling for words ending in the morpheme *-ize* (Breiteneder et al. 2006: 179).

Since OALD7 was heavily used in the transcription process, and the transcribed data served as the basis for the assignment of word class categories, OALD7 was the obvious choice for a reference dictionary in the part-of-speech tagging process. Herein, OALD7 was used to verify the word class categories listed for specific tokens, and was especially relevant in cases for which paradigmatic form and syntagmatic function were different from that in codified English. For example *westbound* in *the westbound,* was listed only as an adjective in OALD7, but also used to function as a noun in VOICE. Or *them* in *them bananas* was listed as pronoun, while it was also used as a determiner in our data. Of course in cases such as these it could be argued that either of these examples might be listed with different functions, i.e. with those that they serve in VOICE, in other dictionaries. For example, while in OALD7 *them* is only listed as pronoun, the Cambridge Online Dictionary lists *them* as determiner (“not standard for those”). However, this is one of the consequences of using a reference dictionary, that it can never describe all possibilities of language as it manifests itself when used in natural conversation, which we had to accept in order to use OALD7 as a reference dictionary. However, it is important to emphasise that the OALD7 was used as a tool for above named reasons, rather than an ‘omnipotent’ source of wisdom, as no dictionary, or other corpus as reference, could fully capture interactive, spoken and multilingual data such as that in VOICE, as it is inherently creative and in flux. Consequently, as with the other sources which were used for referential purposes, there were certain areas for which exceptions were made. With regard to OALD7, these were, for example, discourse markers and proper nouns and names. As these categories are not sufficiently covered, we consulted additional sources to do justice to their occurrences in VOICE data.

In addition, a number of other points of reference were used which come with the application of **tagging tools**. For tagging VOICE, a stacked TBL framework with three different taggers (TreeTagger, Stanford Tagger and LTAG) was used. However, rather than using preconceived lexica which come with taggers, a tagging lexicon was built for VOICE[47]. This was done in order to create a lexicon which encompassed as much of the VOICE data as possible. For this purpose, various sources which are generally acknowledged tools for POS tagging were consulted, i.e. different Penn Treebank corpora (Wall-Street Journal, Brown, SWITCHBOARD, ATIS-3), the lexical databases Wordnet (Fellbaum 1998) and CELEX (Baayen, Piepenbrock & Gulikers 1995), lists from GATE and the Natural Language Toolkit, the BNC and COCA (VOICE Project 2014a: 7). However, as all of these different sources did not sufficiently cover the data in VOICE, there was a large amount of data which had to be tagged manually, such as "discourse markers, multi-word items, long-distance dependencies, words standing alone or at the beginning or end of utterances" (VOICE Project 2014a: 8).

Another point of reference used were different **grammars**. These were consulted for cases which were not sufficiently covered in the points of reference named above (e.g. OALD7 and Penn Guidelines, see below). For example, we consulted Carter and McCarthy (2006b: 539, 541) and Quirk et al. (1997: 316f.) for a definition of collective nouns, which are used fairly flexibly in VOICE, and Biber et al. (1999: 70f.) for a differentiation of different English pronouns as well as for the distinction between relative pronouns and relative adverbs (Biber et al. 1999: 608).

Finally, we also took into account **literature on the tagging of various spoken corpora**, of both English and of other languages, established recommendations for tagging (e.g. EAGLES, Guide to developing linguistic corpora), as well as the tagsets of various corpora. Among other tagsets which were used for certain aspects of the VOICE tagging guidelines, such as the tagset for ICE and the BNC, the tagging of VOICE is heavily based on the guidelines described for the Penn Treebank Project (Santorini 1991), as already mentioned earlier in this in section. As described earlier, we initially worked with TreeTagger in order to test the performance of a tagger on small parts of VOICE. TreeTagger is based on the Penn Guidelines, which is why this tagset served as the starting point for tagging VOICE. These guidelines were created for written data, and later adapted for the spoken SWITCHBOARD corpus (cf. Linguistic Data Consortium 1999). There were four main reasons why we also

[47] cf. The VOICE Part-of-Speech Tagging and Lemmatization Manual VOICE Project (2014a) for a more detailed description of the building of the VOICE lexicon.

chose to adhere to the Penn Guidelines and Tagset in the subsequent process of tagging VOICE:

The first reason was that the tagset had an appropriate size for the data in VOICE. Containing 36 tags only, the categorisation achieved is rather coarse-grained (e.g. compared to CLAWS7 with 137 tags). As already mentioned earlier in this thesis, this can be seen as an advantage for tagging interactive spoken data, which will necessarily contain a large number of items for which unambiguous tags cannot easily be decided. By assuming broader categories, the, potentially unmanageable, number of ambiguous or undecided tags can be reduced. For some areas, such as features characteristic for spoken language or specific to our data, we modified the tagset of the Penn Guidelines to contain more detail. For example, while the Penn Guidelines only list one tag (UH) for spoken features, the VOICE Tagset contains eight different tags which refer to specific aspects of 'spokenness' in VOICE (BR, DM, FI, LA, ONO, PA, RE, UH)[48]. Secondly, the Penn Guidelines contain a reasonable amount of documentation of when to assign a specific tag, as well as explanations and assistance in how to differentiate between certain tag categories. Such documentation provides enough transparency to be taken as a basis for the annotation of untagged corpora. A third reason is that the tagset used for the Penn Treebank is a well-known tagset which uses tags that overlap with those assigned in tagsets for other corpora, e.g. corpora tagged with CLAWS (e.g. BNC, COCA, COLT, MICASE). We could thus expect that people working with other corpora, such as the BNC, would at least be partly familiar with the tag labelling. Finally, as has already been mentioned, the tagger first used on VOICE was based on the Penn Guidelines and hence, even though at a later stage it emerged that a combination of different taggers would be more suitable for tagging VOICE (VOICE Project 2014a: 8), time-efficiency was also a factor in staying with the guidelines, as researchers had become familiar with the guidelines and had already extended and adapted the tagset for the data at hand.

However, having considered the aspects which advocated working with the Penn Guidelines, there were a number of areas for which the Penn Guidelines and Tagset needed to be adapted. In some cases this was because the explanations given were not clear enough to be transferred to the manual tagging decisions which had to be taken. In other cases we found that the tag categorisation was inconsistent or inadequate for describing the data in VOICE. As a result, the Penn Guidelines were extended in terms of additions to the tagset (e.g. spoken features, see above) and to the tagging guidelines (e.g. more detailed guidelines for specific tokens and

[48] See Appendix A for short explanations of VOICE tag abbreviations and Appendix F for a more detailed account.

token sequences) and tagging format (e.g. dual tagging of form and functions, multi-word items, a different symbol for ambiguities than that used in the Penn Guidelines). In addition, some tag categorisations were altered (e.g. for cardinal numbers, spoken markers, relative pronouns and Wh-elements). All of these changes are carefully documented in the VOICE Tagging and Lemmatization Manual (VOICE Project 2014a).

In this context, the question arose of how much it was reasonable to divert from Penn, e.g. whether only the tagset should be extended or whether also tag definitions should be altered. After all, the value of using a well-known tagset is diminished for corpus users the less it resembles this tagset. On the other hand, the extension of tagsets is a common practice in tagging, as has been carried out for e.g. KiDKo (cf. Rehbein & Schalowski 2012), the Gothenburg Spoken Language Corpus (Nivre & Grönqvist 2001: 59) or CHRISTINE (cf. Sampson 2000), as especially for spoken corpora, tagsets initially developed for written data are commonly used. It was decided that we would rely on the Penn Guidelines and Tagset whenever this was in line with our approach to POS tagging VOICE. However, changes such as those explained above were introduced wherever it was felt that this was not the case. The principle was therefore: as few changes as possible with regard to the pre-defined categories but as many as necessary for the nature of the data.

Having discussed various aspects involved regarding the desirability and feasibility of tagging VOICE, as well as the use of external points of reference in the tagging process, I will turn to tagging challenges concerning specific aspects of the VOICE data.

4.3 Issues of particular practice

4.3.1 Non-codified language use[49]

As already pointed out in section 3.2.3.4, one of the issues we encountered in tagging VOICE was that of non-codified or non-canonical language use. As has been demonstrated, such language use is no distinctive feature of an ELF corpus, but also occurs in other types of data, both L1 and L2, standard and dialectal. In fact, Denison (2008: 533f.) argues that interest into non-standard forms is increasing and regards this as "a welcome development, because the bulk of linguistic scholarship has concerned itself, explicitly or implicitly, only with standard forms of language, neglecting what may well be the greater part of English users and usage." (ibid.)

[49] Parts of this chapter have been published as Osimk-Teasdale (2013).

For the annotation of corpora, the term ‘non-canonical’ has been used to refer to different concepts, e.g. items which are unknown to the tagger (Nivre & Grönqvist 2001: 59), to tokens which do not occur in “standard written text” (Rehbein & Schalowski 2012: 238), or to items which cannot be explained within a particular theory within linguistics (Hirschmann, Doolittle & Lüdeling 2007: 1). It seems, therefore, it is necessary to clarify the use of the term for this thesis, in particular in relevance to the present section. In ELF, various terms have been established which serve to differentiate features of ELF from those described in standardized grammars and dictionaries of native English. Dorn (2011: 26), for example, lists the terms “(non-)canonical, (non-)conventional or (non-)codified” as being used in ELF research, using (non-)canonical for her own work on the functions of the progressive in ELF. Hülmbauer (2013: 67), on the other hand, refers to “non-encoded elements”. Yet others use the term ‘non-codified’, e.g. Pitzl (2012: 37) in order to refer to “the creation of new […] linguistic forms and expressions” for her definition of “linguistic creativity” (ibid.). The terms ‘codified’ and ‘non-codified’ are also used by Pitzl, Breiteneder and Klimpfinger (Pitzl, Breiteneder & Klimpfinger 2008: 22) for “words which could not be found in the reference dictionary used for compiling VOICE”, annotated with the <pvc> label (ibid.). For this present thesis, the terms ‘(non-)codified’ and ‘(non-)canonical’ will be used synonymously to refer primarily to the forms and form-function relationships which have not been codified in the reference dictionary. It, hence, needs to be noted that the definition used here differs to some extent from the, above listed, definitions of ‘non-canonical’ which are usually found in the tagging literature.

As a preliminary remark, I would also like to address the issue of the perspective adopted in tagging VOICE, and in fact in the whole process of compilation and annotation of VOICE, with regard to non-codified items. As with other areas, this is a perspective which strives to investigate the ELF data at hand in its own right. While the non-canonical nature of language has been discussed in terms of particular dialects, in L2 user language, non-canonical use is generally classified as ‘error’, even when the cases are very similar. So, while in native corpora non-canonical forms and non-canonical form-function relationships are usually viewed positively as interesting instances of dialectal variation, when produced by L2 language users, they have often been regarded as erroneous approximations to the target language (TL) system that fall short of a native speaker standard, as Hirschmann, Doolittle and Lüdeling (2007: 6) state this difference explicitly,

> Depending on the model and the reason for not fitting into that model, deviations can be categorized differently. In a learner corpus they will mostly be classified as errors, in a spoken language corpus they could be analysed as properties of a spoken register.

Rastelli (2009: 58) describes this as, "systematicity (or lack) in how learners map forms and functions and in how they gradually develop their knowledge of the TL categories out of the available input". As has already been mentioned in section 3.1.5, from an ELF perspective, strategies of learning, as they often are in learner corpus research can also be viewed "as communication strategies" and it has been argued that the main object of interest should be what functions the forms which are uttered by ELF speakers fulfil (Seidlhofer 2009a: 241). The adoption of a perspective which recognises the functional motivation for non-codified forms, therefore, is a key characteristic of ELF research. This includes all the implications of how the users' languages are viewed and how this is implemented in research practices. The recognition that speakers in ELF conversations are language users, making deliberate linguistic moves, rather than being learners applying "compensatory actions […] problematic exchange or helping strategies to make up for non-understanding" (Cogo 2009: 259) as Cogo critically refers to with regard to literature on learner language (ibid.), although not always followed through consistently, generally guides the methodological approaches and presumptions of ELF researchers. It is of interest which strategies, forms and functions the speakers use to achieve these goals, rather than how they fail to do this in exactly the same way that a (rather small) group of native speakers speaking a standard variety would.[50] As outlined in section 3.1.5, extensive research has been carried out which demonstrates distinct communicative functions of different forms used in ELF. Similarly, for most unconventional items of both kinds described above for VOICE, whether newly coined words or non-ENL form-function correspondences, there are indications that the forms are nevertheless functionally motivated. Pitzl, Breiteneder and Klimpfinger (2008: 40ff.) demonstrated that for 247 <pvc> items of a sub-corpus of a total of 250 042 words, four main functions of these forms were to make expressions clearer, more regular, efficient or to coin a word where there is a lexical gap in English.

For tagging purposes, we can posit **two main types** of non-codified items in VOICE data. First, those which fulfil the criteria of a grammatical category (through word-formation processes, e.g. affixation), but happen not to have been codified in our reference dictionary (realisations of the virtual language, annotated as <pvc>s in VOICE). These are referred to as 'non-codified forms'. And secondly, those items with an existing form entry in the reference dictionary which does not correspond with its codified function (not annotated in the untagged versions of VOICE), henceforth **non-codified form-function relationships**. This

[50] This is in line with the theoretical framework of some more recent approaches in SLA which view second language speakers as L2 users (e.g. Cook 2002), address the "monolingual bias" in SLA research (Ortega 2010) and call for investigating the language of these users in its own right. However, this theoretical view is not usually one that informs SLA studies, where a native speaker target is generally still implied.

section will now first deal with non-codified forms, and then with non-codified form-function relationships, explain how these were defined and proceed with how these instances were dealt with in the tagging procedure.

4.3.1.1 Non-codified forms

As already explained, in the VOICE corpus, the annotation <pvc> is used to highlight items which were not defined in the reference dictionary OALD7, and show "[s]triking variations on the levels of phonology, morphology and lexis as well as 'invented' words" (VOICE Project 2007c: 4). In total, 2116 tokens were annotated with the <pvc> tag in VOICE 1.0 (VOICE 2011), in VOICE 2.0 the number of tokens annotated with the label <pvc> is slightly higher with 2195 (VOICE 2013). By capturing any items not found in OALD7, the <pvc> tag was intended to provide a starting point for later analysis (Breiteneder et al. 2006: 181). It is important to note that the decision on the definition of the <pvc> tag was "purely operational" (Pitzl, Breiteneder & Klimpfinger 2008: 5), and that the words captured under the <pvc> tag may include a range of different items, some of which might indeed be found in other corpora or dictionaries[51]. Beside the many newly coined words, such as *luckywise*, also specialised terminology e.g. the medical term *cytokines* or presumed slips of the tongue (examples (11) and (12)) are included. However, the vast majority of these <pvc>s constitute newly coined words which follow the rules of attested word formation processes such as affixation, analogy, reduction, borrowing and addition, to name the most common (Pitzl, Breiteneder & Klimpfinger 2008). At the same time, these <pvc>s have mostly not been codified in ENL (examples (8), (8)(9), (10)).[52]

(8) S1: i was just like (.) <pvc> **putted** </pvc> things in (LEcon565:378, L1=ita-IT)
(9) S1: it was like a surreal <pvc> **inscenation** </pvc> or something (LEcon573:76; L1=ger-DE)
(10) S1: with a diverse <pvc> **linguistical** </pvc> (.) group. (EDwsd303:387, L1=dut-NL)
(11) S2: one could say that er in a period of economic <pvc> **unsiternity** {uncertainty} <ipa> ʌnsɪˈtɜːrnɪtɪ </ipa> </pvc> o:f er (PRpan294:91, L1=slo-SK)
(12) S4: the LAST but one paragraph it says <reading_aloud> a <pvc> **comprison** {comparison} <ipa> ˈkɔmprɪzən </ipa> </pvc> (.) <pvc> **comprision** {comparison} <ipa> ˈkɔmprɪʒən </ipa> </pvc> of the evaluation guidelines to final evaluation reports </reading_aloud> (POmtg403:174, L1= hun-HU)

As the form of these tokens labelled <pvc>, per definition, were not listed in the reference dictionary, it was unclear which tag should be assigned in the tagging process. However,

[51] cf. Pitzl, Breiteneder and Klimpfinger (2008: 22ff.) for a more detailed account of the definition of the <pvc> tag and the guiding principles thereof.
[52] Bold emphasis in all examples taken from VOICE is added.

while the paradigmatic form was non-canonical, in the manual tagging process it emerged that these <pvc>s could often easily be assigned a syntagmatic function. For example *inscenation* in the sequence *a surreal* ***inscenation*** could unambiguously be assigned function of a noun, despite its non-canonical form. It was therefore decided to give a word-class tag according to the syntagmatic function of the <pvc>s. In the dual system of assigning both form- and function-tags to all tokens in the corpus (see section 4.3.1.2 below for a more detailed account of this), the tokens labelled <pvc> were assigned a form-tag named PVC and a function tag with a word-class category, which, in the case of the token *inscenation* from the example above, is *p_inscenation_PVC(NN)*.

In the tagging process, we also had to make decisions regarding a number of tokens which were similar to <pvc> items but had not been marked as <pvc> in the transcription process. These concerned primarily compounded adjectives. They were not marked <pvc> in the transcripts, as their individual components were listed in OALD7, though the full forms were not (VOICE Project 2007d: 25). Such tokens were, among others, *paper-intensive*, *vector-bounded*, *price-off*, *job-oriented*, *blood-building*, *all-in*, *narrow-sighted*, *fee-paying*, *gender-discriminatory*, *hongkong-based*, *asylum-seeking* or *react-to*. Analogous to <pvc>s, these were tagged according to their syntagmatic functions. As they had not been assigned a <pvc> label in the transcription process, they did not receive the form-tag PVC but a form that was identical to their function-tag. For example *react-to* in the sequence *on a react-to basis* was tagged *react-to_JJ(JJ)*.

4.3.1.2 Non-codified form-function relationship

As has been explained earlier in section 3.2.3.4, the terms ‘form’ and ‘function’ in this thesis refer to paradigmatic forms and syntagmatic functions of lexical items. This is different from the use of the terms as it is commonly applied in ELF research. There, the term ‘function’ is often employed in different ways. In some cases it relates to the use of ELF as an international language - as opposed to the function of English as national language (Saraceni 2008: 21), i.e. the contexts in which ELF is used. Another common use is, as has been explained earlier in this thesis, that of forms and their communicative functions, meaning that “functional motives can lead to changes in form“ which can then affect pragmatic strategies (Cogo 2008: 60). On a third level, Seidlhofer, for example, refers to the social functions of a language which motivate different surface forms (Seidlhofer 2009b: 41).

For the purposes of the annotation of VOICE, the definition of a non-codified form-function relationship in tagging VOICE was derived from the definition of a <pvc>. Because of the practical definition of the <pvc> tag, certain variations potentially relevant for assigning part-

of-speech tags are not included if the word form itself is an existing entry in the reference dictionary, such as paradigmatic forms which function as a different word class category than those indicated in OALD7. Pitzl, Breiteneder and Klimpfinger (2008: 26) state, "[a]n 'existing word' may also be used with an entirely new or different meaning or it can be used in another syntactic category. Yet, as long as the word itself can be found in the OALD7 it is not tagged as a <pvc>." (see examples (13) to (19))

The implication of the above quote is, therefore, that the aspect of non-codified ENL form-function mapping, i.e. the form exists but does not correspond with its codified function, is not captured by the <pvc> tag. When dealing with ELF data, this aspect poses possibly an even greater challenge than the non-codified <pvc> forms. This is, firstly, because such non-codified ENL form-function relationships can often only be detected by tedious manual annotation (though there are some automatic means which can assist to help detect similar cases to those found by manual annotaton). A second reason is that the cases found in VOICE related to non-codified form-function relationships display a large amount of variation, i.e. many cases different from each other were found. This variation within the group makes it even more difficult to decide on tags which display some consistency across the different cases. Examples (13) to (19) demonstrate such instances of non-ENL form-function mapping in VOICE data. These include the use of a zero morpheme where Standard English requires 3rd person present tense marking, e.g. (13) and (14), and zero marking of plural nouns as in (15), non-ENL past tense constructions, e.g. (16) and (17), as well as forms of adjectives which occur in typical adverb positions as in (18) and the opposite example of an adverb in adjectival position in (19).

(13) S2: you can TASTE it it **taste** even of milk but it's FINE. (LEcon566:183, L1=ita-IT)
(14) S1: the waiter is NOT somebody who **hate** americans (EDsed31:964, L1=ger-AT)
(15) S2: one can apply (at) this italian agency for (.) two **month** (.) (PRcon550:30, L1=slv-SI)
(16) S1: i don't want erm let's say this way i also didn't **spoke** to [first name34] in france (PBmtg27:527, L1=ger-DE)
(17) S1: in [org11] video camera has **break** down (PBmtg27:653, L1=ger-DE)
(18) S11: but er (1) we are <3> **complete** different </3> (EDsed31:1134, L1=ita-IT)
(19) S10: i just wanted to to give a (1) **partly** answer about the <pvc> baltics </pvc> (.) (POmtg404:924, L1=lav-LV)

This is not to suggest, however, that this phenomenon is unique to ELF or non-native speaker discourse in general, for form-function relationships are often not clear-cut in ENL either. Although paradigmatic forms and syntagmatic functions do mostly agree in ENL, many unclear cases arise, as has been described in section 3.2.3.4, and this has also, in some occasions, been mentioned in tagging literature. Many of the occurrences which we came

across for VOICE data can be found in any corpus of naturally occurring English, as examples (20) and (21) taken from the BNC and COCA show.

(20) Yeah I know our Lorraine **has took** the phone over. (BNC [http://corpus.byu.edu/coca/]: KBE 308)[53]
(21) It didn't hurt when I fell over cos the road's **real soft** and spongy (COCA [http://corpus.byu.edu/coca/]: A74 2982)

Denison (2007), who investigated the way taggers dealt with "five sets of data which fall on or close to word class boundaries" in different L1 corpora, argues that many cases where word categories cannot be safely assigned are due to syntactically vague categories, e.g. mass nouns used countably (Denison 2013a: 28), or the use of verbs ending in *–ing* as adjectives (Denison 2013a: 32), as well as 'borderline cases' such as e.g. adjective-determiner, adverb-adjective or noun-adjective (Denison 2007), and that for these, "corpus mark-up does not recognise word class vagueness even in principle – and maybe it should" (Denison 2013a: 23). He argues that this because some cases "do not have a unique word class, not because of a failure of analysis but because they are genuinely underdetermined: they are syntactically **vague**" (Denison 2013a: 21) and suggests that such vagueness should be "explicitly signal[led]" in tagging systems, either by ambiguous tags or other methods (Denison 2013a: 33). Denison (ibid.) suggests the following tagging procedures for such unclear cases:

> For a word of vague (that is, underdetermined) class, I would prefer tagsets to include tags that explicitly signal indeterminacy between two categories; they could be something like an ambiguity tag in form. In other cases I wish tagging could make distinctions that are deliberately avoided in corpora with which I am familiar.

The examples of syntactic vagueness which Denison names for native varieties also occur in VOICE. However, as ELF data has been shown to contain more variation than corpus data where the speakers share a common L1, this was also to be expected in the area of combination of morphological (i.e. form) and syntactic criteria (i.e. function). Additional to categories which are syntactically vague, as Denison describes it, problems occur because the categories which part-of-speech tags are based on originate from descriptions of codified English. These descriptions are, moreover, heavily based on written categories, which are only partially applicable to spoken ELF. Denison (2007) argues that syntactically vague categories are a normal constituent of language in use, and as such there is a certain discrepancy between real language use and the idea that unambiguous tags can be applied to such language data:

[53]Data cited herein have been extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

> I suspect that among the inconsistencies which have emerged, some are an irreducible consequence of a mismatch between language as actually used and a widely-held conception of grammar, namely that every word in every sentence belongs to exactly one category — no more and no less — selected from a small number of parts of speech. This assumption is common both to traditional grammar and to most modern theoretical approaches. Presumably, it is the working assumption behind POS tagging. […] I suspect strongly, however, that it does not correspond to actual language behaviour and is in consequence unworkable. That last point is, of course, an exaggeration, since categories — even if epiphenomenal — are more often than not quite straightforward to assign, and categorisation can be very useful in practice. Rather, the assumption of universal and unique lexical categoriality is a practical oversimplification of reality which works tolerably well quite a lot of the time — but not in such cases as I have chosen to concentrate on. Note, however, that such cases are not really at the margins of language use: they are a common part of everyday colloquial English. (Denison 2007)

In the quote above, Denison suggests that the oversimplification of tagging procedures, however useful and necessary they are to a certain degree, reach the end of their usefulness when it comes to tokens which cannot be categorised unambiguously. This is exactly what we found for the non-codified form-function relationships in VOICE data. For the many cases of non-codified form-function relationships which were detected in the manual tagging procedure, it was unsatisfactory to give preference to either the paradigmatic form or the syntagmatic function of the token. The essential problem this poses can be illustrated by the following example:

(22) S10: i just wanted to to give a (1) **partly** answer about the <pvc> baltics </pvc> (.) (POmtg404:924, L1=lav-LV)

The question here was should *partly* in the example above be assigned an adverb-tag (RB), according to its morphological characteristics, whereby the suffix *–ly* marks it as an adverb? Or, should it be assigned an adjective-tag (JJ), according to its syntactic environment?[54] And while in the case of *a **partly** answer*, the function of *partly* as an adjective is fairly clear, there are other examples of tokens where the paradigmatic form seems more unambiguous than the syntagmatic function. Consider, for instance, the token *agreed* in the following context:

(23) S2: so this all the issues are should be de- debated o:n the officials' er level to ***agreed*** between diff- er er of <un> xxx </un> (.) AND er norwegian ministry experts (POprc559:104, L1=lav-LV)

While in (24), the form can be identified as a past-tense form (either VVD or VVN) here, the function is much less clear. It could range from a verb base form (VV) or gerund (VVG), to a

[54] Tagging codes with their corresponding word class categories are given in the appendix.

noun (NN), i.e. meaning *level of agreement*. Examples as these are not isolated cases, but occur frequently in the corpus. In total, we found and annotated 152 different combinations of tag with some kind of discrepancy between paradigmatic form and syntagmatic functions. Given the number of cases we came across, and the difficulty to decide between whether to give preference to form or function, it was necessary to decide on a consistent annotation scheme.

As explained in section 3.2.3.4, other corpora have usually given preference to one or the other in such cases. However, for the tagging of ELF data, neither of the options of giving preference to either the paradigmatic form or the syntagmatic function of the token represented a satisfactory solution. This is because the very nature of ELF data means that one cannot rely on conventionalised form-function correspondences for the description of ELF usage, which has been shown to be essentially variable and in flux, and not necessarily bound to codified L1 use. We considered it our task in the tagging process to make this tension between the different categories visible, which arises from this unconventional language use in ELF. However, we did not intend to resolve this tension by choosing a preference for either form or function. This, we felt, was properly to be done by to the researchers working with the tagged corpus data.

In an initial step, these forms were assigned a tag for the paradigmatic form (form-tag) and one for the syntagmatic function (function-tag). These were tagged in the format **FORM-tag(FUNCTION-tag)**, the form-tag being given first, followed by the function-tag in brackets. In the case of *a partly answer* above, *partly* received the form-tag of an adverb and the function-tag of an adjective, formatted *partly_RB(JJ)*. In a second step, it seemed appropriate to carry out this 'dual' tagging procedure not only for those cases which diverged from codified descriptions of English, but to apply it to *all* tokens in the corpus. Hence, it was decided that where paradigmatic form and syntagmatic function agreed, they would receive identical tags for form and function, and where they did not agree, as in the case of *partly* they would be assigned different tags, as indicated in example (24) below. The tokens *a* and *answer* are assigned the same tag for form and function, whereas *partly* is not (note that PA refers to the tag for pause).

(24) a_DT(DT) _1_PA(PA) partly_RB(JJ) answer_NN(NN) (POmtg404:924, L1=lav-LV)

The dual tagging procedure for all tokens, rather than only for those differing from the descriptions in our reference dictionary, was carried out because we operated within a

framework of ELF: to have applied the procedure of assigning individual form- and function tags only to items which were non-codified would have resulted in a certain 'stigmatisation' of these items, which bears similarity to error-tagging procedures carried out for learner corpora. Such procedures were not regarded as appropriate for the tagging of the ELF data in VOICE (cf. Osimk-Teasdale 2013). The decision to give each token in the corpus a tag for form and function, however, was used to display the flexibility of naturally occurring use of English as a lingua franca.

As elaborated on in section 4.2.2, OALD7 was used for deciding on which forms-functions relationships were (non-)codified. Bearing in mind that not all established, and certainly not all possible, functions for a form are listed in OALD7, the use of only one dictionary ensured a better transparency for the corpus user as to why which tags were applied and enabled the annotators to stay consistent in the tagging process. For some areas where the OALD7 lacks detail, exceptions were made. For example, for proper nouns in VOICE data, of which only a small number are listed in OALD7, the dictionary was only used as a point of reference as long as it made sense. Another example are tokens ending in *–ing*. These tokens are usually verbs, but can also function as nouns or adjectives. They are used highly flexibly in VOICE, especially between the categories noun and verb, so that the entries in OALD7 often did not suffice in describing the examples which were found. It was therefore decided that for these examples, OALD7 would not serve as reference, but that we would adopt a 'data-internal' approach, described in more detail in section 4.3.3. below.

Before proceeding to the issue of ambiguous tag assignment, mention needs to be made of a few aspects to be aware of with regard to the display of non-codified form-function relationships. First, it needs to be noted that the non-codified form-function relationships displayed by this format can concern **intra-categorical** (e.g. within verbs, nouns) cases, as *taste, hate, month, spoke* and *break* in examples (13)-(17) or **inter-categorical** cases (e.g. noun to verb, adjective to adverb etc.), as *complete* and *partly* in examples (18) and (19). Secondly, the categorisation of forms and functions was not always straightforward, meaning that often more than one tag was possible for either. These ambiguous cases were an issue not only for non-canonical form-function relationships but generally relevant to the tagging of VOICE and will be discussed in the next section. And finally, the VOICE Tagset determines which form-function distinctions are displayed and which are not apparent. This means that some distinctions are 'hidden' in the tags and our tags are not an all-inclusive list of form-function differences in VOICE. Differences which are 'covert' in the tagset include, for example, the use of singular determiners in plural function as all determiners receive the tag

DT irrespective of whether they are singular and plural. Hence the difference in form and function for the singular determiner *this* functioning as a plural determiner is tagged 'DT' for both form and function-positions and is not made visible by the VOICE tagset (example (25)):

(25) S1: we_PP(PP) close_VVP(VVP) **this_DT(DT)** lists_NNS(NNS)
(POwsd372:246_19, L1=rum-RO)

Despite these limitations, we propose that the format of dual tagging for paradigmatic form and syntagmatic function is a way to at least partly display the vagueness of syntactic categories that Denison calls for in tagging practices. For example, the phenomenon of mass nouns being used countably, which Dension (2013a: 28) mentions, is made transparent by the dual tagging format in VOICE, as demonstrated by the example (26) below, where the uncountable noun *bread* is used countably with the number *two*:

(26) S3: does_VVZ(VVZ) anyone_NN(NN) wants_VVZ(VV) **two_CD(CD)**
bread_NN(NNS) _1_PA(PA) because_IN(IN) (LEcon8:284_5, L1=kir-KG)

However, while we considered it our task in the tagging process to make the tension between the different categories visible, the dual tagging procedure meant this tension was deliberately left for researchers working with the tagged corpus data to resolve for themselves. This analytical step is left to the researchers working with the tagged corpus data.

4.3.2 Ambiguous tag assignment

The term 'ambiguities' is usually further sub-categorised into semantic and syntactic or structural ambiguities (Bussmann 1996: 19, s.v. *ambiguity*). Disambiguation of syntactically and semantically ambiguous items has been a well-known issue in natural language processing of any type of language data. Part-of-speech tagging in itself inherently involves "resolving ambiguity, either semantic or syntactic, based on properties of the surrounding context" (Roth 1998)[55]. Often, the environment of a token does not provide enough information for taggers to disambiguate between different available tagging categories, which results in errors in the automatic tagging procedure (Schmid 2008: 547). In tagging the spoken, interactive and fluid nature of VOICE, a vast number of cases where encountered for which more than one tag was possible within the given environment, and disambiguation was difficult to accomplish with conventional automatic tagging procedures. Many of the cases encountered in VOICE do not fit the conventional categories of structural or semantic ambiguity but were due to a number of factors which are more typical of spoken, plurilingual

[55] No page numbers indicated in pdf.

language use. Consequently, purpose-built ways needed to be found as to how to deal with these cases in VOICE and a large number of these cases had to be tagged manually. In total, 250 different ambiguity tags were assigned (see Appendix B for a list of these). In addition to this number, we encountered numerous cases for which no ambiguous tags, but a 'sequential tagging' procedure was applied (cf. end of this chapter for a more detailed account of this). Before proceeding to elaborate on the main types of uncertain cases of tag assignment found in tagging VOICE, as well as on the solutions which were developed, it will be necessary to focus some attention to the commonly made difference between the terms 'ambiguity' and 'vagueness'. These terms are commonly distinguished in the literature, as for example in Denison (2013a: 21ff.). Denison (2013a: 21) defines 'ambiguity' in part-of-speech tagging as

> the situation when the hearer/reader cannot be sure which of two or more readings was intended by the speaker/writer but does know that it must have been one or other, and the distinction affects the interpretation of the sentence.

Vagueness, on the other hand, is used by Denison when an item is "underdetermined between […] two classes" (Denison 2013a: 21), such as *fun*, which can be either noun or adjective (ibid.), and therefore "the analysis of the containing sentence is also vague" (ibid.)

The distinction between ambiguity and vagueness is drawn as follows:

> Whereas the producer of an ambiguous sentence must have intended one or other of the possible readings, a vague sentence is syntactically underdetermined for both producer and recipient. Vagueness and ambiguity are quite distinct.
>
> (Denison 2013a: 21)

Although these two concepts are questionable when applied to spoken, naturally occurring language use, in general, *ambiguity* may be understood as meaning that utterances are interpretable in clearly distinct ways, a dual interpretation so to speak, whereas *vagueness* refers to simply indeterminate cases, where no clear interpretation is possible. For tagging VOICE, this distinction would mean that while for ambiguous structures, the person disambiguating, e.g. a human doing the manual tagging in VOICE, can choose between a certain number of options when an item is *ambiguous*, whereas in case of a *vague* item, this is not possible, as they are, in Denison's (ibid.) words "underdetermined" for both the speaker who produced the utterance as well as the manual annotator.

However, it needs to be added that Denison's definition is based on well-formed, written language, as it suggests that vagueness and ambiguity are a matter of sentence structure, whereas in the spoken data of VOICE, dual interpretations or indeterminate utterances are due to other causes. This is why both distinctions, that between ambiguity and vagueness, as well

as that between semantic and syntactic ambiguities, are, in many cases, not suitable for the analysis of VOICE data: First, the line between where one or two readings are possible (i.e. ambiguity) and where such a distinction is not possible, is not unproblematic to draw. Arguably, this might be easier when referring to a group of speakers from a single language variety, where the annotators may be able, to some extent, to rely on their intuition to what is indeed a "possible readin[g]" (Denison 2013a: 21) and what is, in turn, "syntactically underdetermined" (ibid.). This, however, is not possible for ELF data, where speakers converse in a language other than their first, and have a range of language repertoires to draw from. It is impossible in many cases for annotators to say if what is uttered would be a "possible readin[g]" or whether it is "syntactically underdetermined". This is because the part-of-speech annotation is based on the transcription and the speakers cannot be consulted in hindsight as to what it is they meant to say. And even if they could be consulted, it is unlikely that the speakers' intuitions about their own language production would always be accurate. Rather, the possible readings are, in addition to the transcriptions, to a large extent determined by the categories specified in the VOICE Tagset, i.e. an annotator is restricted to which categories are available and choose from these in order to describe a specific ambiguity or uncertainty.

Moreover, although some cases of uncertain tag assignment in VOICE can be said to fit the categories of semantic or structural ambiguities (such as groups 1 and 2, see below), most cases of ambiguity found in the process of tagging VOICE do not neatly fit such categories. They often seem to emerge out of a number of factors which are more typical of spoken, plurilingual language use as found in VOICE, such as the interactivity of the data and the flexible use of language forms and functions. These create quite a different situation than when being concerned with written language, for which such categories have often been developed. In many cases, the rapidly alternating turns in the interactive conversations resulted in short, interrupted utterances and clauses which are often not as 'well-formed' as in written language. The resulting lack of co-textual connection, especially at the **beginning or end of utterances**, as well as **tokens in isolation**, made it difficult to disambiguate tokens semantically or syntactically. Together with the many ambiguities which are characteristic for an inflectionally poor language as English, this resulted in cases such as the one for *adult* in the following extract:

(27) S2: <7> yeah this this wo- this </7> would be nice
S3: and the [thing13] it's more (.) in the: (.) yeah in the the <1> teenager </1> grown-up grown-ups adults more <2> in </2> <pvc> sportive </pvc> er (.)
S2: <1> **adult** </1>

S2: <2><soft> uhu </soft></2>
S3: <3> section </3> actually so maybe this will bring some more new input er and will (.)

(PBmtg463:149-153, S2: L1= scc-RS, S3: L1= ger-AT)

The token *adult* uttered by S2, overlapping with S3 saying *teenager*, could be classified either as a singular noun (NN) or as an adjective (JJ), which are both listed in the reference dictionary for the word. It is obviously a contribution to S3's utterance (mentioning *teenager*, *group-ups* and *adults*). The latter, however, does not provide clear evidence for either tag of utterance 151 by S2. Due to the token *adult* **standing alone** (simultaneous speech between S2 and S3), it is difficult to resolve the ambiguity.

Another position which makes disambiguation difficult is for words standing at the **beginning or end of utterances**, such as the tokens *to* and *do*, in examples (28) and (29), respectively:

(28) S4: this makes it much more (1) erm (1) tricky and maybe it it it it requires much more (.) HIGHER and <un> x </un> level of management (1) than in for instance BUSINESS sector when you gain this whatever win-win situation with the MONEY you <5> know </5> (which) <6> is </6> you just buy <7> of </7>
S1: <5> yeah </5>
S1: <6> yeah </6>
S1: <7> it's more </7> more easy **to** =
S4: = anyone (.) you don't have the win-win situation but HERE you really had to i mean =

(POwgd524:943-947, S4: L1= lav-LV, S1: L1= ger, ita)

(29) S5: <2> exactly </2> but I think this is also what [first name1] [last name1] has referred to as so- something very crucial that HE wants to e:r kind of see: er s- to it that we: come up with e:r WHEN does WHICH consortium <3> and WHICH </3> GROUP have to <4> agree.</4> when do rectors for example write <5> something </5> like a letter of a <6> com</6>mitment <7> or so. yah?</7> (.)
S2: <3> **do** what </3>
S4: <3><soft> (consortium) </soft></3>
S2: <4> yeah yeah </4>

(POwgd317:536-539, S5: L1= ger-AT, S2: L1= gre-CY, S4: L1= est-EE)

In example (28), the token *to* cannot be safely disambiguated between an infinitive marker and a preposition (ambiguous tagging: to_IN(IN)/TO(TO)). In example (29), the token *do* could be either classified as a verb present singular, e.g. assuming an elliptical pronoun, or a verb base form, e.g. if functioning as an imperative, e.g. *do* what you need to do (ambiguous tagging do_VV(VV)/VVP(VVP)). In both cases we are dealing with highly interactive stretches of spoken language, in which turns alternate rapidly and clauses are often not as

'well-formed' as in written language. The syntactic environment does not provide enough information in these examples to solve the ambiguity inherent in the tokens *to* and *do.*

As a classification according to the categories semantic and structural is difficult to accomplish for the all the types of unclear tag assignment found in the interactive VOICE data, this section will attempt a categorisation of the five main types of disambiguation challenges encountered in VOICE, which concerned:

1. Semantic ambiguities between **discourse markers/formulaic iteams and their homonyms** in other word class categories
2. **'Real' structural** ambiguities
3. Ambiguities of **paradigmatic form**
4. Ambiguities of **syntagmatic function** (often incorporating an ambiguity of form as well)
5. Ambiguities due to **unclear sequences**:
    i. form/function tag vs. zero realization
    ii. uncertainty as to where (for which word) to assign the conflicting f/f tag
    iii. generally unclear interpretation of longer sequences

All of these were cases of unclear tag assignment for which transparent and consistent tagging guidelines had to be developed. Before these are discussed, however, one needs to note that while this poses a challenge in the categorisation of the data the speakers do not seem to be disturbed by such instances which 'lack in clarity'. In none of these examples the speakers are trying to dissolve the ambiguities which arise. Hence, tagging is an analytic procedure which is remote from the actual experience of users: Analysts identify ambiguities and other features of language that users just do not notice. The problem with tagging is that in uncertain cases, where word class is not explicitly signalled, guesses have to be made about what is going on in the users' minds.

For the first two groups of types of ambiguities mentioned above, the respective tokens were simply ambiguous between two word classes and tag assignment for form and function position was identical for each ambiguous part, e.g. *like_DM(DM)/IN(IN)*, *joke_VV(VV)/NN(NN)*. Groups 2-5, on the other hand, often involved an assignment of a diverging tag for form- and function position, which added to the complexity of the ambiguities. All five groups contain cases which were frequently encountered in the manual tagging procedure and necessitated strategies for dealing with them consistently. These strategies and guidelines for this will be elaborated on at the end of this section.

The first type of ambiguities found in VOICE concerned semantic ambiguities between **discourse markers** or **formulaic items**. These frequently received an ambiguous tag in the manual tagging procedure when tokens had to be disambiguated from their homonyms in other word class categories. As such, they can probably most appropriately be classified as cases of semantic ambiguity, i.e. word forms which have more than one meaning. The disambiguation had to be carried out for the closed list of tokens of single word discourse markers (*like*, *look*, *whatever*, *well*, *so*, *right*) and multi-word discourse markers (*I mean*, *I see*, *mind you*, *you know*, *you see*)[56]. The classification for these tokens could not be automatically corrected by the tagger, as it was part of the VOICE-related changes to the tagset and tagging classifications. This meant that these tokens had to be disambiguated manually according to their syntactic environment. However, while in many cases such a disambiguation could be carried out without major problems, there were also a number of items for which such a distinction was not possible, as exemplified in (30) to (32) below.

(30) S1: it's nice huh?
S2: <smacks lips> very (1) **well** i think it's RE:Ally a bomb (.) @@@ (4) {speakers are eating (4)} this is the so- the sort of dessert that you give to KIDS when (.) before they go out to play all afterNOON. (LEcon566:265-266, S1: L1=ger-DE, S2: L1=ita-IT)
well: → DM/RB

In (30), two interpretations for *well* are possible: either that S2 says *very well* (with a short pause in between, indicated by the annotation *(1)*), in which case *well* would be tagged as adverb (RB), or S2 says *very*, and then starts a new clause with a discourse marker *well*. The pause, characteristic of spoken communication, of course, makes it difficult in this example to choose between the two interpretations.

The tokens *like* and *so* were especially frequently assigned an ambiguous tag. Out of 2589 tokens of *like* having been tagged as discourse marker (DM), as opposed to, e.g. a verb, 438 tokens (16.9%) were classified as ambiguous (tagged *like_DM(DM)/IN(IN)*). The token *like* was often ambiguous between being a discourse marker (tag DM) or expressing similarity (tag IN), as in example (31). In other cases, the person doing the tagging was generally insecure about what was intended by the speaker, e.g. because the *like* stood alone or without much context in an utterance. Much less, but still a regular issue in tagging was the ambiguous tagging of the token *so*. Of 8354 cases of *so* tagged DM, 228 were assigned an ambiguous tag (2.7%). 152 tokens were ambiguous between a discourse marker and an

[56] For a more detailed account of the category of discourse markers in VOICE POS see Appendix F.

adverb, as in (32), tagged *so _DM(DM)/RB(RB)*, 76 between a discourse marker and a subordinating conjunction, as in (33), tagged *so _DM(DM)/IN(IN)*).

(31) S2: <5> have you noticed his </5> writing? (.) he it's not **like** (.) very (4) (EDcon4:647, L1= rum-RO)

→ **like**: DM/IN

(32) S3: that's (**so** i) = (EDsed364:361, L1=bos-BA)

→ **so**: DM/RB

(33) S2: […] i- if if you sort of er er ADD these two columns you probably er end up with er at the at the same height. right it was just sort of er another way of of spending the money e:r **so** that (.) so this (.) reversal er that's mcsherry. (.) (EDsed301:395, L1=ger-DE)

→ **so**: DM/IN

What can be said for ambiguities between discourse markers and their homonyms in other word class categories is also true for formulaic items in VOICE. Consider example (34), where the first occurrence of *man* can be read as a formulaic item (FI), as well as a noun (NN), according to the context:

(34) S3: it's a ritual <@> **man** @@ man's shaving </@>{SX hits table} (.) (EDcon496:222, L1= ind-ID)

→ **man**: FI/NN

As a second issue, we came across a number of examples of what I might call **'real structural ambiguities'**. There, the context provided was sufficient but due to the structure of the utterance, two or more interpretations were possible. These 'real structural ambiguities' are cases which have also been reported on for written language. An example from our data is given in (35), where the token *joke* could be tagged as noun or verb, as either reading is possible according to the structural environment of the utterance.

(35) S4: tend (.) to (.) maybe USE a maltese word or **JOKE** in malte:se or (.) so we DO know it but my friends are even worse than me when it comes to speaking (EDint330:1091, L1= mlt-MT, eng-MT)

→ **joke**: NN/VV

While the first two groups described above are fairly straightforward, in that the descriptions for 'semantic' and 'structural' ambiguities can (at least loosely) be applied, group 3 to 5 are more complex because the ambiguities and uncertainty of tag assignment also involves non-

ENL form-function relationships. The first of these three groups were ambiguities of **paradigmatic form**, as exemplified in (36), (37), (38), (39) and (40).

(36) S1: i thought he was **take** over the marketing position. (PBmtg27:232, L1=ger-DE)

→ **take** *form*: VV/VVP/NN *function*: VVG

(37) S3: WE tell our companies that are **list** that you NEED to be very strong in your home market (PRpan294:8, L1=ger-AT)

→ **list** *form*: VV/VVP/NN *function*: VVN

(38) S6: because i didn't **heard** anything (PBmtg300:1153, L1=ger-DE)

→ **heard** *form*: VVD/VVN *function*: VV

(39) S24: <smacks lips> some kind of INstrument er to com- communicate with **other**. but (1) i am DO really think that we should think about er the QUALITY of lingua franca becaus:e i think we: WE should speak about <smacks lips> (.) about an INstrument. (.) (EDwsd303:379, L1= slo-SK)

→ **other** *form*: JJ/NN, *function*: NNS (most likely)

(40) S1: <to S2><L1kor> xx xxxx xx xxx xx </L1kor></to S2> (2) okay (.) we're very **interest** in (2) (PBmtg3:3774, L1=kor-KR)

→ **interest** *form*: V/NN, *function*: JJ

(41) S4: = e:r two terms yeah hh and it's li:ke it's counted er both semesters? and er: there is a (**except**) <to S6><L1slo> xx xxx </L1slo></to S6> (2) (EDcon250:878, L1=slo-SK)

→ **except** *form:* VV/VVP/IN, *function*: NN

The morphological forms of the tokens *take*, *list*, *heard*, *other*, *interest* and *except* are in themselves ambiguous, as they can refer to more than one morphosyntactic category. The forms *take*, *list* and *interest* can refer to different verb forms (base form or present tense non-third person singuar) or, in theory, also to nouns. The token *heard* can be used to refer to the verb forms preterite[57] or the past participle, *other* can be interpreted as ambiguous between adjective or (pro)noun form, and *except* is listed as verb, preposition and conjuction in OALD7. In all of these cases, again, the lack of syntactic environment is not sufficient to determine which word classes of these forms are actually intended by the speakers. Moreover, speakers with a variety of first and other languages may not always use morphological forms as indicated in codified, ENL-based language descriptions, which renders a classification of the tokens even more challenging.

[57] Terminology from Huddleston and Pullum (2010: 50).

The fourth group of ambiguities refers to even more complex cases, namely such where the **syntagmatic function** was unclear, often in addition to an ambiguity of paradigmatic form. Examples are given in (42), (43), (44), (45), (46), (47) and (48).

(42) S1: he was asking THEM there he **drive** everybody crazy crazy in this er matter (PBmtg27:247, L1=ger-DE)

→ **drive** *form*: VV/VVP, *function*: VVD/VVZ

(43) S1: because the ten minutes are the ten minutes @ (.) okay (1) everybody **understand** (EDsed362:3, L1=eng-US)

→ **understand** *form*: VV/VVP, *functions*: VVZ/VVD/VVP

(44) S6: this is a problem of er **Accept** decisions of other agenc<2>ies </2> (POmtg404:718, L1=cze-CZ)

→ **accept** *form*: VV/VVP, *function*: JJ/VVG

(45) S6: and then the southerners had the british who were really in control of the aden city only? but then the rest are **protect** area? (PRpan1:26, L1=ara-YE)

→ **protect** *form*: VV/VVP, *function*: JJ/NN

(46) S7: the nation**'s** <13><un> x </un></13> (EDwgd6:597, L1=eng-US)

→ **'s** *form*: VBS/VHS/POS, *function*: VBS/VHS/POS

(47) S2: {S2 talks while eating} **australian** is a <swallows> a goal to be opened (PBmtg300:2382, L1= dut-NL)

→ **australian** *form*: JJ/NP, *function*: JJ/NP

(48) S1: o:h the (could) **sure** you er (2) (PRcon534:187, L1=ger-DE)

→ **sure** *form*: RB, *function*: RB/VV

As demonstrated in the examples above, some of these cases concerned already ambiguous verb forms which, in addition, were ambiguous with regard to other verbal functions, as e.g. for *drive* in (42) and *understand* in (43) or across word class categories, e.g. *accept* in (44), and *protect* in (45). Other cases concerned contracted *'s*, exemplified in (46), which could sometimes not be disambiguated between referring to *is*, *has* or a possessive use. In example (47), *australian* is ambiguous in form and function between an adjective and a proper noun. Example (48) demonstrates a case in which the token *sure* has the form of an adverb (or adjective) but might also function as a (base) verb here. However, the syntactic environment does not provide enough information to give an unambiguous tag.

It becomes clear that these cases, where often both form and function are ambiguous, are even more complex than those where only the form is unclear. Hence, it emerged that some kind of simplification measure had to be taken to prevent the tagging becoming too complex. In

particular, decisions needed to be taken with regard to whether to choose the most likely option out of those available or to leave the function open entirely. The solutions developed are discussed at the end of this section.

Fifthly and finally, we encountered stretches which were difficult to interpret as a whole, which I will refer to as **unclear sequences**. Of these we found three types: The first one were cases for which it was difficult to decide whether the assignment of a different tag for form and function was appropriate or a zero-realization of a preceding word was more likely, as illustrated in examples (49) and (50).

(49) S2: and the reasons are alterNAtive **form** of i:nvestments (PRpan294:20, L1=slo-SK)

→ **form** *form*: NN, *function* NNS/NN (assuming a **zero realisation** of determiner *a*/*the*)?

(50) because when i **looking** for books in gender studies (PRpan225:160, L1=ger-CH)

→ **looking** *form*: VVG, *function* VVP/VVG (assuming a **zero-realisation** of *was*)?

In (49), it is unclear whether the token *form* has singular noun form with plural function (tagging: *form_NN(NNS)*), or whether there is a zero-realisation of a determiner, i.e. follows the interpretation ***an*** *alternative form of investments* (tagging: *form_NN(NN)*). In (50), the ambiguity needs a decision between interpreting the sequence as including a zero-realisation of *was*, i.e. *i* ***was looking*** *for books* (tagging: *looking_VVG(VVG)*) or, alternatively, *looking* as having a verb gerund form but a different verb function, such as verb singular present or past tense, i.e. *when i looked/look for books*, for which looking would receive a tag which is different for form- and function-position, i.e. *looking*_VVG(VVD) or *looking*_VVG(VVP).

Another, related, example to this group is illustrated in (51) below, where the decision is whether or not to include an interpretation of a zero-realisation or not, however not involving a conflicting tag for form and function. The issue, which was especially common in VOICE data was that for the token *got*, in many cases it is impossible to establish if there is an elliptical *have* preceding *got* or not. In the example below, speaker S1 says ***got*** *one from palermo*. This case is especially unclear as S1 might be completing S2's *i've* ***got*** *it here*, which includes *have*.

(51) S2: i've got <fast> it here </fast>
S1: **got** one from palermo (.)
S2: n:o that's too small. (9) {beeping sound of the microwave (2)} (you) <un> x </un> (LEcon566:240-242, S2: L1= ita-IT; S1: L1=ger-DE)
→ **got** *form*: VVD, *function*: VVD/VVN

The two tagging options would have been either to include a zero-realisation of *have*, in which case the tagging would have been a past participle tag for *got* (tagging: *got_VVN(VVN)*). Or, alternatively, to exclude the interpretation of a zero-realisation, which would have resulted in a tagging of *got* with a past tense form (tagging: *got_VVD(VVD)*).[58]

The second type of unclear sequences were cases for which it was unclear **where** in a sequence to assign a conflicting tag for form and function (if any), because more than one reading was possible. This is demonstrated in examples (52) to (55).

(52) S4: **do** you **arrived** there? (.) (EDcon4:98, L1= mac-MK)
→ **tagging**: do_VVP(VVD) or arrived_VVD(VVP)?

(53) S24: <7> so:</7> i don't NEED **that rules** (EDwsd303:892, L1=slo-SK)
→ **tagging**: that_DT rules_NNS or that_DT rules_NNS(NN)

(54) S2: so what would be this other perspective so (.) once again we have the idea i guess of cooperation in the information. is not only the **student** that **have** to **motivates themselves** (.) to: get information. university needs to come <4> TO </4> the **student**. no? (EDwsd499:58, L1=fre-FR)
→ **tagging**: singular (because of *student*, *motivates*) or plural (because of *have*, *themselves*) interpretation for sequence?

(55) S5: <5> (yah there) **is** </5> many chinese **restaurant** i think (.) (.) (PRint595:230, L1=ind-ID)
→**tagging**: is_VBZ […] restaurant_NN or is_VBZ(VBP) […] restaurant_NN(NNS)

In example (52) the diverging form-function assignment could be only for *arrive* (i.e. for the interpretation *do you arrive*) or for both *do* and *arrived* (i.e. *did you arrive*). In example (53), it is not clear whether the interpretation should be *that rule* or *those rules*, i.e. whether the differing tag for form- and function-position tag should be assigned to *that* or *rules*. In this case, this would not have been overt in the tag assignment per se, as there is no differentiation

[58] Another option was to add the information that an utterance is elliptical, as has been done by ICE-GB Greenbaum & Yibin (1994: 41).

in tagging singular or plural determiners. However, cases like this also need to be taken into account in order to be consistent with other cases where the tagset does make such decisions overt. In example (54) there is diverging evidence for either a singular noun (*the student*, *motivates*, *the student*) and a plural noun (*have*, *themselves*). The decision needs to be made which interpretation to give preference to and assign diverging tags for form and function-positions accordingly. Example (55) is another slightly different, not uncommon case in VOICE, as there is actually verb-noun agreement between *is* and *restaurant*, which would indicate no need to assign non-identical tags for form- and function-positions. However, the plural meaning of the noun *restaurant* is indicated by the token *many*. Here the question is whether a conflicting form(function) tag should be assigned to both *is* and *restaurant* even though they are in agreement, on account of the plural indicator *many*.

Finally, other cases encountered concerned whole sequences for which the **interpretation was unclear**, often, presumably because of the structure and nature of spoken discourse, but not clearly related to difference in paradigmatic form and syntagmatic function. Consider, for example, the token *SMALL* (capitalisation signalling stress in VOICE) in the following sequence:

(56) S2: ah okay (2) i've got my own
S1: do you want your **SMALL**
S2: i've got <fast> it here </fast> (LEcon566:238-240, S1: L1=ger-DE, S2: L1=ita-IT)

The sequence was not clearly interpretable as such within the rules of codified, written English. With reference to codified norms, however, there are at least two possible interpretations for this sequence, either *yours_PP small_RB* or *your_PP$ small_JJ* (followed by absent noun). For the interpretation 'yours small' the appropriate tag sequence would be *your_PP$(PP) small_RB*, for the interpretation 'your small' plus absent noun, this would be *your_PP$ small_JJ*. In order to represent both interpretations and their ambiguity, including a form- and function tag for all tokens, the tagging would have to be *your_PP$(PP)/PP(PP) small_JJ(JJ)/RB(RB)*. However, while this tagging illustrates two possible readings, other annotators might find additional, equally plausible, readings, as the syntactic environment is simply not sufficient to be certain of which interpretation might be the intended one. In the annotation process of VOICE, examples as these, where there are multiple possible readings, were considered too indeterminate to be represented in the tagging process with ambiguous tags, as it was always possible that there were missing readings which remained unrepresented. Such indeterminate cases, which involve too much ultimately subjective inference, compromise the transparency of the tagging annotation. Additionally, they add an

unnecessary complexity to both the tagging process as well as for the users of the tagged corpus. Hence, we decided to use a 'sequential tagging procedure' in such cases where more than one interpretation is possible (cf. below).

The fundamental question which arose for all these cases was whether we wanted to avoid registering ambiguities in the tag representation, i.e. make a decision for one tag, or instead, include all tags which are likely. Both options obviously have advantages and disadvantages. While the display of only one tag results in a less complex tagging scheme, it will obviously omit a number of readings which are not represented by this one tag. For VOICE, such complete disambiguation was not felt to be suitable, as it would have simplified and misrepresented the fluid and complex nature of the data. On the other hand, the second option, i.e. displaying all likely or indeed, possible, tags for a given token bears the danger of going overboard with possible tags. With the annotation of difficult cases, there are often as many opinions on the interpretation of a sequence as there are annotators. Also, with spoken English, where many forms are potentially ambiguous but without sufficient clues to be disambiguated, it becomes difficult to decide where to draw the line in terms of number of tag possibilities.

In tagging VOICE, this led to a third option, which was to display and so make apparent the flexibility of spoken language through the variability in tag assignment, but, at the same time, to place restraints on the number of options displayed. This meant that in principle, ambiguities should be displayed, as this reflected the nature of the data, and we did not regard complete disambiguation as a desirable aim. However, the complexity of tag representation should be limited in order to be practicable. This resulted in clearly defined guidelines for ambiguous tags. The tagging of uncertain cases as described in this section was done according to **four guidelines** which meant to achieve this goal of striking a balance between leaving options open while avoiding excessive complexity in tag representation:

First, for each token, the **tag possibilities** were restricted to a **number of two** on each side of an ambiguity-signalling slash, while both tags consist of a form-tag and a function-tag, in the format **TAG1(TAG1)/TAG2(TAG2)**. The tags are indicated alphabetically. Hence, *joke* in example (35) was tagged *joke_NN(NN)/VVP(VVP).*

As a second measure, **'generic' tags** were introduced for nouns (common and proper) and verbs, using the tag symbols N and V, respectively. These were used when a token was ambiguous within these word classes, as exemplified for the form-tag in (57), which is given a generic verb-tag (V), and (58), where the function is assigned a generic noun-tag (N).

(57) S2: […] we're determined we will be communicate_V(VVG) with all of you in the near future (PRqas19:3, L1=kor)

(58) S4: again wha- what basic_JJ(N) do you have (.) (POwgd26:1003, L1=nor-NO)

The generic tags helped to reduce the tag complexity in cases where verbs were e.g. ambiguous between VV and VVP (as in example (49)). Moreover, they served to express that even though a sub-specification could not be provided, something could be said about these items on a more general level, as in some cases we felt it could safely be said that a token had a verbal or 'nouny' character, but were not able to determine how this could be sub-categorised within this word class.

Thirdly, an **unknown tag (UNK)** was added to the tagset. This tag was used for tokens which were ambiguous across more than two or more word class categories, and for cases in which tag assignment was highly uncertain. For example as demonstrated in (45), the token ***protect*** in the sequence *the rest are* ***protect*** *area*, could be interpreted to function as an adjective or noun, and is hence tagged UNK for function, i.e. *protect_V(UNK)*[59].

Finally, a **'sequential tagging' procedure** was carried out for passages in which more than one interpretation was possible. Again, this was done in order to reduce tag complexity of the already complex format of form- and function tag for each token. Such cases pertained to those of ambiguous vs. zero-realisation or ellipsis (examples (49) to (51)), uncertainty with regard to the position for which to assign conflicting tags for form and function (examples (52) to (55)) as well as ambiguity of a whole utterance (example (56)). 'Sequential tagging' meant that when tagging these stretches, we favoured a strictly sequential reading from left to right, and excluding readings which required 'looking back' and interpreting what might have happened earlier. There were reasons for doing this: One was that this procedure helped to avoid too much subjective interpretation, as these cases often gave way to not only two but numerous possible readings, as already mentioned above. A second was that spoken language happens chronologically, and hence 'sequential tagging' was viewed as a valid way of interpreting this type of data. Thirdly, this method was also technically feasible. For example, for example (49) *the reasons are alterNAtive* ***form*** *of i:nvestments* this meant that because of the preceeding plural verb *are* (occurring first in a sequential reading), a function tag for plural noun was given (tagging: **are_VBP(VBP)** alternative_JJ(JJ) **form_NN(NNS)** of_IN(IN) investments_NNS(NNS)). The alternative reading, which would have requested returning to reading the sentence again after reading *form* as singular, and therefore possibly

[59] N.B. the form-tag is a generic verb tag (V), as *protect* can be either tagged VV or VVP.

tagging *are* with a singular function tag instead (tagging: are_VBP(VBZ)) was not carried out.

In summary, it is worth noting that even though these measures made the tagging operationally practicable, the assignment of tags in these uncertain cases can only remain an attempt to depict the actuality of language in use. Also, many cases remained not entirely clear and users of the corpus might well disagree with some of the interpretations given for tag assignment in some examples. This is hardly surprising, given that it is impossible to accurately reconstruct the speakers' intentions and use of word class categories, as has already been mentioned, and that the person carrying out the manual tagging will always be influenced by his/her intuitions. It demonstrates that annotation of language, especially spoken language in use, can only ever be an approximation to 'reality' and serve as a tool. This is something which corpus users should be aware of.

4.3.3 Variability and fixed units

In the tagging process, a number of other items were encountered, which had a high number of occurrences in our data but were not sufficiently covered by our reference dictionary OALD7, and for which it became clear that the reference dictionary was far too restrictive to display the character of the data. These occurred in such vast amounts that they appeared to be a integral part of the data. This high number of occurrences made it unsuitable to use the same tagging procedure as for non-codified forms and non-codified form-function relationships but the find other ways to account for the nature of the data in the tagging process.

This concerned on the one hand flexible use of various productive items, such as **collective nouns,** and **tokens** ending in the **suffix -*ing*.** On the other hand, there were a number of items which seemed to act as **fixed units**, such as **compound nouns**, **multiword items**, **formulaic items** and **discourse markers.** In some cases, these units were not sufficiently covered in OALD7, e.g. the frequently occurring compounds *added value*, *good practice*, *member state*, *youth organization*). In other cases, the fixed units were not listed with the word class category found in VOICE, e.g., some pragmatic markers such as *so* and *like* are not listed with the category discourse marker, interjection or exclamation in OALD7 but only with their traditional word class categories, such as preposition, verb, conjunction, noun, adjective, adverb for *like* and adverb, conjunction, noun for *so*.

However, having drawn a distinction between variability and fixed units, it is necessary to mention that these notions overlap, which is particularly the case for compound nouns, whether we refer to those containing a component ending in *–ing* (e.g. *steering crew*,

forwarding industry) or other compounds discussed below (e.g. *member state*, *youth organization*, *lipschitz domain*). Any type of a non-codified compound belongs to both categories, as the coining of the compound displays both the flexible use of language resources as well as the usage of 'fixed units', being inherent in the category of a compound noun. However, as the category of tokens ending in *–ing* is an example of particularly flexible language use, those compounds containing such a token are discussed under 4.3.3.1. All other compounds, being classified as 'multiword items' in the VOICE Part-of-Speech Tagging and Lemmatization Manual (VOICE Project 2014a) are accounted for under the heading of 4.3.3.2 with other types of multiword items in VOICE.

4.3.3.1 Variability

As outlined above, the two issues are discussed under this section heading are the flexible use of **tokens ending in *–ing*** and **collective nouns**. **Tokens ending in *–ing*** can be used highly flexibly and can function as adjectives, verbs or nouns in sentences. This applies to those tokens occurring alone, as exemplified in (59) and (60) below:

(59) S4: for the **sharing** in responsibilities (POwsd372:570_3, L1= mlt-MT)

(60) S5: there was a law **saying** that er if you're young (POwsd372:653_24, L1=nor-NO)

However, in many cases, the occurrences also affected tokens which were part of a compound noun unit, where we had to decide on the word class of the token ending in *–ing*. In such cases, it is often even more tricky to decide between the word classes verb, noun or adjective, than for the 'isolated' tokens ending in *–ing*.

Because of the flexible usage of the '*-ing* category', it is only natural that a reference dictionary would not cover all of its uses. Another consequence of this flexible usage is that it was often difficult to determine whether a particular token is a common use or whether it was a coinage of a particular speaker in VOICE. This issue is different to that of items marked <pvc>, which can be more clearly recognised and labelled as newly coined words as their forms are not listed in OALD7. For some compounds containing one component which ends in *–ing*, OALD7 lists the individual components, but not the full compound. In other cases, similar compounds are listed. For example, *steering committee* is listed as a compound noun in OALD7, while *steering crew* is not (examples (61) and (62)).

(61) S4: but do you need something like an approval from the **steering committee** before (.) going to the er general assembly with this (.) er proDUCTS ? (POmtg541:20, L1=hun-HU)

(62) S1: <6> er </6> (.) the most convenient (.) way for us AND (1) in accordance what we agreed last time the **steering (crew) steering (crew)** would me would be indeed to omit (.) the entire paragraph (1) (POmtg541:1020, L1=fin-FI)

The question is whether these words, which seem to be used synonymously and occur in a very similar context in VOICE (i.e. preceded by a determiner, followed by verb) should be tagged differently. While *steering committee* would quite clearly be tagged as noun for both components, as this occurs as a noun compound in OALD7, this is less clear for *steering crew*. There, the question occurs whether *steering* should be tagged as noun or gerund verb. In this particular case of *steering committee* vs. *crew*, one could argue that as *steering* on its own has an entry as noun in the reference dictionary, *steering* could be tagged noun in both cases, and hence no problem would occur. In any case, a question which also poses itself is whether the different tagging of *steering committee* and *steering crew* would be a legitimate practice for tagging ELF data at all. After all, both compound expressions are used synonymously and constitute equally possible ways of describing the same position. The crucial difference is that one happens to be codified in the reference dictionary, whereas the other does not. And while it could be argued that this is similar to cases annotated <pvc>, which constitute equally legitimate exploitations of the English language, the case is slightly different for the large amount of specialised terminology, for which the tagging is even less clear.

For example, the speakers in VOICE often use words which are specific to particular topics or speech events, and hence are difficult to be covered by a specific (or, in fact, any) dictionary, such as *cloning research* in example (63) below.

(63) S8: (i wish my colleague [S2] [S2/last] would know) i just would like to know if there is any kind of er of new theory combining all kind of <un> xx </un> ethical approaches re- reproductive me- medicine (2) the (.) <pvc> forbiddenness </pvc> or the allowance of cloning and **cloning research** for austria is the first country to deny it? (.) (PRpan13:40, L1=ger-AT)

For *cloning* there is no entry in OALD7 at all, and the entry *clone* is only listed as with the word classes verb and noun, meaning that the only possible tag which could be assigned to *cloning* in VOICE would be that of a verb (derived from the entry of *clone* as a verb), when it is clear from the co-text that a noun-tag would be more likely. For such specialised terminology, the OALD7 proved to be much too restrictive.

Given the high number of seemingly newly coined compounds within the highly flexible *–ing* category, many of which used as part of a specialised terminology, and the reference

dictionary not being sufficient for the classification of such items, we were confronted with a number of highly frequently used items for which there was no point of reference to refer to. This made it difficult to decide on a classification for many of these cases, as it is not often clear what the speakers' intentions were. In some cases, this could be decided from the context, even when the annotators were unfamiliar with the degree of codification of an item, as in *cloning research* above. In other cases, this was more challenging, as demonstrated by the examples below:

(64) S1: […] but then there are two other aspects which you just outlined (.) the first one is that we have different systems in place in: the various **participating countries** so in some cases it is more or less automatic (.) once you have a bachelor you move on to the master (1) (POwgd243:386, L1= ger-DE)

(65) S9: […] there are (.) perhaps two **conflicting tendencies** one is (.) to make (1) a few spots (.) at the european surface which should (.) take (1) the (.) future activities {S1 puts down his coffee cup, making noise (1)} (POwgd14:1149, L1= scr-HR)

(66) S1: it takes long time i mean look to er the same we have in the **forwarding industry**. (.) [org45] [org7] [org46] <9> how </9> long did it take er [org7] and [org46] to come together (1) (PBmtg300:3169, L1=ger-DE)

For example should the token *participating* as part of *participating countries* in example (64) be classified as an adjective, a verb or, if it is part of the compound, noun? Similarly, for *conflicting* in (65) it seems clear that this is not a noun, but whether it is a verb or an adjective is difficult to decide. In OALD7 *participating* and *conflicting* are only listed as verbs. And finally, *forwarding* in *forwarding industry* in example (66) is tricky to classify as it might refer to a specific terminology unknown to the person doing the manual annotation or checking. And while *Freight Forwarding Industry* is a commonly used specialised terminology, the phrasing *forwarding industry* is not listed in the OALD7 or other sources which were consulted. OALD7 lists *forwarding address* as a compound noun, and the entry *forward* is assigned the word class categories noun, verb or adjective. However, as *forwarding* on its own is not listed at all, it means that the only possible tag which could be assigned in VOICE would be that of a verb (derived from the entry of *forward* as a verb), when it is clear from the context that a noun would be more likely.

In total there were 14 011 tokens ending in *-ing* which could be either classified as nouns, adjective or verbs, for which the OALD7 was too restrictive and where a tagging solution had to be sought. It seemed reasonable to treat those forms uniformly to a certain degree. Additionally, as little information as possible should get lost. In contrast to for example the

procedure adopted in LOB, where all such forms were tagged according to function (see section 3.2.3.4), we therefore chose to adopt a very open way of handling these cases by which the tagging was strongly based on the immediate environment of a token ending in *-ing* in a manual tagging procedure. However, in order for the tag assignment to remain traceable, a tagging procedure was carried out in which the form- and function tag differed, as this would allow more flexibility for search options. In summary, it was decided that tokens ending in *-ing* would receive a **FORM-tag VVG**, referring not to gerund or present participle verb, but generically to all forms ending in *-ing.* The function-tag (noun or verb) was assigned manually according to context, as demonstrated in examples (67) and (68).

(67) cos basically **moving_VVG(VVG)** on from (POwsd372:7, L1= rum-RO)

(68) to get **housing_VVG(NN)** is to (EDcon521:1107, L1=dut-NL)

As it was often particularly difficult to decide between verbal and adjectival use of a token, the only options for function-position were either noun or verb-tags. The reference dictionary OALD7 was used solely for adjectival use of tokens ending in *–ing*, such as *amazing*, *interesting*, *overarching*, *ongoing* etc. (cf. VOICE Project 2014a: 19f.), meaning that all items which were listed as adjectives in OALD7 could receive an adjective-tag, as we considered them to be conventionalised. Other exceptions were of course tokens such as *thing*, *wing*, *something*, *nothing* etc. which were not tagged according to the procedure described above for tokens ending in *–ing* but, for example, as nouns. Again, as with all tagging guidelines, there will be cases for which it will seem less appropriate to choose a certain procedure. With the tokens ending in *–ing*, there are a number of cases which are tagged VVG(NN), which are very 'nouny' already, such as *feeling*, *funding*, or *handwriting*, to name just a few. Because the tag VVG, which is normally used for verbs, is also for these tokens established as nouns (e.g. listed in OALD7 as such), which end in *-ing*, it was decided to extend the meaning of the tag VVG "for any word ending in the morpheme *-ing*" (VOICE Project 2014a: 19).

A second issue were **collective nouns**, which were also used more flexibly than in codified descriptions of English. According to the Penn Guidelines, a collective noun is tagged according to whether this is followed by a singular or plural verb (Santorini 1991: 18). However, no list of collective nouns included is given there. The reference dictionary OALD7 did not always provide the information on whether a noun could be used collectively or not either. Hence, in addition to these two points of reference, other grammars were consulted. Carter and McCarthy (2006a: 895) define collective nouns as "[a] type of noun referring to a group of people, animals or things" which "can be followed by either a singular or plural verb form" (2006a: 347). They list 14 such collective nouns, e.g. *team*, *government*, *public*,

community, *committee*, *university* and *company*. Quirk et al. (1997: 316f.) give a more extensive list of 45 collective nouns which can occur with either singular or plural verb.

In VOICE, many of the occurrences matched the lists given in the works of reference named above, such as the collective use of *university*, *government*, *company* (anonymised as *org1*) or *middle class*, as exemplified in (69), (70), (71) and (72).

(69) S3: <6> my </6> my **university have** (.) er sent a FAX ? (EDsve423:33, L1=fre-FR)

(70) S2: but the thing is what i was thinking about the counter-arguments they can give us for THIS (.) module is (.) is that **government pay** and pay and pay <8> and </8> that's that's not realistic neither (.) (EDwgd241:927, L1=ger-CH)

(71) S5: and THEN time by time (.) the project will be approved IF (.) this er project complies with all the (.) regulation that er we are setting NOW (.) i think that er IF (.) **[org1] succeed** (and **do**) this <@> it </@> will be really (a) su<3>ccess </3> (.) (POmtg316:300, L1=ita-IT)

(72) S3: er okay you (.) if you don't want er but i think (1) er the **middle class want** to use internet too (.) (EDwgd241:1029, L1=lav-LV)

However, we encountered also uses of collective nouns which were not listed in the grammars and the reference dictionary. Especially prominent were collective uses of nouns referring to countries and places in general. These were often used with plural verbs. Quirk et al. (1997: 318) notes that nouns labelling countries are only used collectively if they refer to sports teams representative of that country. However, in VOICE the occurrences found do not refer to sports teams. Examples (73), (74), (75), (76) and (77) demonstrate this collective use of place names, such as *state*, *czech republic*, *town*, *venice* and *denmark*.

(73) S5: <3><soft> i mean people </soft></3> the (.) **state know** there will be a (.) (EDwgd6:231: ger-AT)

(74) S8: like i stated that (.) ONE e:r country it was **czech republic** just (.) **have lost their** national identity because of the globalization process they adopted (.) er english really deeply to their (.) er culture? (.) {S3 starts writing on blackboard (22)}<5> and </5> they become kind of (.) (EDwsd304:465, L1=rus-RU)

(75) S7: = <4> it is </4> not so difficult because (.) **each town want** the the money stays in the town. (EDwsd499:901, L1=scr-HR)

(76) S1: **venice** er live with tourism er (.) and is a <pvc> touristic {touristy} </pvc> city = (LEcon405:43, L1=ita-IT)

(77) S3: and then <2> **denmark** give points (.) some points </2> (LEint554:4, L1=spa-ES)

As most of the cases found in VOICE were in line with what was found in the points of reference, the decision made with regard to tagging was a loose application of the examples given in Carter and McCarthy (2006a: 347), Quirk et al. (1997: 316f.) and the OALD7 to VOICE data, meaning that if the examples listed there could be directly or indirectly applied to the collective nouns found in VOICE, they would receive a regular tag. For example, the verb *have* following *european union* in example (78) is tagged with a regular tag for present tense singular (*have_VHP(VHP))* as *european union* could be related to the example of ***commission***/*board*/*association*/*corporation* in Quirk et al. (1997: 316) or to *board*/*committee* in Carter and McCarthy (2006a: 347).[60]

(78) S2: = y:es. er especially since the **european union** <2> **have** </2> accepted it as an official language. (.) (EDint330:599, L1=mlt-MT, eng-MT)

The collective use of place names in the VOICE data, which was not covered by the reference works, was treated differently as it was something which appeared to be particular to the data at hand. In contrast to the forms ending in *–ing*, this concerned not a general 'fuzzyness' of categories, which caused difficulties during the categorisation, but a specific use which could be more clearly marked out. It was felt that this specific issue was more similar to those cases receiving a differing tag for form and function-position. Hence, it was decided that this specific use of collective nouns was to be tagged in line with other cases differing in paradigmatic form and function, and, thus, to apply a differing tag for form- and function-position to the verb position. For example, *want* in *each town want* in (75) is tagged as a generic verb for form (see section 4.3.2), and as a 'present third person singular' verb in function (tagging: *want_V(VVZ)*).

4.3.3.2 Fixed units

The second issue addressed here are fixed units, such as **compound nouns**, **multiwords**[61], **formulaic items** and **discourse markers**. In the tagging process, we felt that some token-sequences acted as units (henceforth also referred to as '**multiword items**') and that these should be tagged as such, as these were rather different from their homonyms which occurred

[60] Note that in VOICE, the tagging of the noun did not change (collective nouns are tagged singular) as in Penn, but only the tag of the verb changed.

[61] Note that in VOICE a difference is made between the term 'multiword item', which refers to all items which act as a unit (e.g. also multiword discourse markers) and 'multiword' which refers to a "**closed list** [of] the most frequent multi-word chunks from the word classes **Adverb, Adjective, Conjunction** and **Preposition** in VOICE" [original emphasis] (VOICE Project 2014a: 31).

separately. So for example, while the sequence *more or less* in (79) would be tagged *more_JJ or_CC less_JJ*, in the sequence in (80) *more or less* seems to act as a modifier of *automatic* and hence it seems more appropriate to treat this as an adverbial unit.

(79) S3: […] i: DON'T think and i will try to discourage [first name7] try to work upon <spel> c p ss </spel> on the country level. because i don't think it's necessary. (1) i mean (.) i think that every country will do: (.) **more or less** (.) (PBmtg269:921, L1=pol)

(80) S1: […] (.) the first one is that we have different systems in place in: the various participating countries so in some cases it is **more or less** automatic (.) once you have a bachelor you move on to the master (1) (POwgd243:386, L1=ger-DE)

While OALD7, which was useful for most other issues as a point of reference, does contain some of these units, the number of those listed was far from covering the full variety of fixed units encountered in VOICE. Relevant issues with regard to token-sequences acting as units were what to include in the definition of a multiword item, which works of reference to consult and how to tag such items.

Let me illustrate this problem by the example of **compound nouns**. In the transcription process, those compound nouns for which the individual parts were listed in OALD7, were not tagged <pvc>, even though they might have been newly coined compound nouns. We encountered a high number of instances of token combinations which we felt acted as compound nouns but were not listed in OALD7 as units, such as e.g. *member state*, *voluntary service* or *youth organization* but also a number of proper names such as *european youth forum*, *erasmus mundus*, *lipschitz domain* or *jungle book*. While OALD7 does have a large number of compound entries, these only cover a small part of the compound cases in VOICE. The problem which emerged was, therefore, that the reference dictionary was much too restrictive to cover the use of compound nouns in VOICE. At the same time, based only on the given co-text, it was sometimes difficult to decide whether two sequential tokens actually acted as a compound, or, whether this was e.g. an adjective followed by a noun. For example, should the frequently occurring instances of *added value* (37 times), *good practice* (31 times) or *joint program*(s) (99 times) be included as compound nouns or be tagged adjective followed by noun?

In order to solve the issue of what to include as a multiword, it was decided that, as OALD7 seemed too restrictive, other references should be consulted in addition. However, the use of other dictionaries was not considered an option, as this would have potentially resulted in

inconsistencies with other tagging decisions for which the OALD7 was used as a reference. And in any case, it was highly unlikely that the use of one or more other dictionaries would have covered more compound nouns in VOICE: many compound nouns are very specific to the speech events and their topics, containing much special terminology. Others are likely to originate from the speakers' exploitation of their individual language repertoires, resulting in the creative use of language. Another reason not to consult other dictionaries was that these often do also not list multiword items, especially those of a spoken nature, i.e. multiword discourse markers and formulaic items. However, other corpora such as the BNC and ICE-GB have also been using multiword items in their tagging procedures and these corpora include spoken data. For fixed units, these corpora and their descriptions thus, presented an additional point of reference to OALD7. Hence, in addition to using all multiword items listed in OALD7, the BNC was consulted for a more comprehensive list of multiwords (adverbs, adjectives, conjunctions and prepositions, (cf. VOICE Project 2014a: 31). The categorisation for formulaic expressions from ICE (cf. Nelson 2005: 13) was used as a starting point for the VOICE list of formulaic items, consisting of single tokens and token sequences (VOICE Project 2014a: 30). For the definition of discourse markers (single and multiword alike), we consulted relevant literature (e.g. Aijmer 2002, Lenk 1998, Müller 2005, Schiffrin 1987) on the topic, as the tagging of discourse markers is often not discussed in detail in the tagging literature. In addition, n-gram lists from our own data were used in order to determine those tokens which most frequently occurred in sequence. This is based on the assumption that frequency is an indicator for grammaticalisation (e.g. Mair 2004: 125) and hence for the 'unit-ness' of token-sequences. These frequency lists were used for all 4 groups of multiword items in VOICE (compound nouns, multiwords, formulaic items and discourse markers). The definition and closed lists of multiword discourse makers and compound nouns in particular were heavily based on the most frequently occurring bi- and trigrams, in addition to using OALD7. The closed list names all multiwords and is found in the appendix of the VOICE Tagging and Lemmatization Manual (VOICE Project 2014a: 31ff.).

Another issue was the **tagging format** of multiword items. Other corpora, such as ICE-GB, have used 'ditto tags' for "compound expressions […] [which] function grammatically as single units" in a way that "[e]ach word in the expression is assigned the tag of the expression as a whole" (Nelson 2005: 3). This means that each token is assigned a separate tag, which is identical for all parts of the compound. Numbers are used to indicate of the number and position of the individual parts of the compound, as demonstrated in example (81) below (example taken from Nelson 2005: 3).

(81) The N(prop,sing):1/4
Duke N(prop,sing):2/4
of N(prop,sing):3/4
Norfolk N(prop,sing):4/4

For the tagging of BNC on the other hand, a single tag was given to the whole multiword expression, rather each token being assigned a separate tag (Leech & Smith 2000, subpage: Guidelines to word class tagging), as demonstrated in (82) (ibid.)

(82) <w PRP>according to (preposition)

For VOICE, this meant two options for tagging compound nouns were essentially the two formats exemplified in (83) and (84) below:

(83) steering_NN1 group_NN2

(84) steering_group_NN

The format in (84) displays the compound as more of a fixed unit, giving one single tag to both parts of the compound, while in (83) a looser connection is displayed by separate tags for both noun components, signalling the connection by numbering. For VOICE, this looser format was chosen as the preferred option in the tagging process, as this left more options open with regard to tagging procedures, e.g. transfer to the format displayed in (83), if this had seemed useful at a later stage. Moreover, for all multiword items but plural nouns (see below), this option meant that all components could be searched individually as well as in combination. Finally, this format was also in line with the general approach adopted in VOICE to display the fluidity of the data. So while these compounded units were considered to be more in connection than other tokens occurring successively, it was felt that the format in (84) would place a too strong emphasis on the fixedness of the unit.

The tagging of multiword items was therefore as demonstrated in the examples below for formulaic items (85), multiwords (86)-(89), multiword discourse markers (90) and compound nouns (91)-(94).

(85) i_PP(PP) 'm_VBP(VBP) sorry_JJ(JJ)

(86) vice_RB(RB) versa_RB(RB)

(87) it was not quite up_JJ(JJ) to_JJ(JJ) date_JJ(JJ)

(88) even_IN(IN) though_IN(IN)

(89) out_IN(IN) of_IN(IN)

(90) you_PP(DM) know_VVP(DM)

(91) critical_NN(NN) mass_NN(NN)
(92) youth_NNS(NNS) organizations_NNS(NNS)
(93) decision-making_VVG(NN) process_NN(NN)
(94) living_VVG(NNS) conditions_NNS(NNS)

Generally speaking, the tagging guideline adopted was to give an identical tag to each component of a multiword item unit, both for form-tag and function-tag, as exemplified above. For multiword discourse markers (example (90)), however, only the function tag received an identical (DM)-tag, as it was considered important that at a later stage, these items could be searched for under their conventional as well as the discourse marker category. Also, it needs to be noted that compound nouns were tagged according to the head noun, meaning that in (92) *youth* is given a plural noun tag (NNS). The same is true for compound nouns with one token ending in the suffix *–ing*, as in (94). As already discussed in section 4.3.3.1, compound nouns ending in *–ing* were tagged differently to other compounds. Because of the inherent flexibility of the form *–ing*, for these items no external points of reference, such as dictionaries, grammars or lists provided by other corpora were used for tagging, but the immediate co-text served as only indicator for the tagging decision. Those compounds with the first word ending in the suffix *–ing* were seen as an 'open list' and were not included in the VOICE list Compound Nouns (a closed list, based on the OALD7 and VOICE n-grams) in the VOICE Tagging and Lemmatization Manual (cf. VOICE Project 2014a: 33ff.).

In conclusion, it can be said that for those items where the reference dictionary did not provide sufficient guidelines, individual solutions were sought. In general, the aim was to strike a balance between an internal and an external approach to the data, meaning that both other external points of reference, e.g. codified descriptions of English or other corpora, as well as corpus-internal information, such as the immediate co-text and n-gram lists were used for the use of both flexible language use as well as fixed units. However, this was done to different degrees. This means that while for the use of the very flexible group of tokens ending in *–ing*, mostly internal information was used, for other issues, such as multiword, closed lists were compiled which also heavily relied on data-external sources, such as lists provided by other corpora and codifications. The purpose of using combination of data-internal and –external sources was to stay compatible with other corpora and codifications of

English language use, but, at the same time, to acknowledge the specific character of the data at hand.

4.3.4 Other issues arising from tagging transcribed data in VOICE

In addition to the issues discussed in the previous sections, we also encountered a number of challenges which arose from the fact that the tagging was based on the already transcribed data. In contrast to some of the issues elaborated on earlier, the implications of these issues were more of a practical, rather than a theoretical nature. Despite this, as well as the fact that some of the aspects discussed below did not occur with high frequencies in the data, they are worth mentioning as they can, e.g. illustrate how decisions taking in previous annotation processes have an impact of the tagging and how these can consequently be handled in the tagging procedure.

In this section, **four groups** of issues will be discussed:

1. Formats which were not easily displayable in the tagging process, e.g. round and angle brackets
2. Forms which were treated non-uniformly in the transcription process, i.e. combining forms and affixes standing alone
3. The format of transcribed spoken language, e.g. resulting in co-dependencies over long stretches of text
4. Reliance on the original transcripts

4.3.4.1 Display of formats used in the transcripts

The format in VOICE Online and VOICE XML and VOICE POS differs in a number of aspects. Two aspects which are displayed in the untagged – but not the tagged – versions are round brackets, signalling uncertain and partly uncertain speech (examples (95),(96)) and pointed brackets, which were used to signal a variety of information, e.g. signalling overlaps, non-verbal feedback, non-English language use, spelling etc. (cf. VOICE Project 2007b). In some cases, the information contained in pointed brackets was displayed differently in the part-of-speech tagged version, as demonstrated in Table 2 (VOICE Project 2013a).

Table 2: Annotation differences between VOICE Online and VOICE POS Online (VOICE Project 2013a)

	Mark-up VOICE Online	Mark-up VOICE POS	POS Tag	Search query in VOICE POS
Anonymization	[S1], [org1], [place1], [last name1]	a_ […] e.g. a_[S1], a_[org1], a_[place1]	NP, NPS	Token: **a_*** POS: **a_*,NP***
Non-English speech	e.g. <L1scc> </L1scc>, <LNfre> </LNfre>,	f_*token* e.g. *f_dire*, *f_wir*	FW (=Foreign word)	Token: **f_*** POS: **FW**
Onomatopoeia	e.g. <ono> kr </ono>	o_ kr	ONO	Token: **o_*** POS: **ONO**
Pause **(up to approx. 0,5 seconds)**	(.)	_0	PA	Token: **_0** POS: **PA**
Pause **(approx. 1,2,3, … seconds)**	(1), (2), (3), …	_1,_2, _3, …	PA	Token: **_1,_2,_3**, … All Pauses: _* POS: **PA**
Pronunciation variations and coinages (PVC**)**	<pvc> </pvc> e.g. <pvc> consumal </pvc>	p_*token* e.g. *p_consumal*	PVC(FUNCTION-TAG) e.g. *PVC(JJ)*	Token: **p_*** POS: **PVC**
Spelling out	<spel> </spel> e.g. <spel> e u </spel>	s_*token* e.g. *s_eu*	Various, e.g. CD,NN,SP	Token: **s_*** POS: **s_*,TAG** e.g. **s_*,CD**

In some cases, however, merely using a different format of representation was not an implementable solution. These cases concerned **uncertain** and **partly uncertain speech** (examples (95) and (96)), **partly unintelligible speech** (example (97)) and **partly foreign speech** (example (98)).

(95) S3: <soft><2> **(yes)** </2> **(please)** </soft> = (PRint30:4, L1=ita)

(96) S5: we have a commitment to to the target we have a variety of **instrument(s)** (PBpan25:12, L1=rum-RO)

(97) S12: but then may i ask <soft> (perhaps i missed) <un> x xx xx xx xx </un></soft> (.) do we have any kind of the basic **<un> xx</un>tion**. (POmtg314:941, L1= scr-HR)

(98) S2: <LNger> **boden</LNger>sea** (LEcon8:305, L1= kir-KG)

All items highlighted in bold in the above examples were displayed without the information contained in the round and angle brackets in VOICE POS, e.g. as *yes please*, *instruments*, *xxtion* and *bodensea*. With regard to the tagging procedure, the question arose which tags these items should receive. For full words which where **unintelligible**, it was decided that these would simply receive the tag which would normally be assigned, irrespective of the degree of certainty, e.g. the (yes) in (95) would be assigned RE, just the same as tokens which

had not been labelled uncertain in the transcription process. However, in cases of **uncertainty** and examples of **partly unintelligible** or **foreign** language usage, the question of the appropriate tag assignment was more difficult to solve. For example, should *instrument(s)* in (96) be annotated with a noun singular tag, as this is the certain transcription, a noun plural tag, taking into account the uncertain transcription in round brackets, or even two tags, signalling ambiguity? Cases of partial unintelligible items, such as *xxtion* in (97) present an even greater challenge, as the first part of the word was unintelligible and hence the tag for the token would have been a result of even greater guesswork. Here, the decision was to either give a tag for unintelligible speech (UNI) or to guess the tag based on the information which was available. However, the information available is ambiguous in itself, as *tion* could of course refer to different word class categories, even though, taking into account the broader co-text, *any kind of the basic* it would probably be most likely that a noun was to follow. Finally, for the partly foreign item the question was whether to give just one tag (either FW for foreign word or NP for proper noun) or two tags signalling ambiguity.

For all these items, the aim was a solution that was both practicable and consistent. Although in principle, ambiguous tags would have been an option, i.e. because the transcriber had signalled that more than one option was possible by marking a token uncertain or unintelligible, such tags were ruled out for these cases as the ambiguity was different from other cases of ambiguity described in section 4.3.2. While in the cases there ambiguity arises because the transcript can be given more than one reading, here it comes about because of the shortcomings of the transcript itself (due to lack of understanding the audio file). Hence, the ambiguous tagging of items is unlikely to be of the same relevance to corpus users as other items for which one or more readings were possible. In addition to that, the ambiguity inherent in the transcription was invisible in the tagged version as the brackets signalling uncertainty were not displayed, and hence, ambiguous tags would have been a potentially confusing representation of what looked like a single token for users of VOICE POS. Therefore, ambiguities were ruled out as an option for these cases and a less complex solution was sought.

In order to maintain a certain level of consistency, the same approach was chosen for **partly uncertain items**, as had been adopted for items which were fully uncertain: The brackets signalling uncertainty were ignored for the sake of the tagging procedure, and hence the only tag that was given was that which conformed to the interpretation without brackets, i.e. plural noun for *instrument(s)* (displayed as instruments_NNS(NNS)). For partly unintelligible tokens it was decided that the amount of information inherent in the syllable was too little as

to assign a tag with sufficient certainty.[62] Therefore, tokens such as *<un> xx</un>tion* received a 'unintelligible' tag (tagging: xxtion_UNI(UNI)), which is the same as those tokens which were fully unintelligible, e.g. xxx_UNI. Finally, for the partly foreign token in (98), it emerged that this was in fact an isolated case, for which an individual solution had to be sought. This was treated as two separate tokens (tagging: f_boden_FW(FW) sea_NN(NN)).

While realising that such differences across different versions of the annotation is not an ideal scenario, in the annotation of VOICE, the issue emerged in order to avoid that the annotation of different parts of the corpus (e.g. the XML version vs. the POS tagged version of VOICE) are too strongly interwoven with one another. As overlapping hierarchies are an inherent problem of any XML resource, the POS tag of VOICE was annotated as a stand-off annotation, while the previous versions were annotated inline. While this means that the different annotations of VOICE are spread over different versions of the corpus, they could theoretically be combined to a single resource in a stand-off annotation (Majewski 2015). In any case, for VOICE we placed a high priority on making the differences in annotation apparent in the documentation files as users of VOICE can and will need to work parallel with different versions in order to access the full information given in the different annotation schemes.

4.3.4.2 Non-uniform treatment of certain forms in transcription: combining forms and affixes

Another issue which was encountered was that certain forms had not been treated entirely uniformly in the transcription process. This concerned especially so-called 'combining forms' and affixes occurring in isolation. For these forms, no suitable tag was available in the Penn Guidelines.

For the first group, which were labelled '**combining forms**' in OALD7, the question which arose was whether to add a tag 'combining form' to the VOICE Tagset. In order to assess this possibility, a list of all combining forms in OALD7 was generated and cross-checked these against VOICE 1.1. Online. This list contained forms such as *–crat*, *audio-*, *bi-*, *mid-*, etc. It emerged that out of 122 different entries for combining forms OALD7, only 6 types occurred in VOICE 1.1. in isolated position, i.e. without attachment to a noun or adjective in VOICE, and of these there are only a small number of instances each. These concerned the tokens of *vice*, *mini*, *mid*, *micro*, *macro* and *iso*.

[62] This was also found for the tagging of ICE-GB. Greenbaum and Yibin (1994: 37) argue that even when the intended meaning of a partially transcribed word can be traced, "it would be inappropriate to tag the word partial by reference to the reconstructed word" and that "[i]n other instances it is impossible to say what the intended word should be".

Most of the other forms which did occur were attached to a noun or adjective by hyphenisation or incomplete words and hence received the tag in accordance to the attached word, e.g. PVC for *quasi-accredited* in example (99), or XX for incomplete words as *anglo-* in example (100).

(99) S14: that WILL require of rectors to agree (.) that these joint programs are (.) <fast> sort of </fast> (.) <pvc> **quasi-accredited** </pvc> by the network (2) (POwgd14:845, L1=ger-AT)

(100) S2: […] an **anglo-** (.) -american (.) er point of view (3) (PRpan225:4, L1=ger-AT)

The instances where combining forms occurred in isolation where either due to compounding, as the VOICE transcription rules (VOICE Project 2007d: 25) imply that in general, compounded nouns were spelt separately, while adjectives are spelt with a hyphen. An example of such compounding with an isolated combining form is given in (101).

(101) S3: […] (.) colleagues were talking about (and) (i'm) concentrating more on the **micro** (1) study case (PBpan581:18)

Some isolated 'combining forms' were due to inconsistencies in the transcription, i.e. compound nouns with combining forms were sometimes annotated with or without hyphenisation, e.g. *semi finals* vs. *semi-final* (both used as nouns), sometimes with or without pvc-tag, e.g. mini festival vs. mini-fantastics vs. *<pvc> mini-dispenser </pvc>*, as exemplified in (102)-(104) below.

(102) S1: <5> yeah </5> (1) there was gig at (market) fes- (.) (market) **mini fest**<6>**ival** </6> thing (.) (LEcon545:725, L1= eng-GB)

(103) S3: <7> the **mini-fantastics** </7> (PBmtg3:3721, L1=ger-AT)

(104) S4: cos <5> it's a SMALL </5> dispenser it's a **<pvc> mini-dispenser </pvc>** (PBmtg3:3740, L1=ger-AT)

Adding to these non-uniformities in the transcription of these combining words was the fact that all but one (i.e. *mini*, see below) of these items were additionally listed under a different word class category than that of a 'combining form' in OALD7. For example, *vice* was listed as noun or combining form.[63] This class of 'combining forms' showed to be too diverse, both in the way it had been treated in the transcription process, as well as in the way it was listed in OALD7, to be treated as one uniform group which can be assigned a 'combining form' tag.

[63] N.B. However, the other word classes assigned in OALD7 to those 'combining forms' in VOICE did not always match the meaning. So even though *vice* was listed as a noun (i.e. "evil or immoral behaviour" cf. OALD7, this was not in the meaning that it occurred in VOICE, namely as e.g. *vice governor*.

As this issue only concerned a small number of cases (around 30 tokens) for which tagging decisions had to be taken, it was decided that individual solutions for the respective tokens would be sought. In some cases, where similar cases had been annotated with <pvc> in the transcription, we unitised these, so that e.g. *mini-festival* and *mini fantastic* were also hyphenated and tagged PVC. In other cases, where combining forms were without hyphen, e.g. as compounds, these were tagged according to other word class categories where this was in uniformity with OALD7, e.g. noun for *macro* in *macro issues* (tagging: macro_NN(NN) issues_NNS(NNS)).

For those forms where the only category indicated in OALD7 was that of the 'combining form', which did not exist in the VOICE Tagset, or a category which did not match the examples in VOICE, individual solutions were sought. This concerned especially the forms *mid* and *mini*. For *mid*, OALD7 lists combining form and preposition as word class categories. However, the examples occurring in VOICE did not match the category of preposition, but had a rather noun-like character, as can be seen in (105). In this case, it was decided that, since OALD7 did provide a word class category which was an available tag in the VOICE Tagset, i.e. preposition this should be dealt with as other cases where a different category was available in OALD7 than that with which the token occurred in VOICE, namely to give a different tag for form- and function-position. *Mid* in example (105) was therefore given preposition for form and noun for function (tagging: *mid_IN(NN)*).

(105) S1: <6> so:</6> we should get the numbers like (.) **mid** or end of april something like that and then we should see the first time this <soft><un> xx </un> (for a year) </soft> (PBmtg280:23, L1=ger)

For the token *mini* the category combining form is the only one listed in OALD7. As mentioned above, this was, in fact, the only token for which no alternative word class categories were listed. For *mini*, first, those cases which had been treated non-uniformly in the transcription process were unified, e.g. *mini festival* and *mini-fantastics* in examples (102) and (103) were hyphenated and annotated as <pvc> in the transcript, in line with *mini-dispenser* in example (104), and hence received the POS-tag PVC. Other examples of *mini* were then tagged according to their respective environment, as no suitable tag in OALD7 was available. Thus, *mini* in (106) was tagged with an adjective tag (JJ) and *minis* in (107) with a tag for proper noun plural (NPS).

(106) S3: […] (.) AND we have many supporters. which are (3) (**mini**) . (2) yeah? (PBmtg269:838, L1=pol)

(107) S5: a:nd it's (.) the fifty-eight it's the **minis** i guess (PBmtg414:1384, L1=ger-AT)

4.3.4.3 Structure of transcribed spoken language

A third issue concerned the format which had been used for the transcriptions in VOICE. In order to display the interactive nature of the ELF used in VOICE, clear guidelines were developed, which are specified in the VOICE transcription conventions (VOICE Project 2007b). Transcripts represent the spoken ELF data orthographically (Breiteneder et al. 2006: 176), with a high level of detail with regard to the spoken nature of the data, and are organised in turns. Each speaker's turn is formatted as a paragraph to ensure good readability (Breiteneder et al. 2006: 174). When a speaker's turn overlaps with that of or more speakers, the first speaker's turn is continued until the occurrence of a pause. The overlaps of the other speaker(s) are then listed below the first speaker's turn. The first speaker's turn continues in the subsequent line (VOICE Project 2008: 7). Such a situation is illustrated in example (108) below.

(108) S1: <5> @ </5> (.) would you keep it no matter what <6> a:nd (.) you you would like </6> to preserve it like something you
S4: <6> yes i WOULD keep it </6>
S4: i wouldn't <7> want </7> to preserve it i've i'd even want to become BETTER in it cos m- i DO LACK (.)
S1: <7> own?</7>
S4: erm <1> m- </1> my maltese lacks a L:OT you know?
S1: <1> mhm </1>
S1: y:es.
(EDint330:1061-1067, S1: L1= scc-RS; S4: L1= mlt-MT, eng-MT)

S1's words *a:nd (.) the you you would like* overlap with S4's *yes i WOULD keep it*, overlaps being marked by the pointed brackets *<6> </6>* (The numbers inside the pointed brackets serve the identification of the corresponding simultaneous speech, also marked *<6> </6>*). After *you*, S4 pauses very briefly (not enough to indicate a *(.)* pause), which is when S1's interruption is transcribed. In the following line, S1's continues until the speaker pauses again (this time indicated by a short pause *(.)*), which is when the next overlap, now by S4, is transcribed, indicated by the pointed brackets *<7> </7>*. In this overlap, it seems that S1 actually continues the utterance started in line 1061, in the overlap indicated by *<7> </7>*, which occurs as S4 already continues with her turn: S4 starts with *you would like to preserve it like something* in utterance 1061 and completes the clause with *own* in utterance 1064. However, because of the notation of the interactive mark up, where such cases are not uncommon, the beginning and the end of S4's clause are 2 utterances apart. In other cases, a

speaker completes a turn for a different speaker. For example, in (109) S1 starts with *so we we need to* and two lines further below this is continued by S2 with *know*.

(109) S1: **so we we need to** (.)
S5: (the const-) =
S2: = <8> **know** </8>
S1: <8> somebody </8> needs to be erm (.) <1> defining </1> the academic content and the structure (2)
S2: <1> involved </1>
(POwgd12:796-800, S1: L1=eng-IE, S2: L1=ita-IT, S5: L1=gre-CY)

Cases like these, which include syntactic co-dependencies over longer stretches of text, are difficult to analyse for an automatic tagger. Seeing as taggers are usually trained on written data, the tagging of interactive speech has to be solved in other ways. The tagger can serve to give a pre-analysis of utterances for which low accuracies are calculated and hence point towards those cases which need correction. However, the workload of annotating the majority of these cases had to be carried out manually.

4.3.4.4 Reliance on the original transcripts

The transcription of VOICE was marked by a high level of detail, precision and consistency. This was ensured through the VOICE transcription conventions, on which all transcripts were based and which transcribers were trained to apply. The transcription tool VoiceScribe, which had been especially developed for transcribing VOICE, ensured that transcribers would use the correct formatting, i.e. by colour-coding different mark-up such as overlaps, as well as by pointing out formatting which did not conform to the VOICE transcription conventions (Breiteneder et al. 2009: 23). Also, all transcripts passed through two stages, in which both the audio recording and the orthography was checked by at least one different researcher than the person who had produced the initial transcript of a speech event. In this process, a high level of consistency and correctness was achieved. Hence, it seemed reasonable to view the transcription and checking process as completed for subsequent annotation. This meant that in the tagging process, we would rely fully on these transcripts without returning to the audio files if uncertainties occurred. Although this might have seemed unsatisfactory at certain points for the person doing manual part-of-speech annotation, especially in uncertain cases, this was a necessary measure to take. Apart from the fact that going back to the audio files would not be feasible time-wise, one needs to bear in mind that at least 2 different researchers had revised the transcript. Passages which had been difficult to transcribe for different reasons (e.g. loud background noise, highly interactive passages) had been marked by uncertain or unintelligible annotation. Therefore, passages which were uncertain then were also very likely

to be uncertain or unintelligible if one of the researchers doing the manual tagging had listened to the audio file again, as two transcribers or taggers are highly unlikely to agree on the same interpretation of a stretch of language which is difficult to identify.

The consequences were two-fold: First, it resulted in passages which seemed implausible or difficult to reconstruct and resulted in a high number of ambiguous tags (cf. section 4.3.2 for an account of how such cases were treated in tagging VOICE). Secondly, a number of typing errors and inconsistencies were detected as the manual tagging was being carried out on the original transcripts. These concerned for example homophones, such as *peace* vs. *piece*, *to* vs. *too* or *two*, *its* vs. *it's*, *whose* vs. *whos's* etc. Part-of-speech tagging is a particularly fruitful way of detecting such errata, as the manual annotator will stumble over such issues in an attempt to assign a word class but fails as this makes no sense due to the error. These issues were corrected in the original transcripts in the tagging process if the alternative spelling seemed much more likely in a given context. Also, a number of inconsistencies were detected, e.g. words which were sometimes spelt with a hyphen and in other cases without, e.g. *hard working* was changed to *hard-working* or items which were sometimes annotated <pvc> and others which were not (see discussion of combining words above). Such cases were collected and unified for the newer versions of VOICE (1.1 and 2.0).

In this section, it was demonstrated that decisions taken in the transcription process influence subsequent annotation stages such as part-of-speech annotation. In some cases, this leads to new challenges in the tagging process, such as how to tag data which has been formatted to represent spoken language, and for which taggers trained on written data are typically not prepared to deal with, e.g. dependencies over longer stretches of text. In other cases, the tagging procedure can help to reduce relevant errata and inconsistencies in the transcript which might have gone unnoticed in the transcription process as stretches of text are analysed by an automatic tagger or a person carrying out the manual tagging, as was shown for e.g. homophones.

4.4 Conclusion: Establishing criteria for part-of-speech tagging VOICE

In this chapter, the main challenges which were encountered in the part-of-speech tagging of VOICE have been highlighted. In the course of dealing with these issues, a number of guidelines emerged, which resulted in 4 general guidelines for the tagging of VOICE. These are also elaborated on in the VOICE Part-of-Speech Tagging and Lemmatization Manual (cf. also VOICE Project 2014a: 9f.). The first was to give priority to an ELF perspective, meaning that the first question was always how certain tags should be assigned in agreement within an

ELF framework. The practical implementation of these decisions was subordinate to these questions[64]. Secondly, it was decided that external points of reference should be considered as far as they were compatible with our approach to tagging VOICE. Where they were not, they were adapted and extended. A third guiding principle was that the tagging scheme in VOICE should be intuitively accessible to users of the corpus, and hence using points of references such as word class categories, or the Penn Guidelines. At the same time however, the variable character of VOICE should be displayed. In order to achieve this, it was attempted to balance the use of external reference with a more data-internal approach, e.g. for compiling the multi-word items lists, formulaic items or discourse markers, commonly acknowledged sources such as dictionaries and sources from other corpora were taken into account while at the same time considering VOICE-internal frequency lists. Finally, it was attempted to "strike a balance between inevitable interpretation, i.e. leaving options open on the one hand, and avoiding potentially excessive complexity for corpus users, on the other" (VOICE Project 2014a: 10). The display of different options was achieved by for example, the tagging format of forms- and function tags or the display of ambiguities, while a number of measures, such as detailed guidelines for ambiguous tagging (e.g. sequential tagging procedure) or closed lists for multiword items restricted the complexity of the annotation.

As for any tagged corpus, which will inevitably focus on certain issues, the tagging of VOICE has a number of strong points, and, at the same time, some weaker aspects. VOICE tagging is especially strong in the display of the spoken and variable character of the data. In order to achieve this, the Penn Guidelines were adapted in two ways: In some cases, categories were added, on the others, categories were changed to better represent the data at hand. In total, 30 tags were added. These were 10 tags typical for the spoken interactions in VOICE (partly for already existing mark-up in the transcriptions), i.e. **BR** (breathing), **DM** (discourse marker), **FI** (formulaic item), **FW** (foreign word), **LA** (laughter), **ONO** (onomatopoeic noises), **PA** (pause), **PVC** (pronunciation variation and coinage), **RE** (response particle) and **SP** (spelling out), 15 tags sub-categorising verbs, e.g. **N** and **V** (generic noun and verb tag), and individual tags for verbs forms of be, have and all other verbs (e.g. adding **VV** and **VH** to the tag VB which has been used in the Penn Guidelines). Other tags which have been added are **PRE**

[64] The fact that we operated within an ELF framework seems to be worth stressing here, as this differs greatly to other approaches with which L2 data has been approached, e.g. in learner corpora. Although it has been suggested that tagging and tagsets should be as theory-neutral as possible (cf. Güngör 2010: 206), this has de facto often not been carried out for non-native language data. On the contrary, traditional approaches to tagging L2 data have, from an ELF perspective, viewed the learner language derogatively and in strict comparison to a perceived L1 variety. In contrast, the investigation of ELF data 'in its own right' has been at the very heart of ELF research, this is also something which needs to be foregrounded in the tagging procedures and prioritised over 'theory-neutrality' in tagging procedures.

(pronoun), **UNI** (unintelligible) and **UNK** (unknown). Category changes to the Penn Guidelines include those of Wh-words (distinguishing between **PRE**, **WDT**, **WP** and **WRB**) and those for discourse markers (distinguishing **DM** and **UH**)[65]. As the detailed and consistent annotation of discourse markers was given a high priority in the transcription process of VOICE, these can be searched for more specifically as well, e.g. backchannels, hesitations. In some cases, a search query combining the annotation during transcription with the part-of-speech tags can be used to create very specific search queries for various discourse markers. For example, all hesitation markers can be traced by entering the search "UH,er* (yielding all hesitations/fillers which have been consistently transcribed as *er* and *erm*). The spoken phenomena included in the VOICE Tagset are not represented in most other tagsets discussed earlier in this thesis. As, for example, Rehbein and Schalowski (2012: 239) report for the tagset STTS, "hesitations, backchannel signals, question tags, onomatopoeia, and non-words" are not listed, all of which are present in the VOICE Tagset. The tags in VOICE also cover all the extensions proposed by Gibbon, Mertins and Moore (2000: 30) to the EAGLES category of 'interjection', not all, but most with separate tags. As has been discussed in section 3.2.3.2, most other corpora do not contain such detailed tags for markers of spoken discourse. The other aspect, the display of the variability of the data, was achieved by the dual tagging procedure of form- and function tags, as well as the use of ambiguous tags wherever this seemed necessary.

Another strong point of the VOICE tagging, I would argue, is the detailed documentation and transparency of the tagging decisions taken. From the very beginning, it was deemed of high importance that all tagging decisions, e.g. tag categorisation, description of different categories and points of references used, were documented in what was finally published as the VOICE Part-of-Speech Tagging and Lemmatization Manual (VOICE Project 2014a). On the one hand, the detailed documentation from the beginning served to ensure consistency across tagging decisions. On the other hand, the tagging manual gives detailed information to users of the corpus regarding the approach adopted in tagging VOICE and the tag categorisation, as this is by no means always accessible by intuition. As Greenbaum and Yibin (1994: 34) write, "[i]t is essential that those using tagged corpora should read the accompanying manuals carefully and not rely simply on their understanding of what the names of categories may cover".

However, as probably most tagging enterprises, VOICE had to be carried out within a certain time frame and limited financial resources and therefore, we had to focus on some aspects

[65] A description of the distinguishing criteria for all categories is given in VOICE Project (2014a).

more than on others. Some categorisation is more difficult to carry out than others, and hence exceeded the resources within the project. For example, the tagging of VOICE is weaker in the point of distinguishing between particles and adverbs/prepositions, e.g. differentiation between phrasal verbs and non-phrasal verbs. Other corpus compilers have also reported that distinctions as these are difficult to make. For example, for Penn, guidelines are given for the distinction between the categories preposition, particle and adverb as these are found to difficult to carry out in tagging (Santorini 1991: 9f, 21). Similarly, for CHILDES (Sagae et al. 2010: 709) the categories adverbs and particles into one tag (ADV) for the same reason only distinguished between adverbs and prepositions, which was also found to be problematic.

Another category which could be described in more detail is that of foreign words. In the current version of VOICE POS, all of these are subsumed under one tag (FW). The tagging of individual foreign words was tested within the initial phase of POS tagging VOICE, as the plurilingual language use is something which is characteristic for ELF and, therefore, it would have been desirable to include a more detailed annotation for utterances in other languages. However, it was found that, considering the amount of utterances in other languages, many of which are also transcribed as such (as opposed to being marked ‘unintelligible’), would have taken much effort in relocating people who could speak these languages and assign part-of-speech categories to them. Additionally, it would have meant that the tagset of VOICE might have needed to be expanded considerably for categories that exist in other languages but not in English. All of this it was unfortunately not feasible to carry out within the current project.

However, in the little time that VOICE has been available, it has already been positively received by e.g. Rehbein & Schalowsi (2013: 201f.) and been worked with (cf. Ray Carey ELFA blog 2014). The following chapter will give one example of the use of VOICE POS based on a case study investigating word class shifts in ELF data.

5 APPLICATIONAL STUDY: FORMS AND FUNCTIONS IN ELF

5.1 Introduction[66]

The focus in the previous chapter was on the description of theoretical and practical issues concerning the part-of-speech tagging of spoken language in use, and of spoken ELF in particular. As a result, for the tagging of VOICE, 4 main criteria were developed which were considered useful in the dealing with the challenges and issues emerging when tagging ELF data. This chapter will demonstrate the application of these criteria on an issue which was frequently encountered in the tagging of VOICE, namely the flexible use of lexical items across word class categories in the data.

Previous research has shown that ELF users transcend and exploit boundaries of conventional language use for their own purposes in creative ways (Pitzl 2012) and in ways which are at times unconventional but effective (e.g. Breiteneder 2009; Hülmbauer 2010; Mauranen 2012; Osimk 2010). Seidlhofer (2011b: 99) writes that, "ELF users can be observed — usually quite self-unconsciously — pushing the frontiers of Standard English when the occasion, or the need, arises." As mentioned earlier in this thesis, such extensions of the boundaries of language codifications fulfil a variety of communicative functions (cf. section 3.1.5). In VOICE, numerous cases were encountered in which ELF users exploit language by transcending conventionally established boundaries of word class categories in codified English by change of word class but without any change of morphological form. While these cases were apparent in the transcription process, it was not until the part-of-speech tagging of VOICE, that these cases were captured and investigated systematically. In fact, during the phase of POS–tagging, it became essential to find ways of dealing with such cases, as word class categories needed to be assigned to individual tokens.

The focus of the present section is on the usefulness of the part-of-speech tagging scheme developed for VOICE for gaining insights into the issue of word class shifts in ELF. It will be demonstrated how the criteria discussed in the previous section were applied to the specific case of word class variation, e.g. by application of a dual tagging procedure which gives individual tags to form and function of each token. Moreover, this is an example for one aspect of the data which, without such an annotation, would be tedious, if not to say impossible, to investigate systematically. It hence illustrates how the tagging scheme developed for VOICE can aid in investigating issues which are more tedious to analyse within other tagging schemes.

[66] Parts of this chapter have been published as Osimk-Teasdale (2014).

5.2 Word class change re-assessed: Conversion, multifunctionality and their applicability to ELF data

The phenomenon of a change of word class without a change of morphological form has been discussed predominantly using the two terms *conversion* and *multifunctionality*. As these two approaches have been considered useful when describing inner and outer circle varieties of English, as well as learner English, the section below will discuss these theories and their underlying assumptions. However, we will see that dealing with ELF data calls at least for a partial reconsideration of these conventional approaches in a number of ways, which will be discussed subsequently.

Conversion is usually addressed within the field of morphology and is regarded to be a word formation process which "creates new items, although while it simply changes the word-class of an already existent element, its form remains unaltered" (Balteiro 2007: 13). Plag (2003: 107) describes conversion as "the derivation of a new word without any overt marking", and Pavesi (1998: 213) defines it as "the morphological process by which a word is formed without any explicit derivational mark – e.g. *milk* → *to milk*." There is general consensus that conversion is "an extremely productive process" with few constraints (Plag 1999: 219). However, it has also been observed that the morphological productivity of conversion is language-specific, i.e. that it occurs more in certain languages than in others. This holds true not only for L1- but also for non-L1 language use, where it has been found that if the process of conversion is generally common in a particular language it is also highly productive for all levels of L2 use of this language. For example, while it has been widely assumed that conversion is normally more productive in initial stages of the L2 acquisition process (Plag 2009: 345), Pavesi (1998: 223) found that for English, where conversion is a generally productive process, it is also used frequently by more proficient speakers, alongside other word formation processes.

As criteria for conversion, Balteiro (2007: 76) lists the "extension of the functional potential […] beyond the limits of its word-class", no morphological change of form, and semantic difference between the original and the derived category. For Plag (2003: 107) it is crucial that the "pairs of words […] are derivationally related and are completely identical in their phonetic realization". Furthermore, conversion tends to be regarded as something which starts at speech level but becomes part of the language (Balteiro 2007: 67), and hence "nonce-formations" (Balteiro 2007: 67) "hapaxes" (Plag 2003: 54) or "some weird ad-hoc invention by an innovative speaker" (ibid.) are usually excluded from the analysis.

The literature varies with regard to the types of change typically included in the definition of conversion, from listing only changes across word class categories (Plag 2003, Pavesi 1998) to also including changes between word classes (cf. Balteiro 2007: 77)[67]. For changes across word class categories in L2 English, Plag (2009: 346) lists "verbs […] into nouns of action and result" as well as changes into adjectives, and occasional changes from nouns into verbs. In his study of different L2 Englishes, Biermeier (2008: 87) also mentions various changes across word classes, with changes between nouns and verbs, adjectives and verbs, nouns and adjectives (always occurring in both directions) being the most frequent. Furthermore, it has been observed that in learner language and creoles, changes are favoured when the original and the derived word are similar in meaning (Plag 2009: 346), which stands in contrast to the criteria usually defined for conversion in L1 varieties.

A second common concept to explain word class change is multifunctionality. While both conversion and multifunctionality are based on the assumption of a word class change without altering of morphological form, the concepts differ in a number of ways (illustrated in Figure 5). For multifunctionality, no word formation process as such is assumed (Pavesi 1998: 218), but rather that one lexical item can be freely used in different word class categories (Balteiro 2007: 56), involving no specific direction from A to B (Plag 2009: 346). The forms are assumed to be semantically close, rather than distinct (ibid.). The concept of multifunctionality has often been used to explain conversion-like phenomena in creoles and pidgins (Pavesi 1998: 218) but has also been suggested to account for word class change in learner language (Plag 2009: 346). However, as Pavesi (1998: 218) suggests, multifunctionality and conversion are not necessarily distinct but might also be regarded as two aspects of the same process, stating that, "[i]t is still unclear whether multifunctionality and conversion are two unrelated phenomena similar only at the surface level, or [whether] there is a relationship between the initial free use of lexical items and the later word formation process".

[67] Balteiro (2007: 77f.) also draws a further distinction between *partial* and *total conversion*. In total conversion, all characteristics of the new word-class category are adopted, e.g. *hammer* (noun→verb), whereas partial conversion refers to the adaptation of only some characteristics, meaning that the word belongs to both word-classes, e.g. the *boy king*, where *boy* is both adjective and noun but the comparative and superlative of the adjective cannot be formed

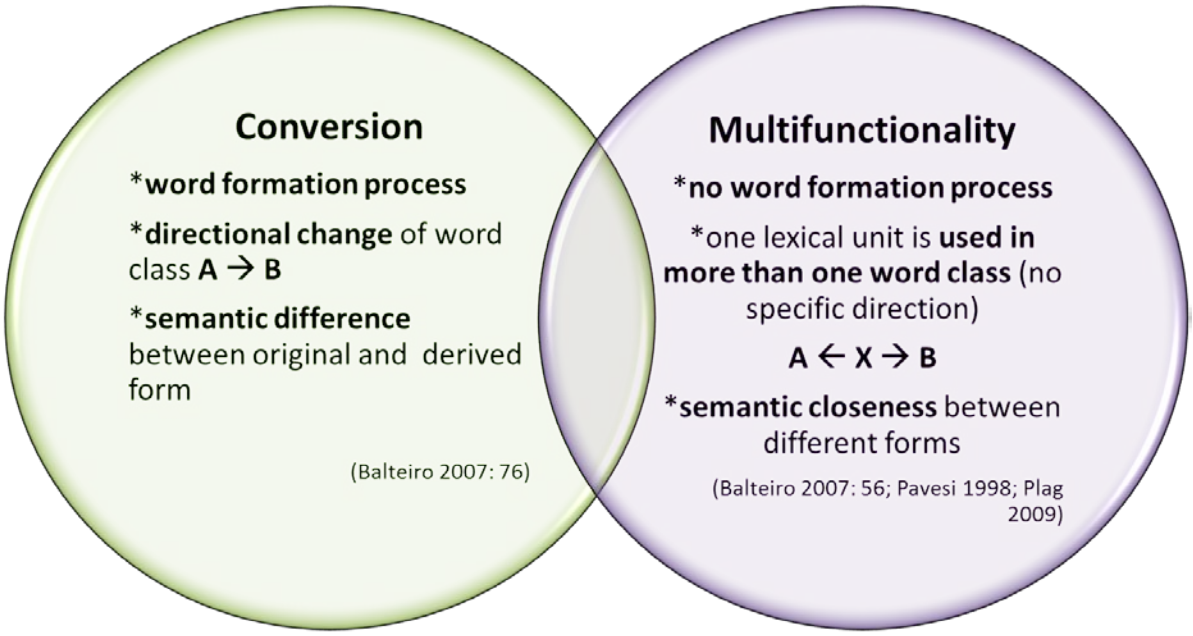

Figure 5: Comparison of the concepts of conversion and multifunctionality

The application of the concepts of conversion and multifunctionality, which have been explored for L1 and L2 varieties, as well as learner language, to describing word class variation in ELF data is problematic in a number of ways. These are rooted in the nature of describing online ELF data, which is different to the description of L1 and L2 varieties of English.

First of all, in contrast to the data referred to in most of the studies of conversion, the ELF data at hand involves mainly speakers who are not part of a stable speech community. Concepts such as "community", along with "variety" or "competence" are, as Seidlhofer (2011b: 95) argues, "essentially arbitrary constructs designed for convenience" which "represent a dynamic process as a fixed state of affairs". In other words, to view ELF in use in terms of these concepts would mean to give an inaccurate picture of fluidity of the interactions observed. And while this is true for any language use, the fluidity of language becomes especially apparent in the dynamics of online ELF interactions. In the literature on conversion and multifunctionality, which is heavily based on concepts such as "community", it is assumed that word class shifts are only valid as such when they recur and become stabilised in the language. There, word class *change* is used to refer to the result or product of the word class shift. However, the unit of analysis for online ELF interactions in this paper is the speech event, which has a clearly defined start and endpoint, and where speakers might not be communicating beyond this particular speech event. Unlike in the traditional concepts

for L1 and L2 varieties, it is therefore irrelevant — if not impossible — to determine if a particular word class change is adopted by other speakers and extended to other contexts outside of a particular communicative situation. As their establishment beyond the immediate environment of the speech event cannot be determined, the word class shifts which occur in VOICE will per definition be what is disregarded as ad-hoc or nonce-formations in the literature on conversion and multifunctionality. Within VOICE they form an important part of developing an interaction and will provide the basis for the analysis in this paper. The word class shifts in online ELF interactions are more suited to be viewed as a *process*, rather than a *product* of a change, and may better be described as word class *variation* or *shifts* than as *changes*, as these terms better describe the fluid and process-like nature of the use of word class categories observed in VOICE data. Hence, in this chapter, I will refer to word class *variation* or *shifts* when making reference to such processes in ELF interaction, and to *change* only when relating to the products referred to in the literature on conversion and multifunctionality.

A second problem with applying the concepts of conversion and multifunctionality to ELF data is that in the latter, where speakers of different first language backgrounds interact, often in the absence of native speakers and far from any native speaker community, the reference to a codified standard might not always be useful, as it cannot be assumed that any native speaker norms are in force. Therefore, issues such as identical prosodic realisation of the start- and endpoint of a word class shift, which Plag (2003: 110) regards as a defining criterion, as well as the issue of directionality, might not be relevant. For example, while Plag (2003: 110) does not regard *let down* and *a let down* as a conversion because of their difference in stress, such a prosodic relationship would not be as clear for an ELF context, because phonological realisations are by no means always performed according to codified norms and it cannot be presumed that prosodic norms are necessarily relevant in ELF communication. Similarly, for ELF it can often be difficult to determine which the 'basic' word class category is, and which is derived. While this is already problematic to decide for more stable speech communities which refer to a common standard (cf. problems listed by Plag 2003: 46, 108-110), this is even more difficult in ELF communication because there might not be one reference but various, as speakers have different first languages and language repertoires and will have learnt English in diverse contexts. Nevertheless, for practical reasons, word class shifts can of course only be described if start- and endpoints are defined. While the prosodic criterion was not considered in the analysis for the data at hand, OALD7 was used as the external point of reference was defined in order to determine which word classes for each token served as starting points. This will be further explained in the following section.

Thirdly, it has been argued that multifunctionality phenomena can be found mostly in "initial learners" (Plag 2009: 346) as for these speakers, the "word class knowledge is not well entrenched" and the change of word class is "a convenient way to fill lexical gaps" (ibid.). Pavesi (1998: 217) writes about one word class change of *food* from noun into a verb: "Given the simplicity of the grammatical system in which *food* is embedded, it is questionable whether the word used in a new functional context is actually the result of a word formation process." The approach of multifunctionality then assumes that 'underspecified' items are the result of unintentional use rather than a conscious process of word formation. In VOICE data, speakers with different proficiencies converse with each other and communicate through English. A criterion for including speakers in the VOICE corpus was "self-selected participation" (VOICE Project 2013d), meaning that the speakers chose to converse in English themselves because they felt competent enough to do so. Hence, for the analysis of word class shifts in VOICE, the problem is that firstly, the level of the speakers' proficiency cannot be determined and secondly, as the word class shifts occur within the setting of spontaneous, interactional language use, many other situational factors may play a role as to why word classes are varied, such as accommodation to other users. It can therefore be difficult to determine whether the word class variation is the result of the speakers' language proficiency or due to other factors. Rather, as mentioned above, it is better viewed in terms of a process, which is embedded in the situational setting, than as a product.

We have seen in this section that conversion and multifunctionality are two slightly different approaches to the phenomenon of the coining of new words with the same morphological form but a different syntagmatic function. Neither of these approaches is unproblematic when applied to ELF data. Hence, for the case study in this paper, these concepts are not regarded to be a valid starting point. Instead, a rather broad definition of the term word class variation will be employed, taking into account all cases in which a differing tag for paradigmatic form and syntagmatic function was assigned because of a shift of category across word classes (henceforth *inter-categorical word class shifts*, cf. below). This means that only those cases will be considered which constitute a change across different word classes, rather than within a word class. Hence, for example, all cases of singular noun forms functioning as plural noun forms, e.g. *two solution_NN(NNS)* are excluded, while nouns functioning as a verb, as in *a kind of summarize _V(NN)* are included in the analysis.

5.3 Capturing word class shifts in VOICE: Tagging forms and functions

The dual tagging of the non-codified form-function relationships, which is described in detail in section 4.3.1.2 forms the basis for the investigation of word class variation in VOICE. This

dual tagging procedure illustrates an application of the criteria developed for tagging VOICE: An ELF perspective was given priority by the development of the dual tagging in order to display the character of ELF data, in particular its variable nature. Hence it was not decided on a single tag for those items where form and function diverged in some way, but on the display of two tags (or sometimes more, if the dual tagging was used in combination with ambiguous tags), which were felt to describe more adequately the variable use of language categories in ELF. This was done despite the increased effort it took to assign tags in tedious largely manually carried out tag assignment and despite the efforts it took to implement and display this format technically. Also, it meant that while external points of reference were considered, i.e. the reference dictionary OALD7 which was consulted to decide on word classes for each of those items, other points of reference were extended to suit the nature of our data. As such, the Penn Guidelines, which served as a starting point, were extended with the dual tagging procedure for VOICE. In this way, it was ensured that external points of reference were used as far as possible, e.g. to make the tagging more intuitively accessible to corpus users, they were also extended where this was necessary to do justice to the data at hand. Finally, the choice to display more than one option, but, at the same time, to restrict these options with regard to the options listed in OALD7 as well as in terms of limited number of possible tags (see section 4.3.2) demonstrates the guideline of keeping a balance between keeping possibilities open but limiting complexity in the tag assignment.

Having illustrated the application of the VOICE tagging guidelines to the issue of word class variation, it is necessary to consider two implications of the dual tagging procedure: First, it needs to be noted that not all cases covered by the differing form-function tags refer to word class shifts, as this format has also been used for other purposes, e.g. to signal multiword items, e.g. multiword discourse markers such as *you_PP(DM) know_VVP(DM)*, vagueness of the category of tokens ending in *–ing*, e.g. in the sequence *vomiting _VVG(NN) pill*, or the use of a pronunciation variation and coinage (PVC), such as in *p_approvement _PVC(NN)*. Such cases obviously do not refer to real word class variation as discussed in this paper and had to be excluded from the analysis. When the differing form-function tags do relate to word class shifts, they can refer to different types. They can concern *intra-categorical cases* (e.g. within nouns, such as count- vs. non-count nouns), or *inter-categorical* cases (e.g. noun to verb, adjective to adverb etc.). Secondly, as has been mentioned in section 4.3.1.2, certain form-function relationships are not represented as individual tags in the VOICE Tagset and hence, cannot be displayed and analysed. Thirdly, as often more than one tag-option was possible for either form- or function-tag, frequent use was made of ambiguous tags. Such cases were, however, not included in the analysis for word class shifts in VOICE (see below).

5.4 Case study: Exploring word class shifts in VOICE

The aim of this case study is to systematically explore word class variation in ELF by considering which types of word class shifts without morphological marking were found in VOICE, and to compare their occurrence in inner- and outer circle varieties of English. Furthermore, a characterisation of word class variation in ELF data will be provided, considering the most frequent examples of word class shifts in VOICE.

The case study was first carried out on a preliminary version of VOICE POS, the part-of-speech tagged version of VOICE, and was presented at the 5th International Conference of English as a Lingua Franca (Osimk-Teasdale 2012). The results presented in this chapter are based on that paper. However, while the results are essentially identical, the numbers indicated in this case study differ from Osimk-Teasdale (2012), as the present results refer not to the preliminary version, but to the final version of VOICE POS 2.0. (VOICE Project 2013b).

5.5 Detection of word class shifts in VOICE

As elaborated above, non-codified form-function relationships, which form the basis for detection of word class shifts, were not annotated in the transcription stage (Osimk-Teasdale 2013, Pitzl, Breiteneder & Klimpfinger 2008: 26) and thus needed to be dealt with during the tagging process. At the same time, detecting these cases is a laborious process as they are not marked by e.g. a non-codified form and could not be retrieved fully automatically. Instead, in a first step, these cases had to be identified manually. This initially took place in the manual tagging of 10,000 words of training material for VOICE POS (see section 4.2) and over a course of two years (2010–2012) while manually annotating and checking large parts of the corpus material. In order to gain consistency, cases similar to those detected manually needed to be captured systematically for the entire corpus. This was conducted using automatic tagging procedures, which assisted further in finding similar cases than those detected in the manual tagging. Even though the tagging was a long process in recurring phases of manual and automatic annotation and corrections, it cannot be guaranteed that each and every case of a token eligible for a differing form- and function-tag was found, because in both the automatic and the manual annotation, different interpretations for sequences are possible. Even though this will be true for any tagged corpus, regardless of the tagging procedure, this is something to be kept in mind.

5.6 Analysis of the data

For the analysis of word class shifts in VOICE, first all combinations of form-tags and function-tags were extracted from the text files, derived from the current version of VOICE POS XML 2.0 (VOICE Project 2013b). Of these, those which had an identical tag for form and function, e.g. JJ(JJ), were excluded, resulting in a list with only those tags which differed in form and function, e.g. JJ(RB), i.e. those which had the formal characteristics of one word class category but functioned as another in the given syntactic environment. This list contained a total number of 152 tag combinations. For the analysis, only those types with differing tags for paradigmatic form and syntagmatic function were considered which had been assigned the tags unambiguously, such as in *bright _JJ(VV)*. Excluded were those differing tags occurring as part of an ambiguous tag, e.g. *sure_JJ(VV)/RB(VV)*, or those where either form- or function tag was tagged unknown (UNK), meaning that this was considered ambiguous across two or more word class categories, e.g. *come backs _UNK(RB)*. However, the analysis did not exclude minor types of ambiguities as displayed by the generic tags V and N. The reason for this was that even though it was sometimes impossible to say whether a form concerned, e.g. a verb base form or a verb non-third person singular form, it was still clear that a verb form was used in a noun-like function, as in *understand* in the example of e.g. *had no understand_V(NN) back then*. Therefore, despite the ambiguous verb form, these cases could still be identified as cases of word class shifts and were included in the analysis. In a further step, of those cases with differing tags for form and function, tag types were excluded which served a different purpose than signalling a shift of word class (e.g. multiword discourse markers, tokens ending in *–ing*, pronunciation variation and coinages, see above).

The remaining cases were then sorted into those cases of inter-categorical and those of intra-categorical word class shifts. In total, 32 types and 370 tokens of cases were found which concerned changes across word class categories. For those being assigned a differing tag for form and function within a word class (e.g. form: noun plural, function: noun singular, as in NNS(NN)), a much larger number were identified, namely 52 different types and 1440 tokens. These, however, were not considered for the present study.

5.7 Findings

Table 3 below displays the types of word class shifts found in VOICE POS 2.0. The first column, from left to right, indicates the type of word class shifts encountered, the second column the types of tags represents the respective word class shift, the third column the

number of the individual tag occurrences in VOICE, and the forth column the total number of occurrences of the respective type of word class shift.

Table 3: Types of word class shifts in VOICE

[Explanation] Adj=Adjective, Adv=Adverb, Card Num=Cardinal number, Det=Determiner, Formulaic=Formulaic Item, N=Noun, Ord Num=Ordinal Number, Pers Pron=Personal Pronoun, Prep=Preposition, Sub Con=Subordinating Conjunction, V=Verb. For correspondences between tag and word classes see Appendices A and F. All shifts which are bi-directional are marked in colour.

Type of change	Tag types included	Number of occurrences for each tag type	Total number of occurrences for word class shifts
Adj→Adv	JJ(RB)	98	
	JJS(RBS)	1	99
Adv→Adj	RB(JJ)	52	52
Adj→N	JJ(N)	1	
	JJ(NN)	32	
	JJ(NP)	8	41
N→Adj	NN(JJ)	27	
	NP(JJ)	8	35
V→N	V(N)	1	
	V(NN)	28	
	V(NP)	1	
	VVZ(N)	2	32
V→Adj	V(JJ)	26	
	VVZ(JJ)	1	27
N→V	NN(VVP)	4	
	NN(VV)	18	22
Car Num→ Ord Num	CD(NN)	17	
	CD(NNS)	1	18
Adj→V	JJ(VV)	6	
	JJ(VVN)	2	
	JJ(VVP)	4	12
Adv→Det	RB(DT)	8	8
Adv→N	RB(NN)	7	7
Card Num→Adj	CD(JJ)	4	4
Prep/Sub Con→N	IN(NN)	4	4
Det→Adv	DT(RB)	2	2
Det→N	DT(NN)	2	2
N→Adv	NN(RB)	1	1
Adj→Ord Num	JJ(CD)	1	1
Card Num→Det	CD(DT)	1	1
Pers Pron→Det	PP(DT)	1	1
Formulaic→Adj	FI(JJ)	1	1

In total, 20 different types of inter-categorical word class shifts could be identified via the analysis of tokens which had been assigned a different tag for form- and function position. In this paper, only those types of word class shifts occurring 12 or more times (from Adj→V) in the table, are considered. Analysis of the data showed that all the types of word class shifts occurring less than 12 times either concerned no of word class shift as this term is understood in this paper, or they are very infrequent, with only 4 occurrences or less.

5.7.1 Adj→Adv

The most frequent type of word class shifts is that from adjective to adverb, as exemplified in (110) to (114) below, which lists the most frequently occurring word class shifts from adjective to adverb. Most frequently, the tokens particular (8 times), total (5 times), absolute, complete and usual (each 4 times) occurred in this type of change.

(110) S2: this this also yeah yeah **particular** _JJ(RB) AUSTRALIA (PBmtg300:1175, L1=dut-NL)[68]

(111) S1: yeah _1 this is a **total**_JJ(RB) special way _0 and if we consider that a lot of cargo will not move via luxembourg any more _1 yeah erm _0 s- strength like a_[S4] has for example developed in in time _1 p_virtuality _1 yeah er and those _0 are for more value (PBmtg27:1213, L1=ger-DE)

(112) S1: yeah this was **absolute**_JJ(RB) great _0 (POmtg447:56, L1=rum-RO)

(113) S11: but er _1 we are **complete**_JJ(RB) different (EDsed31:1134, L1=ita-IT)

(114) S7: but **usual**_JJ(RB) they get the recognition because they have someone from foreign country they have there 's something happening _0 the organization like _0 the work of the _0 (POwgd524:929, L1=ger-AT)

Most of the occurrences of word class shifts from adjective to adverb are individual instances, occurring twice at the most in one speech event, and are usually only used by one speaker.

5.7.2 Adv→Adj

A reverse word class shift from adverb to adjective is the second most frequent, with 52 occurrences in total and *really* (12 times) and *funnily* (3 times) occurring most frequently in the data, as exemplified below. Again, most instances are ‘one-offs’ and are scattered across a large number of speech events.

[68] Examples in this chapter are displayed with a reduced mark-up that was used for the POS-tagged version of the corpus. Tags are given after the underscore for the token where the word class shift occurs, e.g. particular _JJ(RB). The symbols *_0*, *_1*, etc. refer to pauses of different lengths.

(115) S3: because in in in p_hungaria _0 people do n't speak a **really**_RB(JJ) german _0 you are from er _0 erm _0 south tyrol (PRcon29:141, L1=ger-AT)

(116) S2: before that 's **funnily**_RB(JJ) enough huh _0 (POmtg541:146, L1=dan-DK)

5.7.3 Adj→Noun

The most frequent occurrences of word class shifts from adjective to noun concern the words *win-win* (5 times), *eastbound* (4 times) and *westbound* (3 times), as demonstrated below in examples (117) and (118).

(117) S4: this is you know _0 er it 's a **win-win**_JJ(NN) for for her _0 this is the reason she 's interested i- interested in this er project (POmtg439:296, L1=fre-LU).

(118) S2: a_[S1] the danger is quite nearer _0 look _1 the **westbound**_JJ(NN) is now okay _0 but the **eastbound**_JJ(NN) er as we all know rates are whhhh @@ _0 and _0 you see this imbalance _0 is getting closer to the the break even more and more because the _0 (PBmtg300:1423, L1=dut-NL)

Eastbound and *westbound* are event-specific terms in PBmtg300, referring to flight routes in the conversation and are, apart from one instance, used as nouns there.

The token *win-win* occurs in three different speech events (PBmtg462, POmtg439 and POwgd524). While in PBmtg462 and POmtg439 there are no other occurrences of the *win-win* apart from the two instances of the converted adjective into a noun in these speech events, in POwgd524, the word occurs 14 times overall in the conversation and is used by three different speakers, namely S2, S1 and S4. S2 uses it first (3 times), in the phrase *win-win situation* in utterances 31, 535 and 537. This is also the collocation listed in OALD7, the reference dictionary. Then, in utterance 734, S1 first uses *win-win* without clear marking as an adjective or noun, as exemplified in (119).

(119) S1: not **win-win**_JJ(JJ)/JJ(NN) yeah but that 's the question _2 (POwgd524:734, L1= ger, ita)

Over the course of the conversation, the word is used more often in combination with *situation*, both by S1 and S4 (utterances 748, 757, 771, 774, 780, 785, 947 and 954). In the last two occurrences of the speech event, however, S1 uses *win-win* again (examples (120) and (121)), but without *situation* and with the clear noun-markings *this* and *the*, signalling the word class shift more clearly than before.

(120) S1: so i think we i- it 's very important to underline this **win-win**_JJ(NN) _1 (POwgd524:960, L1= ger, ita)

(121) S1: and and and i think it 's quite _0 and this is where the **win-win**_JJ(NN) _1 depends because it 's very much as you said it depends very much on the expectation and the motivation and if if those do n't come together (POwgd524:980, L1= ger, ita)

Therefore, the phrase *win-win situation*, i.e. the conventional phrase, is present in the conversation and it seems that S1 also knows it or uses it but then decides to convert it into a noun – first without clearly marking it as such (119), then with clear marking in the last two instances (120) and (121). The word class shift is not taken up by the other speakers.

5.7.4 Noun→Adj

Examples (122), (123), (124), (125) and (126) list the most common examples for word class shifts from noun to adjective, i.e. *solidarity* (3 occurrences) and *fanatic*, *so-and-so*, *technician* and *belgium* (each 2 occurrences):

(122) S5: yes _1 so it 's democratic it 's very soli- _0 **solidarity**_NN(JJ) (EDwgd241:869, L1=pol-PL)

(123) S1: hm _1 i do n't know _2 i think i like celtic music but erm not so as er er a_[firstname8] is **fanatic**_NN(JJ) about this kind of music @ and she loves ireland and _0 i never been there _0 you (LEcon229:374, L1=ger-AT)

(124) S5: so for example _0 you know for spanish you could say hh _0 **so-and-so**_NN(JJ) number of speakers in spain **so-and-so**_NN(JJ) number of speakers in the world (POmtg444:700, L1=eng-GB)

(125) S3: like the **technician**_NN(JJ) university or some (PRint30:39, L1=ita)

(126) S9: what about the **belgium**_NP(JJ) frenc- s- french-speaking community _0 (POmtg404:260, L1=nor-NO)

While the instances of *fanatic*, *so-and-so* and *belgium* occur in isolation in the speech events, uttered as ‘one-offs’ by individual speakers, the situation is different for *solidarity* and *technician*.

In the speech event EDwsd302, the token *solidarity* occurs 67 times, predominantly in its conventional function as a noun. This conversation takes place at a student conference where students have to decide on five values which they consider of primary importance. ‘Solidarity’ is one of the values discussed and is used by various different speakers as a noun. In utterance 1734, however, S5 uses it as an adjective, as demonstrated in example (127) below.

(127) S5: but i do n't wan na be **solidarity**_NN(JJ) we- so- it 's er _0 er f_solidarisch with with everybody (EDwsd302:1734, L1=ger-DE)

In this example, it is fairly clear that S5 uses solidarity intentionally as an adjective, underlining this intention by switching into German to demonstrate that *solidarisch* is what she wanted to say and signalling that this word is presently not available in English to her. The word class shift from noun to adjective without morphological marking is this 'lexical gap' which is not a lexical gap in the English language per se, but for S5 at that given moment. At a later point in the conversation, solidarity is used again by a female speaker, who, however, could not be clearly identified as either S5 or a different speaker:

(128) SX-f: i _0 am _0 **solidarity**_NN(JJ) with you (EDwsd302:2038, L1=unidentified)

As with *solidarity*, the occurrence of *technician* is also a re-occurring topic with regard to universities in the speech event PRint30. In this interview between three speakers, we can also observe some variability as to how the adjective *technical* is realised: we find the conventional form *technical*, the varied form *technician*, as well as the German equivalent *technisch*. S3 first uses *technical* in the coinage *polytechnical*. Two utterances on, S3 says *technician university* (129), converting the noun *technician* into an adjective.

(129) S3: like the **technician**_NN(JJ) university or some (PRint30:39, L1=ita)

This is used in the same way by S1 a little further on, in utterance 42 (example (130)).

(130) S1: but which universities are yours then _1 not the **technician**_NN(JJ) ones (PRint30:42, L1=ger-AT)

Sometime later, S1 refers to the type of universities by using the conventional term *technical university* (utterance 95). The final occurrence of the word, this time referring to a museum rather than a university, S3 who initially introduced the word class shift of *technician* into an adjective, switches into the L1 when referring to the museum, saying *technisch museum*, rather than *technician museum* (example (131)).

(131) S3: we have to _0 we we have a list of a _0 the _0 the **f_technisch**_FW(FW) museum _0 (PRint30:290, L1=ita)

It is worth noting that, unlike in the example of *solidarity* above, the language S3 switches to is not S3's first language but that of S1, the interviewer's. As an example (127), the word

class shift (here from earlier in the conversation) is supported and made clearer by the use of an additional language.

5.7.5 Verb→Noun

Of the 32 cases of a word class shift from verb to noun, all but three cases (i.e. *assess* with 5 occurrences and *proofread* and *believe* with 2 occurrences each) concern individual tokens, which occur across all domains and various speech event types. Two of these are exemplified in (132) and (133) below.

(132) S1: which is we i had no **understand**_V(NN) _0 back then i did n't understand bristol it 's very cool hh _0 (LEcon573:118, L1=ger-DE)

(133) S3: so they were keeping our **suggest**_V(NN) to the very very end _0 and the guys were afraid (PBmtg269:441, L1=pol)

Of most frequent occurrences, i.e. *assess* and *believe* and *proofread*, only the last can be classified as word class shift which is morphologically motivated, as with the other two, a phonological reason for the word class shift can be assumed (*internet assess* vs. *internet access, believe* vs. *belief*). What can be noted, however, is that for this group, mostly uninflected, rather than inflected, forms are shifted.

5.7.6 Verb→Adj

In general, what is noteworthy about the word class shift from verb to adjective is that two thirds (18 tokens of 27) of base forms were shifted, e.g. *a understate level, my english is so complicate, phd level er is becoming very restrict*, only in 9 instances an inflected form (ending in –ed or –s) is converted into an adjective (e.g. *this is not equals to the licence, the subcontracted work*).

The most frequent changes concern the tokens *compacted* (5 occurrences) and *interest* (5 occurrences). While all instances of *compacted* occur in speech event PBmtg269, *interest* as adjective (demonstrated in (134)) can be found across 4 different speech events.

(134) S1: er the reason why i 'm **interest**_V(JJ) in is the er _1 the _0 biggest barrier _0 what 's the price point (PBmtg3:3744, L1=kor-KR)

The adjective *compacted* is introduced by S3 in PBmtg269 and is only used by this speaker in the conversation. This speaker gives a power point presentation where ‘compactness’ is a key characteristic of the product S3 is talking about. In addition to using *compacted*, the speaker also uses the canonical adjective *compact* (utterance 62). This is always used in the environment of *compact [thing1]* (utterances 62, 164, 827) by S3 (and in utterances 769 and

785 by S6). *Compacted* as an adjective, however, is used in the context of *more compacted* (utterance 60, utterance 831), *the compacted form* (utterance 827), *thirty-three per cent compacted product* (utterance 833). In addition, the different speakers also talk about this characteristic of the product using the noun coinage *compaction*[69], and the terms, *compacting, supercompact, compacts,* and *supercompacts*.

Two aspects which can be noted here are the following: Firstly, S3 obviously deliberately alternates between *compacted* and *compact* as adjectives for different uses. Secondly, the characteristic of the product seems to be open to debate in the conversation as so many different terms (noun, adjective) are used to refer to it. It becomes especially apparent in utterance 728 where S5 says the following, using *compacting* and *compact*:

(135) S5: i i believe we should avoid this **compacting** word _0 **compact** _2 word or (PBmtg269:728, L1=slv)[70]

5.7.7 Verb→Noun

Of the 22 cases of word class shifts from verb to noun only one occurs more than once, namely *sticker* (example (136)), which occurs twice. All other cases are individual occurrences, some examples given in (136) and (137).

(136) S3: _0 it 's not possible to bring it in because er i cannot **sticker**_NN(VV) this small packaging (PBmtg463:387, L1=ger-AT)

(137) S2: then we can **corridor**_NN(VV) this this kind of possibility _0 (PBmtg300:481, L1=dut-NL)

(138) SX-6: _0 and they didn't **guilt**_NN(VV) that they _1 have aids (EDwsd302:1612, L1=lav-LV)

5.7.8 Cardinal Number→Ordinal Number

This word class shift concerns 18 cases of cardinal numbers being shifted to function as ordinal numbers, as exemplified below. They occur in 4 different speech events (LEcon351, PBmtg27, POmtg541, EDcon521) and concern mainly individual instances of word class shifts.

[69] The use of *compaction* in this context seems to be filling a lexical gap as the codified noun *compact* in OALD7 has a different meaning, e.g. "a small car[,] a small flat box with a mirror, containing powder that women use on their faces [and] a formal agreement between two or more people or countries" Hornby et al. (2005: 305).

[70] In this example S5 does not refer to the linguistic level of avoiding the word, but to the characteristic of the product, e.g. vs. sustainability.

(139) S7: we have on the **twenty-four**_CD(NN) of december it 's er _0 santa claus (LEcon351:191, L1=spa-AR)

(140) S3: the er _1 the er _1 the basic documents will need to be sent before our next meeting _0 before before may t- **twenty-six**_CD(NN) absolutely _1 (POmtg542:269, L1=fin-FI)

5.7.9 Adj→Verb

These 12 cases are also mainly individual occurrences of changes from adjective into verb, as exemplified below.

(141) S3: and and i always **poor**_JJ(VVP) a little bit him because he never _0 speaks any word in maltese _0 (EDint330:902, L1=mlt-MT)

(142) S19: yeah but but still all countries have _0 some countries just **proud**_JJ(VVP) their history and **proud**_JJ(VVP) their culture it's like when it (EDwgd241:257, L1=rus-RU)

5.8 Discussion: Word class shifts with a purpose

There are a number of general observations which can be made with regard to word class variation encountered in VOICE data. These and possible implications of these will be discussed in the following section. For the most frequent types of word-class shifts (12 occurrences or more), tendencies can be observed with regard to the type of *forms* which are shifted, the *directionality* of the shifts, and the *environments* in which shifts occurs.

5.8.1 Forms

Most commonly, morphologically simple forms are shifted rather than morphologically more complex ones. This can be observed for all forms which have morphologically simple and more complex counterparts. For example, for *nouns*, common and proper, no clearly identifiable noun plurals (NNS, NPS) are shifted but always noun singular forms (NN, NP). Only one form is shifted which could not be clearly identified as either a singular or plural noun (N). This is also observed for *verbs*, where word class shifts from a base form or uninflected form are much more common than from an inflected form. Of 59 verb forms which are either converted into a noun or adjective, 45 are converted from a base form (see example (143)) and only 14 from an inflected form (see example (144)).

(143) S3: so they were keeping our **suggest**_V(NN) to the very very end _0 […] (PBmtg269:441, L1=pol)

(144) S1: okay to be very honest a_[firstname45] _0 has all the abilities you would need _0 commercial **knows**_VVZ(N) the system knows everything would be a fine and easy replacement _2 […] (PBmtg27:730, L1=ger-DE)

Thirdly, if we consider *adjectives* and *adverbs* as related word classes, a similar observation can be made with regard to the preferred word class shifts of simpler forms rather than more complex ones. The results show that word class shifts starting from adjective to adverb are twice as frequent as vice versa, i.e. Adj→Adv occurring 99 times and Adv→Adj occurring 52 times (cf. Table 3). When considering the overall occurrences of plain forms of adjectives and adverbs in the corpus (the tag RB for adverb was assigned 76,794 times, the tag JJ for adjective 45,467 times), we can see that 0.2% of adjectives were converted into adverbs, but only 0.07 % of adverbs into adjectives. We can, therefore, see that the shifts from the 'simpler' adjectival form into the more complex adverbial function is more than twice as frequent as the opposite direction. Finally, one of the frequent changes illustrated in Table 3 concerned the change from *cardinal* into *ordinal number*. Again, this is a change from a simpler form without an affix into a more complex form with an affix. The observation that morphologically simple forms seem to be a common feature in ELF interactions is of course nothing new and has been shown particularly for the flexible use of forms within word class categories. For example, Breiteneder (2009: 259) discusses the use of zero-realisation of third person –s and argues, speakers utter instances such as "it last three years" in order to "exploit the internal redundancy of the language" (Breiteneder 2009: 264). This has also been shown for singular noun forms used with plural meaning, such as "next three point" or "in most university" (Osimk & Breiteneder 2011), where the singular form of the noun commonly co-occurs with some marker of plural, such as a number, but also a plural verb or an adverbial plural marker. The information inherent in a preceding phrase or word is not repeated in the noun with zero-s-plural-realisation. On the one hand, it therefore seems that sometimes a reason for the use of less inflected forms is the reduction of redundancy. Another way of explaining the preferred use of less complex forms might be that the ELF users in the data seem to rely on the underlying conceptual category of a word, namely the uninflected form, which is then used flexibly, and that more superficial features, such as inflection, are disregarded. According to Pavesi (1998: 214), "conversion is […] characterized by semantic ambiguity as the same form has different meanings which can be identified only through syntagmatic environment". This makes it less "transparent" (ibid.) than other word formation processes, e.g. affixation. It can be argued that such simple forms, which are preferred by ELF speakers, display more "semantic ambiguity" than inflected forms. While perhaps reducing

clarity, they seem to be useful to ELF speakers in that their ambiguity also increases flexibility. Seidlhofer (2011b: 80) argues that environments where English is spoken as a lingua franca are characterised by such flexibility, which actually "strengthens the communicative robustness" (ibid.) of these speech situations.

5.8.2 Directionality

A second issue emerging from the results concerns the directionality of the word class shifts. Of those changes occurring 12 or more times in VOICE, all apart from one change, namely that from cardinal to ordinal number, occur in both directions, as illustrated in Figure 6.

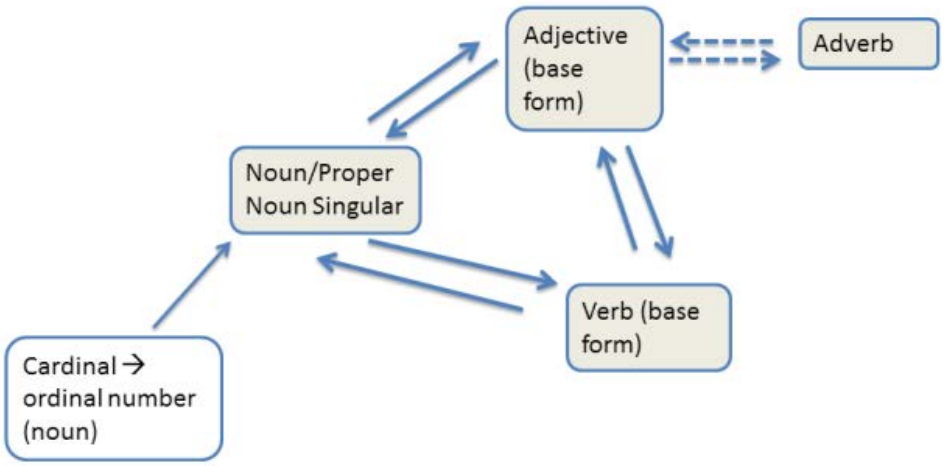

Figure 6: Direction of word class shifts in VOICE

First, what can be shown by this figure is that in VOICE data, (mostly simple, uninflected) forms of nouns, verbs and adjectives are easily moved and that the most frequently occurring types of word class shifts tend to go both directions, i.e. are bi-directional. Moreover, what becomes apparent is that word class shifts occur mostly between categories which express semantic content, rather than syntactic functions. As with the preferred use of morphologically simple forms, what this suggests is that what is prominent in ELF conversation is the conveying of a message, or, in other words, going for communicative essentials, while leaving aside other features, especially those which are contextually or co-textually redundant. The focus is on the expression of a semantic content, while other features, such as morphosyntactic properties are disregarded.

If we then compare the types of word class variation occurring in VOICE to the types of conversion mentioned in Plag (2003: 108) and Balteiro (2007) we see that not all of these types of conversion are mentioned for L1 Englishes. While some of the word class shifts found in VOICE are mentioned also in Plag (2003) (i.e. noun→verb, verb→noun,

adjective→verb, adjective→noun) the following five changes are not mentioned: noun→adjective, verb→adjective, cardinal→ordinal number, adjective→adverb, adverb→adjective. This is demonstrated in Figure 7.

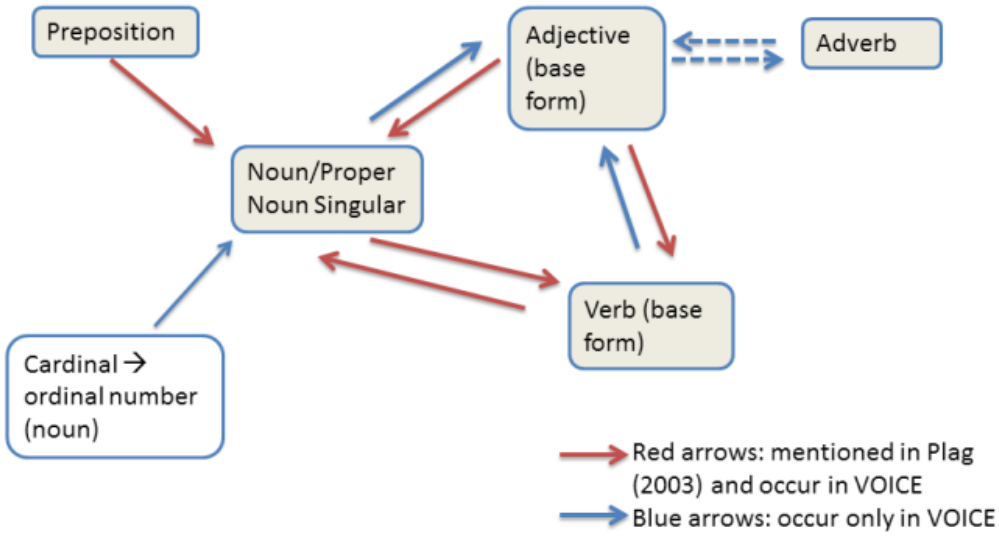

Figure 7: Direction of word class shifts in Plag (2003) and VOICE compared

Balteiro (2007) in her book on English conversion on the other hand, does mention all types found for VOICE (and more), apart from the shift verb→adjective (although some are only regarded as partial conversion by Balteiro). Compared to other L1 and L2 varieties, taken from Kortmann and Schneider's (2006) interactive map of *Varieties of English* and Biermeier's (2008) study on conversion of different L2 varieties of English from the ICE corpora, we see that all but two types of word class shifts (adverb→adjective and cardinal number→ordinal number) found in VOICE also occurred in L1 and L2 varieties of English mentioned in one of those sources, as shown in Table 4.

Table 4: Types of word class shifts in VOICE compared to those mentioned in Kortmann and Schneider (2006) and Biermeier (2008)

[Explanation] All changes which are bi-directional are marked in colour.

Word class shift in VOICE	Number of occurrences in VOICE	Kortmann and Schneider	Biermeier
Adjective→Adverb	99	✓	-
Adverb→Adjective	52	-	-
Adjective→Noun	41	-	✓
Noun→Adjective	35	-	✓
Verb→Noun	32	-	✓
Verb→Adjective	27	-	✓
Noun→Verb	22	-	✓
Car Num→Ord Num	18	-	-
Adjective→Verb	12	-	✓

The word class shift from adverb to adjective is not mentioned in Biermeier (2008), Kortmann and Schneider (2006) or Plag (2003). Only Balteiro (2007: 90ff.) mentions this. However, in the ELF data in VOICE, this word class shift is the second most frequent type encountered. What is noteworthy about this shift is that in VOICE, of the 52 different word class shifts from adverb to adjective, 12 concerned the adverb *really* (examples (145), (146) and (147))). All other tokens, e.g. *exactly*, *typically*, *funnily* etc. only occurred between 1 and 3 times each. Moreover, 7 of the tokens are accompanied by other adverbs expressing quantity, as exemplified in (148), (149), (150) and (151), such *more* and *very*.

(145) S12: […] _0 and trust building _0 is an a **really**_RB(JJ) issue _0 so maybe we could in the next _0 next half an hour (EDwsd499:959, L1=cze-CZ)

(146) S1: […] _0 er second question _0 that is the idea of structures we _0 feel that we have no **really**_RB(JJ) problems _0 about _0 different structures _0 […] (POwgd14:695, L1=swe-SE)

(147) S3: because in in in p_hungaria _0 people don't speak a **really**_RB(JJ) german _0 you are from er _0 erm _0 south tyrol (PRcon29:141, L1=ger-AT)

(148) S5: no no no no no _0 no no no no i think you are great in that i act- _0 really _0 i think you should do that _1 i think you are **very fluently**_RB(JJ) in english _0 you can present yourself very well you 've that shown a lot of times here _0 i think you can do that very well _0 (EDwsd242:363, L1=ger-AT)

(149) S1: but like with russians _0 they have to be _0 like **very slowly**_RB(JJ) very _0 (LEcon560:1184, L1=pol-PL)

(150) S1: i mean but you could have **more liquidly**_RB(JJ) _1 than this (LEcon566:92, L1=ger-DE)

(151) S2: […] _0 at least slovak and czech er eh slovak and czechs _0 have a _0 quite bad experience with er credit with capital market in terms of er both x organizations that er was _0 **more fairly**_RB(JJ) in slovakia than _1 in the czech republic but still _0 (PRpan294:110, L1=slo-SK)

Although it is impossible within the limits of this thesis to give a definite explanation of the frequency of this word class shift, I would suggest a possible way of interpreting the examples above: All of them concern adjectival functions of tokens which express a characteristic fairly strongly. It seems that the strength of the adjective is enhanced by the adverbial form, which, through the suffix *–ly* is longer and hence more prominent than an adjective, as in the 12 cases of the adverbial form *really*, which is used instead of the adjectival form *real*. In the cases of examples (148) to (151), additional adverbs, e.g. *more* and *very* are used, which seems to

strengthen the word class shift even more. This might be a hint, therefore, that in the case of word class shift from adverb to adjective, this may happen in order to enhance clarity, as has been suggested for other types of word formation processes in ELF data (Pitzl, Breiteneder & Klimpfinger 2008). A reason why this happens in ELF, but has not been observed for many other varieties of English, might be that what has been argued for ELF communication is also true here: that conventions are not as fixed there and therefore linguistic repertoires are more exploited in order to converse and convey a message. It has been suggested that second language *use* gives particular rise to linguistic innovations (Siegel 2004, Siegel 2008) and that this may be applied to word formation processes (Plag 2009: 358). The data in VOICE concerns exactly such – predominantly – non-native language use, and we have seen that this exhibits linguistic innovation such as word class variation. It may, therefore, be reasonable to assume that some of the structures which occur for VOICE, where we are dealing with, but are not mentioned for e.g. L1 data (cf. Figure 5) and/or other L2 data, such as the adverbial forms being used in adjectival positions, may occur in and through the particular combination of circumstances in which ELF is actually used.

5.8.3 Environment

All of the word class shifts with 12 occurrences or more were investigated within their particular context of the speech event. The results showed that often occurrences of word class shifts occurred only as single instances in a speech event by a particular speaker and remained uncommented by the other speakers. In some cases, however, the shifts occurred more than once in a conversation and played an integral part to the conversation topic, as demonstrated above by the examples of *win-win* in POwgd524, *solidarity* in EDwsd302, *technician* in PRint30 and *compacted* in PBmtg269. From these examples, a number of aspects can be observed as to how word class shifts function in a particular speech situation in ELF.

First of all, words which are converted or related lexical items are often key terms in their respective conversations. For example, *win-win* occurs 14 times in total in the conversation and *solidarity* 67 times. For *technician* we find 2 different occurrences with co-occurrences of *technical*, and *technisch* in the conversation, and *compacted* occurs 5 times, with co-occurrences of *compact*, *compaction*, *compacting*, *supercompact, compacts,* and *supercompacts.* In these conversations, therefore, the words which are frequently converted are not unknown to the other speakers. It seems therefore, that the frequent occurrence of the words in these conversations gives rise to a number of creative uses of the language resources of the speakers.

In the cases of *solidarity* and *technician*, the shift is also supported by a switch into another language, either the L1 of the speaker for the first, or the L1 of an interlocutor for the latter, signalling that the speakers are filling a lexical gap with the shift of word class. The fact that a different language is also used to clarify the meaning suggests that the word class is shifted consciously with a clear intention in mind. Similarly, in PBmtg269, S3 clearly alternates between *compacted* and *compact* as adjectives for different uses, which again indicates an intentional use of the word class shift.

Irrespective of whether the words occur as 'one-offs' or frequently in a conversation, what can be noticed is that the word class shift is rarely taken up by other speakers. In the examples given, this only happens once for *technician* used as an adjective, which is taken up by S1, and possibly for *solidarity*, though the second time it is used, the speaker could not be identified. What has become clear though is that the word class shifts in these situations are effective, intentional and fulfil the function of getting a message across, even though they are not adapted by the other speakers or, indeed, a whole speech community. This resonates with Štekauer's (2002: 98) understanding of nonce-formations, which he sees as "regular, structurally transparent, predictable naming units generated by (potentially) productive word-formation rules" and which occur because of "the interplay of linguistic and extra-linguistic factors" (2002: 99). For ELF data, this suggests that that so-called nonce-formations should not be ignored but rather investigated within the contexts in which they occur.

5.9 Conclusion: Towards a characterisation of word-class variation in ELF

The issue of different form-function relationships was one of the challenges which had to be dealt with in the tagging of VOICE, because of the very nature of the data. Engaging with these challenges necessarily focuses attention on characteristics of ELF data that might otherwise escape notice, and raises questions about what significance they might have for understanding how linguistic forms function in the ELF interactions recorded in VOICE. The analysis in this chapter prompts several observations.

First, we can observe a number of general trends regarding word class variation in spoken ELF data in VOICE. Generally, it can be said that the ELF users in VOICE make use of word class shifts without change of morphological form. This does not appear to happen randomly, but with some clear tendencies: it is *bidirectional* and it concerns the movement of *un-inflected forms* between *content words*. This suggests that what has been shown for other areas of ELF research is also true for variable word class use in the data at hand: ELF users seem to foreground semantic, rather than morphosyntactic information by using mainly shifts

between categories which express content rather than function, and prefer the flexibility of simple forms over the less ambiguous inflected forms. Moreover, the types of word class shifts found in the ELF data at hand are mostly not 'new', but can also be found in L1 and L2 varieties of English, which points toward the naturalness of the phenomenon in ELF communication.

Returning to the different possible explanations of change of word class without morphological modification, *conversion* and *multifunctionality*, it can be said that both concepts are potentially useful for describing and explaining the cases observed in VOICE data. However, they need to be adapted in a number of ways when applied to ELF data (see Figure 8).

From the point of view of conversion, we can view the word class shifts as examples of a word formation process. Word formation processes require a certain intentionality, which, in the examples given, was signalled by switching into another language as support for the word class shift used, or by alternation between a codified and a non-codified word (*compact* and *compacted*). However, rather than being changes which are established in a certain speech community, as in literature on conversion, we are dealing with — as we may call it — *situational contingency*, meaning that words are produced on the occasion, and rather than being conventionally established, they might only ever be used in a certain communication situation. This is different to a word being established in a more stable language community, as in L1 and L2 varieties of English. As such, many cases discussed in this paper are so-called nonce- or ad hoc-formations. However, while such cases are usually not included in traditional analyses of conversion, the analysis in this paper certainly shows that general patterns of conversion in ELF data can be described based on such word formations which arise out of a situation and are often only used by one speaker. Such trends relate to the forms, directionality and environment of word class shifts in VOICE. The study of ELF interactions enables, even obliges, us to think of the immediate process of conversion as it happens online, whereas traditional analyses focus attention on the results of this process which have become established as language change.

The concept of multifunctionality might also be useful in order to explain some aspects of the word class shifts in the VOICE data. We have seen that the most frequent word class shifts in VOICE are between those categories which express semantic content rather than grammatical function, which resonates with the idea that the multifunctional use of lexical items happens when they are semantically close (Plag 2009: 346). However, in VOICE, word class shifts are used across all domains in different speech event types and by speakers of very different

levels of proficiency. Hence, within ELF interactions, such multifunctional use can be reconsidered as a resourceful and creative exploitation of the potential of the virtual language, related to the concept of virtual categories (cf. Seidlhofer & Widdowson 2009, Seidlhofer 2011b, Widdowson 1997), rather than a phenomenon which occurs in 'simple' grammatical systems as has been suggested (e.g. Plag 2009).

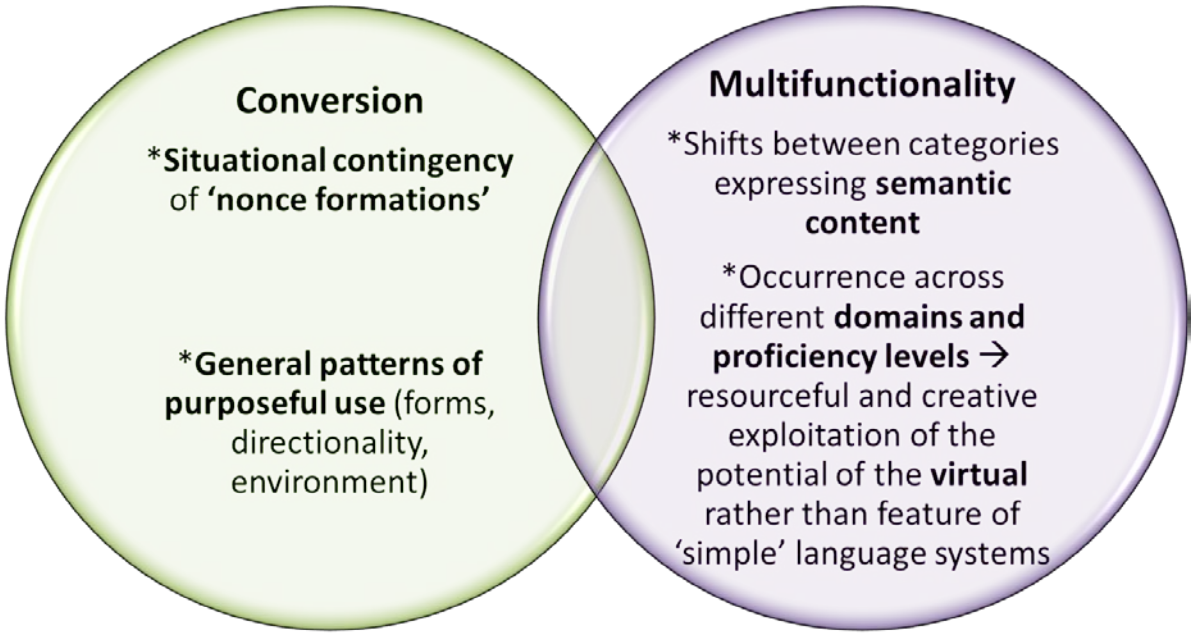

Figure 8: Application of the concepts of conversion and multifunctionality redefined for ELF

In the light of these findings we can argue that research into ELF can prompt a re-evaluation of traditional concepts to describe linguistic innovation, such as conversion and multifunctionality. As Seidlhofer (2011b: 95) argues, common concepts such as community, variety or competence are convenient constructs to describe language but may not always be appropriate to describe language in use. The ELF data at hand, which is based on recorded, naturally occurring and complete speech events, can help to reconstruct the process of innovation in interactive, spoken language in use. Moreover, it demonstrates that linguistic innovations are not limited to descriptions of language in stable speech communities. It also suggests that, in fact, notions of community, variety or competence need to be reconsidered not only for ELF data, but for language use in general, as any natural language use is likely to yield innovative language use when the need arises.

Finally, it becomes clear that initial 'problems' in corpus annotation can be important starting points for gaining insights into the data at hand. The investigation of word class variation in ELF was only possible because the discrepancy between non-codified forms and function was

not glossed over in the tagging process by displaying both paradigmatic form and syntagmatic function of a token. As such, an attempt was made to fit the necessarily pre-defined tagging categories to our data, rather than vice versa. This resonates with the approach that not all words can be assigned to exactly one grammatical category but that in 'real life', many words are "underdetermined" or "syntactically vague" (Denison 2013a: 21) argues that an oversimplification in tagging procedures, however necessary it may be, reaches the end of its usefulness when it comes to tokens which cannot be categorised unambiguously. Hence, Denison (2013a: 33) suggests that such vagueness should be "explicitly signal[led]" in tagging systems, either by ambiguous tags or other methods (ibid.) in order to mirror language as it is actually used (Denison 2007). For tagging the ELF data at hand, this was attempted in the way outlined above. As a result, the natural process of word class variation was made visible, rather than disguised in the annotation process.

6 CONCLUSIONS TO BE DRAWN FROM TAGGING VOICE

At the beginning of this thesis, two initial questions, or focus points, were addressed: The first was whether the transcripts of spoken ELF interactions in VOICE could meaningfully be assigned part-of-speech tags. The second was which theoretical implications arose from this practical enterprise.

6.1 The practical outcome: feasibility and limits of categorisation

The first question can, in principle, be answered clearly positively, as the finished, published products VOICE POS Online 2.0 and VOICE POS XML 2.0 demonstrate. However, this question also needs to be answered in the light of the lack of applicable tagging strategies for such data, as well as the challenges faced and compromises involved in assigning pre-defined categories to ELF data. With regard to this it emerged that the tagging of VOICE could in many respects not be based on available tagging solutions, as these do not, on the one hand place as much focus on displaying the spoken nature of their data, and on the other hand, because they do not address a number of issues related to the spoken and variable nature of VOICE data. Moreover, the feasibility of tagging VOICE needed to be viewed in the light of the challenges, both of a practical and theoretical nature, which were faced throughout the tagging process. Hence, the naturally occurring, transcribed data in VOICE was in many respects not straightforward to tag, and the solutions found were not always the easiest to implement. In sum, it was thus found that categorisation with conventional categories can indeed be applied to a large part of the transcripts in VOICE but that at the same time, this categorisation also has limits as not all aspects could readily be categorised with conventional POS categories. What has been pointed out in this thesis as crucial, however, is that the challenges this poses point to interesting aspects of the data on the one hand, and, on a more general level, to issues concerning the discrepancy between language categorisation and actual language use. As such, the tagging of ELF data is a discovery procedure as it sheds light on issues about ELF discourse and more generally, about processes involved in language use that the tagging of conventional written data would not engage with. It is argued that precisely those aspects of the data which posed challenges in that not a single tag could be assigned unambiguously, and for which no sufficient point of reference was available, refer us to certain characteristics, e.g. the fluidity, of spoken ELF data.

The result of tagging VOICE provides added value for corpus users on a practical level in two aspects in particular. Firstly, the tagging of VOICE makes the investigation of forms used in ELF data easier in general. For example, while before the tagging of VOICE, the data for investigating the use of the progressive or the zero-realisation or 3rd person -s in ELF had to

be carried out manually (separating the progressive from e.g. gerund/adjective forms), this can now be conducted automatically with the POS-tagged version (Osimk-Teasdale, Dorn & Schekulin 2014), and thus more time-efficiently and systematically. Secondly, it makes insights into the data possible which are difficult to envisage without POS tags, such as research into the transcendence of word class categories, as has been shown in chapter 5 of this thesis.

6.2 Theoretical implications of a linguistic approach to tagging

The issue that a certain proportion of tokens in VOICE could not be tagged unambiguously has a number of theoretical implications which go beyond the mere practical tagging task and outcome. Central to the meaningful application of POS tags to VOICE data, as well as an important pre-requisite to answering those theoretical issues raised by the challenges in tagging VOICE, was the prioritisation of a 'linguistic' rather than a 'computational' perspective on the tagging. This meant on the one hand that it was regarded a high priority to do justice to the data at hand, and hence display all the characteristics of spoken ELF. Such features are, even though they are reported to be present in other spoken corpora, absent or under-represented in the majority of existing tagsets. On the other hand this meant that problems were welcomed rather than glossed over and theoretical issues were foregrounded, as has been done in a small number of other approaches to tagging, e.g. by Rahman and Sampson (2000) or Rehbein, Schalowski and Wiese (2012a). These two characteristics of a linguistic approach, i.e. attempting to display actual language use accurately and linking practical tagging challenges to theoretical issues are evident in a number of individual corpus projects, such as KidKo, CHRISTINE or the corpus of spoken Irish. Interestingly, these projects also deal with naturally occurring and interactive, rather than elicited or monologic data. It is needless to say that naturally occurring language data is more remote from written forms than other types of spoken data, e.g. monologic or elicited, and cannot be as easily described with pre-defined linguistic categories. It might, therefore, be that conceptual issues present more pressing issues in corpora dealing with online, spoken data and that they, therefore, are more likely to be considered.

One of the main realisations in the tagging of VOICE was that, as mentioned earlier, the assigning of word classes to actual language in use, especially those parts which are more variable and non-codified, is often arbitrary. Hence, within the tagging of VOICE, the approach was confirmed that categorisation was not regarded a given, but always overtly viewed as subject to interpretation. Nevertheless, every decision was preceded by a thorough thought process in order to do justice to the display of actual language use within a pre-

defined set of word class categories. A result of this is that all tagging strategies and decisions made were documented and discussed at length in the *VOICE Part-of-Speech Tagging and Lemmatization Manual*. As such, the detailed description of the tagging of VOICE exists for a number of reasons: First, it is meant to make the data more accessible, and the reasons for tagging decisions more traceable for corpus users, so that they might be able to exploit the full possibilities of the POS tagged version of VOICE. Additionally, it might point to issues which might be worth investigating further within the corpus data. A third reason for documenting the products of our thought processes regarding the POS tagging of VOICE is that it might stimulate discussion regarding the categorisation of such language data as well as assist corpus compilers of similar types of complex data in the annotation of similar data.

In essence, three theoretical issues were raised by the tagging of VOICE:

The first issue concerns the applicability of external points of references to ELF data and the investigation of ELF data in its own right. This is an issue of general relevance to ELF research, which has been addressed multiple times before. However, in the tagging process of VOICE, these questions, often raised only in theory in the literature, became pressing issues in need of practical solutions. One the one hand, a clear definition of what terms such as *standard usage* or *native speaker usage* meant with regard to the practical tagging task was needed. On the other hand, the issue of how to display the ELF data in its own right was a very practical question which needed to be addressed, as decisions needed to be taken with regard to how each individual token in VOICE should be treated in the tagging procedure. In this context, it emerged that in ELF research, notions such as *standard usage*, *native speaker (usage)* and *prescriptive language* have not always been clearly defined, something that became necessary in the tagging of VOICE. Moreover, it was discussed how external points of reference can be dealt within ELF research and which purpose they can fulfil. It was pointed out that external references should be concretely defined, as opposed to remaining abstract and simplified notions. The listing and naming of the individual points of reference was a high priority in the tagging of VOICE. It has the advantage that by specifying these, their potential but also limits are highlighted, as well as the fact that notions such as ‘Standard English’ remain, to a certain extent, artificial and idealised constructs. At the same time, however, it was argued, that the use of references is justified also when aspiring to do justice to the variable nature of ELF data. This is because the use of external points of reference renders the tagging process feasible and enhances the usability of the tagged corpus as it is something which is widely known to corpus users. Moreover, the type of comparison drawn by the use of external points of references is useful for the description of ELF data and is

justified as some notion of Standard English is part of the ELF speakers' linguistic repertoires. In this context, it also emerged that the tagging standards such as EAGLES, as well as existing tagsets do not take into account features central to naturally occurring language use such as ELF. However, to display these is highly useful for the investigation of such features. Hence, in order to enhance comparability between corpora, it would be necessary to extend existing tagging standards with regard to issues such as displaying ambiguities and uncertain tag assignment, non-codified language use, in particular that of non-codified form-function relationships, as well as variability in spoken discourse.

A second issue that emerged and is of relevance to ELF, but also to SLA and learner corpus research is the distinction between language learners and users and their data, and the annotation of this data. In the POS tagging process of VOICE this was an issue as in corpus annotation, similar sets of data have been treated very differently and it needed to be decided if these procedures were justified from an ELF perspective and where VOICE data fitted in. I have tried to argue that the crucial difference between language learners and users is one of perspectives, rather than the communicative setting or language proficiency, and that this has major implications. These implications relate to the way the language output of users and learners is described in the literature, where language learners are generally viewed in direct comparison to the target language speakers whereas for users' language is assumed to be communicatively purposeful and to fulfil certain functions. This, in turn, has an influence on the different aspects of language output which are focussed on in the literature, depending on whether the speakers are defined as learners or users of the language, and on the conclusions which are drawn. Unfortunately, it also means that literature on learners and users is often not taken into account mutually, even though it has been argued that both fields could profit from such an exchange.

The third issue is that the existence of a clear bias towards written, standard language in linguistics is mirrored by the available tagging strategies. It is demonstrated by a lack of detailed representation of typical features of naturally occurring, spoken discourse in most tagsets and by the assimilation of such typical spoken characteristics to forms which are more neatly described with codifications of language, by strategies such as 'normalisation'. Such a bias can be said to exist towards more formal types of spoken language, which are more similar to written data. However, it is especially apparent with regard to spontaneous spoken, and especially non-native language. The investigation of non-native language has, in fact, been discarded as not worthy of investigation in many areas of linguistics and also with regard to part-of-speech tagging. And although a large body of ELF research has emerged

which demonstrates the desirability of investigating ELF data since the turn of the last century, it has been suggested even very recently that issues such as word class shifts in non-native data are not interesting because of their frequent occurrence of nonce-formations (cf. Denison 2013b: 177). However, as research on ELF has demonstrated, there is also something to be said about communication which takes place with a 'situationality factor' and amongst people who do not share a first language. This thesis suggests that precisely the display of such forms typical of data as it occurs in VOICE, which is spoken, naturally occurring, variable and plurilingual, is a necessary prerequisite for the investigation of such forms. Moreover, the study of word class variation in chapter 5 of this thesis demonstrates that the investigation of variable, spontaneous and non-native language data is worthwhile, and thus, adds another reason for the benefits of investigating ELF data.

In general, it has be shown in this thesis that a linguistic perspective on tagging has a number of advantages, as it establishes links between the practical task of POS tagging on the one hand, and theoretical questions of the categorisation of spoken language discourse on the other. It can, for example, contribute to the discussion of the suitability and limits of existing categories for the description of actual language use, and especially for ELF, which represents an especially dynamic form of language use.

In addition, the inherent openness of the linguistic approach to challenges in categorisation can also influence the finished tagging product positively. For example, in VOICE this led to the display of aspects of the data which are not easy to categorise with conventional methods. While a purely practical tagging activity would aim at removing or glossing over difficulties, or disregarding them as 'messy' components of actual language use, within a linguistic approach the motivation is to make these difficulties visible in the tagging scheme while at the same time recognising that it will never be possible to capture the reality of language use with pre-defined categories. Such a detailed display of the characteristics of spoken, plurilingual language data, cannot only point to aspects worthy of further exploration, but it can also facilitate the investigation of these. Such areas are not only the transcendence of word class categories as demonstrated in chapter 5 of this thesis. Areas are which can be, and indeed have been, facilitated by POS tagging are, for example, research into spoken aspects such as discourse markers, non-codified forms and form-function relationships, vagueness and ambiguity of forms and functions in ELF, the investigation of certain language forms which are used flexibly in VOICE, such as collective nouns, and tokens ending in the suffix *–ing* and that of fixed units, such as compound nouns, multiword items, formulaic items and multiword discourse markers.

Finally, it can be argued that the tagging of VOICE can serve as a starting point for the investigation of other, similar types of data, in terms of the linguistic approach adopted, as well as the key principles and practical tagging solutions which emerged from this approach. As explained above, the use of language in a globalised world, be it in English or other languages, is characterised by plurilingual language use and the exploitation of language resources. It is ever changing and fluid and not restricted to the use of English as a lingua franca, as recent research corpus compilation projects such as KiDKo show. For the investigation of such language use researchers need different tools and practical solutions to the conventional methods, as hithero applied tagging strategies of normalisation of all features of such language use, or the oversight of such data altogether, are insufficient. This is not to say that the tagging strategies applied to VOICE data are the most optimal or straight forwards solution to tagging such discourse. However, they present ways to approach such language use within pre-defined categories with strategies to deal with non-codified language use, ambiguities, variability and fixed units, as well as issues relating to basing the tagging on already existing transcripts. As such, the tagging of VOICE can hopefully contribute to the discussion of categorising naturally occurring, transcribed language data for research purposes, as well as, ultimately, to the discussion of the categories used for linguistic investigation in the light of actual language in use, in general.

REFERENCES

Aarts, Bas. 2004a. "Conceptions of gradience in the history of linguistics". *Language Sciences* 26, 343-389.

Aarts, Bas. 2004b. "Introduction: the nature of grammatical categories and their representation". In Aarts, Bas; Denison, David; Keizer, Evelien; Popova, Gergana (eds.). *Fuzzy grammar: a reader*. Oxford: Oxford University Press, 1-28.

Aarts, Bas. 2006. "Conceptions of categorization in the history of linguistics". *Language Sciences* 28, 361-385.

Aijmer, Karin. 2002. *English discourse particles: evidence from a corpus*. Amsterdam: Benjamins.

Allen, Will; Beal, Joan C.; Corrigan, Karen P.; Maguire, Warren; Moisl, Hermann L. 2007. "A linguistic 'time capsule': the Newcastle Electronic Corpus of Tyneside English". In Beal, Joan C.; Corrigan, Karen P.; Moisl, Hermann L. (eds.). *Creating and digitizing language corpora. Volume 2: diachronic databases*. Basingstoke: Palgrave Macmillan, 16-48.

Allwood, Jens; Björnberg, Maria; Grönqvist, Leif; Ahlsén, Elisabeth; Ottesjö, Cajsa. 2000. "The spoken language corpus at the department of linguistics, Göteborg University". *Forum: Qualitative Social Research* 1(3), 1-20.

American National Corpus Project. 2012. "American National Corpus". http://www.anc.org/ (10 July 2014).

Andersen, Gisle; Stenström, Anna-Brita. 1996. "COLT: a progress report". *ICAME journal* 20, 133-136.

Anward, Jan. 2000. "A dynamic model of part-of-speech differentiation". In Vogel, Petra Maria; Comrie, Bernard (eds.). *Approaches to the typology of word classes*. Berlin: de Gruyter, 3-45.

Aston, Guy. 1998. "Learning English with the British National Corpus". Paper presented at the 6th Jornada de Corpus, UPF, Barcelona, May 1998. http://www.sslmit.unibo.it/~guy/barc.htm (20 January 2015).

Atwell, Eric. 1996. "AMALGAM homepage: the Lancaster/IBM Spoken English Corpus (SEC) tag-set". http://www.comp.leeds.ac.uk/ccalas/tagsets/sec.html (17 January 2015).

Atwell, Eric. 2008. "Development of tag sets for part-of-speech tagging". In Lüdeling, Anke; Kytö, Meria (eds.). *Corpus Linguistics: an international handbook. Volume 1*. Berlin: de Gruyter, 501-527.

Atwell, Eric; Demetriou, George; Hughes, John; Schiffrin, Amanda; Souter, Clive; Wilcock, Sean. 2000. "A comparative evaluation of modern English corpus grammatical annotation schemes". *ICAME journal* 24, 7-23.

Baayen, R. Harald; Piepenbrock, Richard; Gulikers, Leon. 1995. "The CELEX Lexical Database". [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Baird, Robert; Baker, Will; Kitazawa, Mariko. 2014. "The complexity of ELF". *Journal of English as a Lingua Franca* 3(1), 171-196.

Baker, Will. 2011. "Intercultural awareness: modelling an understanding of cultures in intercultural communication through English as a lingua franca". *Language and Intercultural Communication* 11(3), 197-214.

Balteiro, Isabel. 2007. *A contribution to the study of conversion in English*. Münster: Waxmann Verlag.

Bas, Martha. 2010. *Identity negotiation among the users of English as a lingua franca in continental Europe*. MA thesis, University of Southampton.

Beal, Joan C.; Corrigan, Karen; Smith, Nicholas; Rayson, Paul. 2007 "Writing the vernacular: transcribing and tagging the Newcastle Electronic Corpus of Tyneside English (NECTE)". http://www.helsinki.fi/varieng/journal/volumes/01/beal_et_al/ (17 January 2015).

Beal, Joan C. 2009. "Creating corpora from spoken legacy materials: variation and change meet corpus linguistics". *Language and Computers* 69(1), 33-47.

Bellini, Daniele; Stefan, Schneider. 2006. "Spoken Italian online: the Banca dati del italiano parlato (BADIP)". In Kettemann, Bernhard (ed.). *Planing, gluing and painting corpora: inside the applied corpus linguist's workshop*. Frankfurt am Main: Lang, 13-26.

Biber, Douglas; Johansson, Stig; Leech, Geoffrey; Conrad, Susan; Finegan, Edward (eds.). 1999. Longman grammar of spoken and written English. Harlow: Longman.

Biermeier, Thomas. 2008. *Word-formation in new Englishes: a corpus-based analysis*. Münster: LIT.

Bjørge, Anne Kari. 2007. "Power distance in English lingua franca email communication". *International Journal of Applied Linguistics* 17(1), 60-80.

Björkman, Beyza. 2009. "From code to discourse in spoken ELF". In Mauranen, Anna; Ranta, Elina (eds.). *English as a lingua franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 225-251.

Böhringer, Heike. 2009. *The sound of silence: silent and filled pauses in English as a lingua franca business interaction*. Saarbrücken: VDM.

Böttger, Claudia; House, Juliane; Stachowicz, Roman. 2013. "Knowledge transfer on English as a lingua franca in written multilingual business communication". In Bührig, Kristin; Meyer, Bernd (eds.). *Transferring linguistic know-how into institutional practice*. Amsterdam: Benjamins, 117-136.

Brants, Thorsten. 2000. "TnT - A statistical part-of-speech tagger". In Nirenburg, Sergei (ed.). *Proceedings of the Sixth Applied Natural Language Processing Conference*. Seattle, WA: ACL, 224-231.

Breiteneder, Angelika. 2005. "The naturalness of English as a European lingua franca: the case of the 'third person -s'". *Vienna English Working PaperS* 14(2), 3-26.

Breiteneder, Angelika. 2009. "English as a lingua franca in Europe: an empirical perspective". *World Englishes* 28(2), 256-269.

Breiteneder, Angelika; Klimpfinger, Theresa; Majewski, Stefan; Pitzl, Marie-Luise. 2009. "The Vienna-Oxford International Corpus of English (VOICE) - A linguistic resource for exploring English as a lingua franca". *ÖGAI-Journal* 28(1), 21-26.

Breiteneder, Angelika; Pitzl, Marie-Luise; Majewski, Stefan; Klimpfinger, Theresa. 2006. "VOICE recording - methodological challenges in the compilation of a corpus of spoken ELF". *Nordic Journal of English Studies* 5(2), 161-188.

Brill, Eric. 1992. "A simple rule-based part of speech tagger". In ACL (ed.). *Proceedings of the 3rd conference on Applied Computational Linguistics*. Trento: ACL, 112-116.

Brill, Eric. 1995. "Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging". *Computational Linguistics* 21(4), 543-565.

Brill, Eric; Wu, Jun. 1998. "Classifier combination for improved lexical disambiguation". In ACL (ed.). *Proceedings of the 36th annual meeting of the ACL and 17th International Conference on Computational Linguistics. Volume 1*. Stroudsburg, PA: ACL, 191-195.

Bürki, Dominique. 2013. *Necessity and obligation in English as a Lingua Franca: a corpus-based grammatical variationist analysis*. MA thesis, University of Bern.

Burnard, Lou. 2010. "British National Corpus: the BNC in numbers". http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=numbers (18 August 2014).

Burt, Martina K. 1975. "Error analysis in the adult EFL classroom". *TESOL quarterly* 9(1), 53-63.

Bussmann, Hadumod (ed.). 1996. *Routledge dictionary of language and linguistics*. London: Routledge.

Cambridge Dictionaries Online. 2015 "them". http://dictionary.cambridge.org/dictionary/british/them_2. (3 February 2015)

Capone, Marzia. 2010. *Lexical innovations in ELF: the professional business domain in the VOICE corpus*. MA thesis, Università degli Studi di Verona.

Carey, Ray. 2014. "ELFA project: a research blog from the University of Helsinki's ELFA project". http://elfaproject.wordpress.com (27 August 2014).

Carter, Ronald. 1998. "Orders of reality: CANCODE, communication, and culture". *ELT Journal* 52(1), 43-56.

Carter, Ronald; McCarthy, Michael. 2006a. *Cambridge Grammar of English: a comprehensive guide to spoken and written English usage*. Cambridge: Cambridge University Press.

Carter, Ronald; McCarthy, Michael. 2006b. *Cambridge Grammar of English: a comprehensive guide to spoken and written English usage.* [CD-ROM]. Cambridge: Cambridge University Press.

Church, Kenneth W. 1988. "A stochastic parts program and noun phrase parser for unrestricted text". In ACL (ed.). *Proceedings of the 2nd conference on Applied Natural Language Processing*. Stroudsburg, PA: ACL, 136-143.

Clayson-Knollmayr, Beate. 2010. "'Drop me an e-mail when draft is ready.': register and style in ELF business e-mails". Paper presented at the Third International Conference of English as a Lingua Franca, University of Vienna, Vienna, 22-25 May 2010.

Cogo, Alessia. 2008. "English as a Lingua Franca: form follows function. Some clarifications on the nature of ELF". *English Today* 95(3), 58-61.

Cogo, Alessia. 2009. "Accommodating difference in ELF conversations: a study of pragmatic strategies". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 254-273.

Cogo, Alessia. 2012. "ELF and super-diversity: a case study of ELF multilingual practices from a business context". *Journal of English as a Lingua Franca* 1(2), 287-313.

Cogo, Alessia; Dewey, Martin. 2006. "Efficiency in ELF communication: from pragmatic motives to lexico-grammatical innovation". *Nordic Journal of English Studies* 5(2), 59-93.

Cook, Vivian. 2002a. "Background to the L2 user". In Cook, Vivian (ed.). *Portraits of the L2 user*. Clevedon: Multilingual Matters, 1-28.

Cook, Vivian (ed.). 2002b. *Portraits of the L2 user*. Clevedon: Multilingual Matters.

Cresti, Emanuela; Moneglia, Massimo. 2005. *C-ORAL-ROM: integrated reference corpora for spoken Romance languages*. Amsterdam/Philadelphia: Benjamins.

Crystal, David. 1967. "English". *Lingua* 17, 24-56.

Cutting, Doug; Kupiec, Julian; Pedersen, Jan; Sibun, Penelope. 1992. "A practical part-of-speech tagger". In ACL (ed.). *Proceedings of the 3rd conference on Applied Computational Linguistics*. Trento: ACL, 133-140.

Davies, Mark. 2008-. "Corpus of Contemporary American English (COCA)". http://corpus.byu.edu/coca/ (8 July 2014).

Davies, Mark. 2009. "The 385+ million word *Corpus of Contemporary American English* (1990–2008+): design, architecture, and linguistic insights". *International Journal of Corpus Linguistics* 14(2), 159-190.

Dawa, Idomuco; Husal, L. Y.; Ming, Yue Yao; Uulang; Cheng, Bai Shuang; Batsaihan, Y.; Mitsunaga, M.; Isahara, Hitoshi; Nakamura, Satoshi. 2006. "Multilingual text–speech corpus of Mongolian". In *Proceedings of the International Symposium on Chinese Spoken Language Processing*. Kent Ridge, Singapore: ISCA, 759-770.

Denison, David. 2007. "Playing tag with category boundaries". In Meurman-Solin, Anneli; Nurmi, Arja (eds.). *Studies in variation, contacts and change in English: annotating variation and change. Volume 1*. University of Helsinki: VARIENG.

Denison, David. 2008. "Clues to language change from non-standard English". *German Life and Letters* 61(4), 533-545.

Denison, David. 2013a. "Grammatical mark-up: some more demarkation disputes". In Bennett, Paul; Durrell, Martin; Scheible, Silke; Whitt, Richard J. (eds.). *New methods in historical corpora*. Tübingen: Narr, 17-35.

Denison, David. 2013b. "Parts of speech: solid citizens or slippery customers?". *Journal of the British Academy* 1, 151-185.

Dewey, Martin. 2009. "English as a Lingua Franca: heightened variability and theoretical implications". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 60-83.

Díaz-Negrillo, Ana; Meurers, Detmar; Valera, Salvador; Wunsch, Holger. 2010. "Towards interlanguage POS annotation for effective learner corpora in SLA and FLT". In Moreno Jaén, María; Pérez Basanta, Carmen (eds.). Language Forum 36(1-2), 139-154.

Diemer, Stefan; Schmidt, Selina; Brunner, Marie-Luise. forthc. "Ja, also, can you see me now? Designing and compiling CASE: a corpus of computer-mediated international academic spoken English". *ICAME Journal. Special Issue on Corpus Compilation*.

Dorn, Nora. 2011. *Exploring* –ing*: the progressive in English as a lingua franca*. Saarbrücken: VDM.

Dorn, Nora; Rienzner, Martina; Busch, Brigitta; Santner-Wolfartsberger, Anita. 2014. "'Here I find myself to be judged': ELF/plurilingual perspectives on language analysis for the determination of origin". *Journal of English as a Lingua Franca* 3(2), 409-424.

Ehrenreich, Susanne. 2009. "English as a Lingua Franca in multinational corporations - exploring business communities of practice". In Mauranen, Anna; Ranta, Elina (eds.).

English as a Lingua Franca: studies and findings. Newcastle upon Tyne: Cambridge Scholars Publishing, 126-151.

Eisikovits, Edina. 1987. "Variation in the lexical verb in Inner-Sydney English". *Australian Journal of English* 7, 1-24.

Fellbaum, Christiane (ed.). 1998. *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press.

Garside, Roger. 1995. "Grammatical tagging of the spoken part of the British National Corpus: a progress report". In Leech, Geoffrey; Myers, Greg; Thomas, Jenny (eds.). *Spoken English on computer*. London: Longman, 161-167.

Garside, Roger. 1996. "The robust tagging of unrestricted text: the BNC experience". In Thomas, Jenny; Short, Mick (eds.). *Using corpora for language research: studies in the honour of Geoffrey Leech*. London: Longman, 167-180.

Garside, Roger; Leech, Geoffrey; McEnery, Tony (eds.). 1997. *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

Garside, Roger; Smith, Nicholas. 1997. "A hybrid grammatical tagger: CLAWS4". In Garside, Roger; Leech, Geoffrey; McEnery, Tony (eds.). *Corpus annotation: linguistic information from computer text corpora*. London: Longman, 102-121.

Gibbon, Dafydd; Mertins, Inge; Moore, Roger K. (eds.). 2000. *Handbook of multimodal and spoken dialogue systems: resources, terminology, and product evaluation*. Norwell, MA: Kluwer.

Giesbrecht, Eugenie; Evert, Stefan. 2009. "Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus". In Inaki, Alegria; Leturia, Igor; Sharo, Serge (eds.). *Proceedings of the fifth web as corpus workshop*. San Sebastian: Elhuyar Fundazioa, 27-35.

Gilquin, Gaëtanelle. 2008. "Hesitation markers among EFL learners: pragmatic deficiency or difference". In Romero-Trillo, Jesús (ed.). *Pragmatics and corpus linguistics: a mutualistic entente*. Berlin: de Gruyter, 119-149.

Gilquin, Gaëtanelle; Cock, Sylvie de. 2011. "Errors and disfluencies in spoken corpora: setting the scene". *International Journal of Corpus Linguistics* 16(2), 141-172.

Giménez, Jesús; Marquez, Lluis. 2003. "Fast and accurate part-of-speech tagging: the SVM approach revisited". In Nicolov, Nicolas (ed.). *Recent advantages in natural language processing III: selected papers from RANLP 2003*. Amsterdam: Benjamins, 153-163.

Gimpel, Kevin; Schneider, Nathan; O'Connor, Brendan; Das, Dipanjan; Mills, Daniel; Eisenstein, Jacob; Heilman, Michael; Yogatama, Dani; Flanigan, Jeffrey; Smith, Noah A. 2011. "Part-of-Speech tagging for Twitter: annotation, features, and experiments". In ACL (ed.). *Proceedings of the 49th annual meeting of the Association for Computational Linguistics Human Language Technologies. Volume 2: short papers pages*. Madison, WI: Omnipress, 42-47.

Godfrey, John J.; Holliman, Edward C.; McDaniel, Jane. 1992. "SWITCHBOARD: telephone speech corpus for research and development". In *IEEE International Conference on Acoustics, Speech, and Signal Processing.* San Francisco, CA: IEEE, 517-520.

Goutsos, Dionysis. 2010. "The corpus of Greek texts: a reference corpus for Modern Greek". *Corpora* 1(5), 29-44.

Granger, Sylviane. 2008. "Learner corpora". In Lüdeling, Anke; Kytö, Merja (eds.). *Corpus linguistics: an international handbook*. Berlin: de Gruyter, 259-275.

Granger, Sylviane; Rayson, Paul. 1998. "Automatic profiling of learner texts". In Granger, Sylviane (ed.). *Learner English on computer*. London: Longman, 119-131.

Greenbaum, Sidney; Nelson, Gerald. 1996. "The International Corpus of English (ICE) project". *World Englishes* 15(1), 3-15.

Greenbaum, Sidney; Svartvik, Jan. 1990. "The London-Lund Corpus of Spoken English". In Svartvik, Jan (ed.). *The London Corpus of Spoken English: description and research.* Lund: Lund University Press.

Greenbaum, Sidney; Yibin, Ni. 1994. "Tagging the British ICE corpus: English word classes". In Oostdijk, Nelleke; Haan, Pieter de (eds.). *Corpus-based research into language: in honour of Jan Aarts*. Amsterdam: Rodopi, 33-45.

Guido, Maria Grazia. 2012. "ELF authentication and accommodation strategies in crosscultural immigration encounters". *Journal of English as a Lingua Franca* 1(2), 219-240.

Güngör, Tunga. 2010. "Part-of-speech tagging". In Indurkhya, Nitin; Damerau, Fred J. (eds.). *Handbook of natural language processing*. (2nd edition). Boca Raton: CRC Press, 205-235.

Gustafson-Capková, Sofia; Hartmann, Britt. 2006. "Manual of the Stockholm Umeå Corpus version 2.0". http://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf (6 February 2013).

Halteren, Hans van. 1999a. "Performance of taggers". In Halteren, Hans van (ed.). *Syntactic wordclass tagging*. Dordrecht: Kluwer, 81-94.

Halteren, Hans van (ed.). 1999b. *Syntactic wordclass tagging*. Dordrecht: Kluwer.

Hamaker, Janna S. 1999. "Towards building a better language model for SWITCHBOARD: the POS tagging task". In *ICIIS '99 Proceedings of the 1999 International Conference on Information Intelligence and Systems*. Piscataway, NJ: I.E.E.E. Press, 579-582.

Haser, Verena; Auer, Anita; Botma, Bert; Elenbaas, Marion; Gyuris, Beáta; Allan, Kathryn; Mehl, Seth; Vobornik, Erin; Anderwald, Liselotte; Schröder, Anne; Miličević, Maja; Kraš, Tihana; Schekulin, Claudio; Dorn, Nora; Callies, Marcus; Montoro, Rocío. 2014. "English language". *The Years Work in English Studies* 93(1), 1-174.

Haser, Verena; Auer, Anita; Botma, Bert; Elenbaas, Marion; Wurff, Wim van der; Gyuris, Beáta; Allan, Kathryn; Vobornik, Erin; Anderwald, Liselotte; Schröder, Anne; Lozano, Cristóbal; Hülmbauer, Cornelia; Santner-Wolfartsberger, Anita; Schekulin, Claudio; Ollinger, Astrid; Callies, Marcus; Montoro, Rocío. 2013. "English language". *The Year's Work in English Studies* 92(1), 1-180.

Haser, Verena; Auer, Anita; Botma, Bert; Elenbaas, Marion; Wurff, Wim van der; Gyuris, Beáta; Allen, Kathryn; Cross, Erin; Anderwald, Liselotte; Schröder, Anne; Seidlhofer, Barbara; Dorn, Nora; Callies, Marcus; Montoro, Rocío. 2012. "English language". *The Year's Work in English Studies* 91(1), 1-177.

Hasselgren, Angela. 2002. "Learner corpora and language testing: smallwords as markers of learner fluency". In Granger, Sylviane; Hung, Joseph; Petch-Tyson, Stephanie (eds.). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: Benjamins, 143-173.

Hirschmann, Hagen; Doolittle, Seanna; Lüdeling, Anke. 2007. "Syntactic annotation of non-canonical linguistic structures". In Davies, Matthew; Rayson, Paul; Hunston, Susan; Danielsson, Pernilla (eds.). *Proceedings of the Corpus Linguistics Conference CL2007*. Birmingham: University of Birmingham, 1-15.

Holzschuh, Elisabeth. 2013. *Discourse marker use in ELF conversations: the case of 'you know'*. MA Thesis, University of Vienna.

Hornby, Albert Sidney; Wehmeier, Sally; McIntosh, Colin; Turnbull, Joanna; Ashby, Michael. 2005. *Oxford advanced learner's dictionary*. (7th edition). Oxford: Oxford University Press.

Huddleston, Rodney D.; Pullum, Geoffrey K.; Bauer, Laurie. 2010. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Hudson-Ettle, Diana M.; Schmied, Josef. 1999. "Manual to accompany The East African Component of The International Corpus of English ICE-EA: background information, coding conventions and lists of source texts". http://clu.uni.no/icame/manuals/ICE_EA.PDF (24 January 2015).

Hüllen, Werner. 1982. "Teaching a foreign language as 'lingua franca'". *Grazer Linguistische Studien* 16, 83-88.

Hülmbauer, Cornelia. 2009. "'We don't take the right way. We just take the way that we think you will understand' – the shifting relationship between correctness and effectiveness in ELF". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 323-347.

Hülmbauer, Cornelia. 2010. *English as a Lingua Franca between correctness and effectiveness: shifting constellations*. Saarbrücken: VDM.

Hülmbauer, Cornelia. 2013. "From within and without: the virtual and the plurilingual in ELF". *Journal of English as a Lingua Franca* 2(1), 47-73.

Hülmbauer, Cornelia; Seidlhofer, Barbara. 2011. "Common sense and common practice: ELF research and conceptions of multilingualism". *Unpublished Working Paper DYLAN 5*, 174-220.

Hülmbauer, Cornelia; Seidlhofer, Barbara. 2013. "English as a lingua franca in European multilingualism". In Berthoud, Anne-Claude; Grin, François; Lüdi, Georges (eds.). *Exploring the dynamics of multilingualism: the DYLAN project*. Amsterdam: Benjamins, 387-406.

ICE Project. 2013. "POS-tagged and semantically tagged Corpora". http://ice-corpora.net/ice/taggeddownload.htm (7 February 2015).

ICE Corpora. 2009a. "Indexpage". http://ice-corpora.net/ICE/INDEX.HTM (8 July 2014).

ICE Corpora. 2009b. "The design of ICE Corpora @ ICE-corpora.net". http://ice-corpora.net/ice/design.htm (4 February 2015).

Ide, Nancy; Christiane, Fellbaum; Baker, Collin; Rebecca, Passonneau. 2010. "The manually annotated sub-corpus: a community resource for and by the people". In ACL (ed.). *Proceedings of the ACL 2010 conference short papers*. Stroudsburg, PA: ACL, 68-73.

Ingvarsdóttir, Hafdís; Arnbjörnsdóttir, Birna. 2013. "ELF and academic writing: a perspective from the Expanding Circle". *Journal of English as a Lingua Franca* 2(1), 123-145.

Jendryczka-Wierszycka, Joanna. 2009. "Collecting spoken learner data: challenges and benefits". In Mahlberg, Michaela; González-Díaz, Victorina; Smith, Catherine (eds.). *Proceedings of the Corpus Linguistics Conference CL2009*. Liverpool: University of Liverpool.

Jendryczka-Wierszycka, Joanna; Rayson, Paul; Hoffmann, Sebastian. 2009. "Spoken learner corpus & its POS tagging".

http://www.ling.lancs.ac.uk/groups/crg/files/CRG09_wk30_JJW_slides.pdf (24 January 2015).

Jenkins, Jennifer. 2000. *The phonology of English as an international language: new models, new norms, new goals*. Oxford: Oxford University Press.

Jenkins, Jennifer. 2005. *World Englishes: a resource book for students*. London: Routledge.

Jenkins, Jennifer. 2006. "The spread of EIL: a testing time for testers". *ELT Journal* 60(1), 42-50.

Jenkins, Jennifer. 2009. *World Englishes: a resource book for students*. (2nd edition). London: Routledge.

Jenkins, Jennifer; Cogo, Alessia; Dewey, Martin. 2011. "Review of developments in research into English as a lingua franca". *Language Teaching* 44(3), 281-315.

Jenkins, Jennifer; Leung, Constant. 2013. "English as a Lingua Franca". In Kunnan, Antony John (ed.). *The companion to language assessment*. Hoboken, NJ: Wiley, 1605-1616.

Jenks, Christopher. 2013. "'Your pronunciation and your accent is very excellent': orientations of identity during compliment sequences in English as a lingua franca encounters". *Language and Intercultural Communication* 13(2), 165-181.

Jespersen, Otto. 1924. *The philosophy of grammar*. London: Allen & Unwin.

Johannessen, Janne Bondi; Hagen, Kristin; Priestley, Joel James; Nygaard, Lars. 2007. "An advanced speech corpus for Norwegian". In Nivre, Joakim; Kaalep, Heiki-Jaan; Muischnek, Kadri; Koit, Mare (eds.). *Proceedings of NODALIDA-2007*. Tartu: University of Tartu, 29-36.

Johannessen, Janne Bondi; Jørgensen, Fredrik. 2006. "Annotating and parsing spoken language". In Henrichsen, Peter J.; Skadhauge, Peter R. (eds.). *Proceedings of NODALIDA 2005 special session: treebanking for discourse and speech*. Copenhagen: Samfundslitteratu, 83-103.

Johannessen, Janne Bondi; Priestley, Joel; Hagen, Kristin; Vangsnes, Øystein Alexander. 2009. "The Nordic Dialect Corpus – an advanced research tool". In Jokinen, Kristiina; Bick, Eckhard (eds.). *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. Odense: NEALT, 73-80.

Joos, Martin. 1950. "Description of language design". *The Journal of the Acoustical Society of America* 22(6), 701-707.

Jørgensen, Fredrik. 2007. "Strategies for tagging transcribed spoken Norwegian". *StatMet Term Paper*, 1-7.

Kachru, Braj. 1985. "Standards, codification and sociolinguistic realism". In Quirk, Randolph (ed.). *English in the world*. Cambridge: Cambridge University Press, 11-34.

Kankaanranta, Anne. 2006. "'Hej Seppo, could you pls comment on this!': internal email communication in Lingua Franca English in a multinational company". *Business Communication Quarterly* 69(2), 216-225.

Kaszubski, Przemysław. 2003. "TOSCA-ICLE tagset". http://ifa.amu.edu.pl/~kprzemek/corpora/TOSCA-ICLE_tagset.htm#FULL%20LIST (19 August 2014).

Kaur, Jagdish. 2009. "Pre-empting problems of understanding in English as a lingua franca". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 107-123.

KiezDeutsch-Korpus. 2015. "Index page". http://www.kiezdeutschkorpus.de/ (11 July 2014).

Kirkpatrick, Andy. 2010. "Researching English as a Lingua Franca in Asia: the Asian Corpus of English (ACE) project". *Asian Englishes* 13(1), 4-18.

Kirkpatrick, Andy. 2013. "The Asian Corpus of English: motivation and aims". *Learner Corpus Studies in Asia and the World* 1, 17-30.

Klimpfinger, Theresa. 2009. "'She's mixing the two languages together' – forms and functions of code-switching in English as a Lingua Franca". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 349-371.

Klötzl, Svitlana. 2014. "'Maybe just things we grew up with': linguistic and cultural hybridity in ELF couple talk". *Journal of English as a Lingua Franca* 3(1), 27-48.

Knapp, Karlfried. 1985. "Englisch als internationale lingua franca und Richtlinien [English as international lingua franca and guidelines]". In Bausch, Karl-Richard; Christ, Herbert; Hüllen, Werner; Krumm, Hans-Jürgen (eds.). *Forschungsgegenstand Richtlinien* [Object of research guidelines]. Tübingen: Narr, 84-90.

Knapp, Karlfried. 1987. "English as an international lingua franca and the teaching of intercultural communication". In Lörsche, Wolfgang; Schulze, Rainer (eds.). *Perspectives on language in performance*. Tübingen: Narr, 1022-1039.

Kokkinakis Johansson, Sofie. 2003. *En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS [A study on relevant factors in wordclass tagging based on the example of tagging Swedish text with EPOS]*. Göteborg: Göteborg University.

Kortmann, Bernd; Schneider, Edgar W. (eds.). 2006. "Varieties of English: a multimedia reference tool". http://www.varieties.mouton-content.com (29 August 2013).

Labov, William. 1973. "The boundaries of words and their meanings". In Bailey, Charles-James N.; Shuy, Roger W. (eds.). *New ways of analyzing variation in English*. Washington: Georgetown University Press, 340-373.

Lee, David. 2010. "Bookmarks for corpus-based linguists". http://www.uow.edu.au/~dlee/CBLLinks.htm (17 July 2014).

Leech, Geoffrey. 1997. "A brief users' guide to the grammatical tagging of the British National Corpus". http://www.natcorp.ox.ac.uk/docs/gramtag.html (17 August 2014).

Leech, Geoffrey. 2000. "Grammars of spoken English: new outcomes of corpus-oriented research". *Language Learning* 50(4), 675-724.

Leech, Geoffrey. 2004. "Adding linguistic annotation". In Wynne, Martin (ed.). *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books, 17-29.

Leech, Geoffrey. 2007. "Corpora". In Teubert, Wolfgang; Krishnamurthy, Ramesh (eds.). *Corpus linguistics: critical concepts in linguistics*. London: Routledge, 3-17.

Leech, Geoffrey; Garside, Roger; Bryant, Michael. 1994a. "CLAWS4: the tagging of the British National Corpus". In ACL (ed.). *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*. Kyoto, Japan: ACL, 622-628.

Leech, Geoffrey; Garside, Roger; Bryant, Michael. 1994b. "The large-scale grammatical tagging of text: experience with the British National Corpus". In Oostdijk, Nelleke; Haan, Pieter de (eds.). *Corpus-based research into language*. Amsterdam: Rodopi, 47-63.

Leech, Geoffrey; Smith, Nicholas. 1999. "The use of tagging". In Halteren, Hans van (ed.). *Syntactic wordclass tagging*. Dordrecht: Kluwer, 23-36.

Leech, Geoffrey; Smith, Nicholas. 2000. "Manual to accompany The British National Corpus (Version 2) with improved word-class tagging". http://ucrel.lancs.ac.uk/bnc2/bnc2postag_manual.htm (12 August 2014).

Lenk, Uta. 1998. *Marking discourse coherence: functions of discourse markers in spoken English*. Tübingen: Narr.

Lichtkoppler, Julia. 2007. "'Male. Male.' — 'Male?' — 'The sex is male.': the role of repetition in English as a lingua franca conversations". *Vienna English Working PaperS* 16(1), 39-65.

Lindsay, Jean; O'Connell, Daniel C. 1995. "How do transcribers deal with audio recordings of spoken discourse?". *Journal of psycholinguistic research* 24(2), 101-115.

Linell, Per. 2005. *The written language bias in linguistics: its nature, origins and transformations*. London: Routledge.

Linguistic Data Consortium. 1999. "Addendum to the part-of-speech tagging guidelines for the Penn Treebank project (Modifications for the SwitchBoard corpus)". http://www.cis.upenn.edu/~bies/manuals/tagguid2.pdf (4 February 2015).

Lo Bianco, Joseph. 2014. "Dialogue between ELF and the field of language policy and planning". *Journal of English as a Lingua Franca* 3(1), 197-213.

Lüdeling, Anke; Kytö, Meria (eds.). 2008. *Corpus Linguistics: an international handbook. Volume 1*. Berlin: de Gruyter.

Luraghi, Silvia; Parodi, Claudia. 2008. *Key terms in syntax and syntactic theory*. London: Continuum.

Luzón, María José; Campoy, Mari Carmen; Del Mar Sánchez, María; Salazar, Patricia. 2007. "Spoken corpora: new perspectives in oral language use and teaching". In Campoy, Mari Carmen; Luzón, María José (eds.). *Spoken corpora in applied linguistics*. Bern: Lang, 3-30.

MacWhinney, Brian. 2009. "Enriching CHILDES for morphosyntactic analysis". *Department of Psychology* (Paper 175), 1-57.

MacWhinney, Brian. 2012a. "Morphosyntactic analysis of the CHILDES and TalkBank corpora". In Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry; Doğan, Mehmet Uğur; Maegaard, Bente; Mariani, Joseph; Odijk, Jan; Piperidis, Stelios (eds.). *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: ELRA, 2357-2380.

MacWhinney, Brian. 2012b. "The CHILDES Project: tools for analyzing talk – electronic edition. Part 1: the CHAT transcription format". http://childes.psy.cmu.edu/manuals/chat.pdf (14 Feburary 2013).

MacWhinney, Brian. 2012c. "The CHILDES Project: tools for analyzing talk – electronic edition. Part 2: the CLAN programs". http://childes.psy.cmu.edu/manuals/clan.pdf (14 Februar 2013).

Maekawa, Kikuo; Koiso, Hanae; Furui, Sadaoki; Isahara, Hitoshi. 2000. "Spontaneous speech corpus of Japanese". In Gavrilidou Maria; Crayannis George; Markantonatu Stella; Piperidis, Stelios; Stainhaouer, Gregory (eds.). *Proceedings of LREC-2000*. Athens, Greece: ELRA, 947-952.

Mair, Christian. 2004. "Corpus linguistics and grammaticalization theory". In Lindquist, Hans; Mair, Christian (eds.). *Corpus approaches to grammaticalization in English*. Amsterdam: Benjamins, 121-150.

Majewski, Stefan. 2011. *Design and implementation of a research infrastructure for a corpus of spoken ELF*. MA Thesis, University of Vienna.

Majewski, Stefan, 2015. Email correspondence, 18 January.

Manning, Christopher D. 2011. "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?". In Gelbukh, Alexander F. (ed.). *Computational linguistics and intelligent text processing. 12th International Conference, CICLing 2011. Lecture Notes in Computer Science 6608.* Berlin: Springer (6608), 171-189.

Martinez, Angel R. 2012. "Part-of-speech tagging". *Wiley Interdisciplinary Reviews: Computational Statistics* 4(1), 107-113.

Märzinger, Katrin. 2012. *The use of phrasal verbs in English as a Lingua Franca.* MA thesis, University of Vienna.

Mauranen, Anna. 2006. "Signalling and preventing misunderstanding in ELF communication". *International Journal of the Sociology of Language* 177, 123-150.

Mauranen, Anna. 2009. "Chunking in ELF: expressions for managing interaction". *Intercultural Pragmatics* 6(2), 217-233.

Mauranen, Anna. 2011. "Learners and users – who do we want corpus data from?". In Meunier, Fanny; Granger, Sylviane; Gilquin, Gaëtanelle; Paquot, Magali (eds.). *A taste for corpora: in honour of Sylviane Granger*. Amsterdam: Benjamins, 155-171.

Mauranen, Anna. 2012. *Exploring ELF: academic English shaped by non-native speakers*. Cambridge: Cambridge University Press.

McArthur, Tom. 1998. *The English languages*. Cambridge: Cambridge University Press.

McEnery, Tony; Hardie, Andrew. 2012. *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press.

McNamara, Tim. 2012. "English as a lingua franca: the challenge for language testing". *Journal of English as a Lingua Franca* 1(1), 199-202.

Metashare. "The Corpus of Ingrian Finnish". http://urn.fi/urn:nbn:fi:lb-2014073032 (21 August 2014).

Meurers, Detmar; Wunsch, Holger. 2010. "Linguistically annotated learner corpora: aspects of a layered linguistic encoding and standardized representation". http://www.sfs.uni-tuebingen.de/~dm/papers/meurers-wunsch-10.pdf (27 January 2015).

Mukherjee, Joybrato. 2007. "Exploring and annotating a spoken English learner corpus: a work-in-progress report". In Volk-Birke, Sabine; Lippert, Julia (eds.). *Anglistentag 2006 Halle: Proceedings*. Trier: WVT, 365-375.

Müller, Simone. 2005. *Discourse markers in native and non-native English discourse*. Amsterdam: Benjamins.

Murata, Kumiko; Jenkins, Jennifer. 2009. *Global Englishes in Asian contexts: current and future debates*. Basingstoke: Palgrave.

Nakagawa, Tetsuji; Taku, Kudo; Yuji, Matsumoto. 2002. "Revision learning and its application to part-of-speech tagging". In ACL (ed.). *Proceedings of the 40th annual meeting on Association for Computational Linguistics (ACL '02)*. Stroudsburg, PA: ACL, 497-504.

Nelson, Gerald. 2005. "The ICE tagging manual". ice-corpora.net/ice/TaggingManual.doc (27 January 2015).

Nivre, Joakim; Grönqvist, Leif. 2001. "Tagging a corpus of spoken Swedish". *International Journal of Corpus Linguistics* 6(1), 47-78.

Nøklestad, Anders; Søfteland, Ashild. 2007. "Tagging a Norwegian speech corpus". In Nivre, Joakim; Kaalep, Heiki-Jaan; Muischnek, Kadri; Koit, Mare (eds.). *NODALIDA 2007 Conference proceedings*. Tartu: University of Tartu, 245-248.

Oostdijk, Nelleke. 2000. "The spoken Dutch corpus: overview and first evaluation". In Gavrilidou Maria; Crayannis George; Markantonatu Stella; Piperidis, Stelios; Stainhaouer, Gregory (eds.). *Proceedings of LREC-2000*. Athens, Greece: ELRA, 887-894.

Ortega, Lourdes. 2010. "The bilingual turn in SLA". Plenary delivered at the Annual Conference of the American Association for Applied Linguistics, Atlanta, GA, 6-9 March 2010. http://www2.hawaii.edu/~lortega/Ortega_AAAL_2010.ppt (27 January 2015).

Osimk, Ruth. 2009. "Decoding sounds: an experimental approach to intelligibility in ELF". *Vienna English Working PaperS* 18(1), 64-89.

Osimk, Ruth. 2010. "Testing the intelligibility of ELF sounds". *Speak Out!* 42, 14-18.

Osimk, Ruth; Breiteneder, Angelika. 2011. "Multiple voices: highlighting aspects of form and function in ELF". Paper presented at the 16th World Congress of Applied Linguistics, Beijing Foreign Studies University, Beijing, 23-28 August 2011.

Osimk-Teasdale, Ruth. 2012. "'You can actually perform good in another language if you put your mind to it': conversion in ELF". Paper presented at the 5th International Conference of English as a Lingua Franca, Boğaziçi University, Istanbul, 24-26 May 2012.

Osimk-Teasdale, Ruth. 2013. "Applying existing tagging practices to VOICE". In Mukherjee, Joybrato; Huber, Magnus (eds.). *Corpus Linguistics and variation in English: Focus on non-native Englishes (Proceedings of ICAME 31).* Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/13/osimk-teasdale/ (7 February 2015)

Osimk-Teasdale, Ruth. 2014. "'I just wanted to give a partly answer': capturing and exploring word class variation in ELF data". *Journal of English as a Lingua Franca* 3(1), 109-143.

Osimk-Teasdale, Ruth; Dorn, Nora. forthc. "Categorising the unconventional: potential and limits of existing points of reference for POS-tagging VOICE". *International Journal of Corpus Linguistics.*

Osimk-Teasdale, Ruth; Dorn, Nora; Schekulin, Claudio. 2014. "POS-tagging the Vienna-Oxford International Corpus of English: strategies and rewards". Paper presented at OLINCO Conference, Palacký University, Olomouc, 4-7 June 2014.

Pavesi, Maria. 1998. "Same word same idea: conversion as a word formation process". *International Review of Applied Linguistics* 34(3), 213-231.

Pietikäinen, Kaisa S. 2014. "ELF couples and automatic code-switching". *Journal of English as a Lingua Franca* 3(1), 1-26.

Pirk, Ana Monika. 2013. "Construction grammar and 'non-native discourse'". *Theory and practice in English Studies* 6(1), 55-72.

Pitzl, Marie-Luise. 2005. "Non-understanding in English as a lingua franca: examples from a business context". *Vienna English Working PaperS* 14(2), 50-71.

Pitzl, Marie-Luise. 2010. *English as a lingua franca in international business: resolving miscommunication and reaching shared understanding*. Saarbrücken: VDM.

Pitzl, Marie-Luise. 2011. *Creativity in English as a lingua franca: idiom and metaphor*. PhD Thesis, University of Vienna.

Pitzl, Marie-Luise. 2012. "Creativity meets convention: idiom variation and remetaphorization in ELF". *Journal of English as a Lingua Franca* 1(1), 27-55.

Pitzl, Marie-Luise; Breiteneder, Angelika; Klimpfinger, Theresa. 2008. "A world of words: processes of lexical innovation in VOICE". *Vienna English Working PaperS* 17(2), 21-46.

Plag, Ingo. 1999. *Morphological productivity: structural constraints in English derivation.* Berlin: de Gruyter.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.

Plag, Ingo. 2009. "Creoles as interlanguages: word-formation". *Journal of Pidgin and Creole Languages* 24(2), 339-362.

Pravec, Norma A. 2002. "Survey of learner corpora". *ICAME journal* 26(81), 81-114.

Pullin Stark, Patricia. 2009. "No Joke—this is serious! Power, solidarity and humour in Business English as a Lingua Franca (BELF)". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 152-177.

Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffrey; Svartvik, Jan. (ed.). 1997. *A comprehensive grammar of the English language*. (14th edition). London: Longman.

Radeka, Michael. in prep. *VOICE: a parallel stacked TBL approach*. PhD Thesis, University of Vienna.

Rahman, Anna; Sampson, Geoffrey. 2000. "Extending grammar annotation standards to spontaneous speech". In Kirk, John M. (ed.). *Corpora galore: analyses and techniques in describing English: papers from the Nineteenth International Conference on English Language Research on Computerised Corpora*. Amsterdam: Rodopi, 295-311.

Ramsay, Alan M. 2010. "From computational linguistics to language engineering". In Malmkjær, Kirsten (ed.). *The Routledge linguistics encyclopedia* (3rd edition). Abingdon: Routledge, 92-94.

Ranta, Elina. 2009. "Syntactic features in spoken ELF - learner language or spoken grammar?". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 84-106.

Rastelli, Stefano. 2009. "Learner corpora without error tagging". *Linguistik online* 38(2), 57-66.

Ratnaparkhi, Adwait. 1996. "A maximum entropy part-of-speech tagger". In Brill, Eric; Church, Kenneth W. (eds.). *Proceedings of the First Conference on Empirical Methods in NLP*. Somerset, NJ: ACL, 133-142.

Rehbein, Ines; Schalowski, Sören. 2012. "Extending the STTS for the annotation of spoken language". In Jeremy Jancsary (ed.). *Proceedings of the Conference on Natural Language Processing (KONVENS) 2012*. Vienna: ÖGAI, 238-242.

Rehbein, Ines; Schalowski, Sören. 2013. "STTS goes Kiez–Experiments on annotating and tagging urban youth language". *JLCL* 28(1), 199-227.

Rehbein, Ines; Schalowski, Sören; Wiese, Heike. 2012a. "Annotating spoken language", In Haugh, Michael; Ruhi, Şükriye; Schmidt, Thomas; Wörner, Kai (eds.). *Best practices for speech corpora in linguistic research: workshop Programme of the LREC-2012*. http://www.corpora.uni-hamburg.de/lrec2012/Proceedings_Complete.pdf (17 January 2013).

Rehbein, Ines; Schalowski, Sören; Wiese, Heike. 2012b. "Erweiterung des STTS für gesprochene Sprache [Extension of STTS for spoken language]". Paper presented at the STTS Workshop at IMS Stuttgart. http://www.ims.uni-stuttgart.de/veranstaltungen/stts-workshop/pdfs/Rehbein_stts_spoken_potsdam2.pdf (4 February 2013).

Reiter, Riccarda. 2013. *The use of* so *as a discourse marker in ELF conversations: a VOICE-based study*. MA Thesis, University of Vienna.

Robins, Robert Henry. 1967. "Yurok". *Lingua* 17, 210-229.

Rögnvaldsson, Eiríkur. 2005. "The Corpus of Spoken Icelandic and its morphosyntactic annotation". In Henrichsen, Peter J.; Skadhauge, Peter R. (eds.). *Treebanking for discourse and speech: proceedings of the NODALIDA. Special session on treebanks for spoken language and discourse*. Copenhagen: Samfundslitteratur

Rooy, Bertus van; Schäfer, Lande. 2002. "The effect of learner errors on POS tag errors during automatic POS tagging". *Southern African Linguistics and Applied Language Studies* 20(4), 325-335.

Rooy, Bertus van; Schäfer, Lande. 2003. "An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus". In Archer, Dawn; Rayson, Paul; Wilson, Andrew; McEnery, Tony (eds.). *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster: Lancaster University, 835-844.

Rosch, Eleanor. 1973. "Natural categories". *Cognitive Psychology* 4, 328-350.

Ross, John Robert. 1972. "The category squish: Endstation Hauptwort". In Peranteau, Paul M.; Levi, Judith N.; Phares, Gloria C. (eds.). *Papers from the eighth regional meeting Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, 316-328.

Ross, John Robert. 1973. "A fake NP squish". In Bailey, Charles-James N.; Shuy, Roger W. (eds.). *New ways of analyzing variation in English*. Washington: Georgetown University Press, 96-140.

Ross, John Robert. 2004. "Nouniness". In Aarts, Bas; Denison, David; Keizer, Evelien; Popova, Gergana (eds.). *Fuzzy grammar: a reader*. Oxford: OUP, 351-422.

Roth, Dan. 1998. "Learning to resolve natural language ambiguities: a unified approach". In AAAI (ed.). *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, Calif: AAAI Press, 806-813.

Roth, Dan; Zelenko, Dmitry. 1998. "Part of speech tagging using a network of linear separators". In ACL (ed.). *Proceedings of the 36th annual meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Stroudsburg, PA: ACL, 1136-1142.

Russell, Bertrand F. R. S. 1923. "Vagueness". *Australasian Journal of Psychology and Philosophy* 1(2), 84-92.

Sagae, Kenji; Davis, Eric; Lavie, Alon; Macwhinney, Brian; Wintner, Shuly. 2007. "High-accuracy annotation and parsing of CHILDES transcripts". In Buttery, Paula;

Villavicencio, Aline; Korhonen, Anna (eds.). *Proceedings of the workshop on cognitive aspects of computational language acquisition*. Stroudsburg, PA: ACL, 25-32.

Sagae, Kenji; Davis, Eric; Lavie, Alon; MacWhinney, Brian; Wintner, Shuly. 2010. "Morphosyntactic annotation of CHILDES transcripts". *Journal of Child Language* 37(3), 705-29.

Sampson, Geoffrey. 2000. "CHRISTINE Corpus: documentation". http://www.grsampson.net/ChrisDoc.html (10 July 2014).

Santner-Wolfartsberger, Anita. 2012. "Parties, persons and one-at-a-time: some fundamental concepts of conversation analysis revisited". *Vienna English Working PaperS* 21, 1-24.

Santorini, Beatrice. 1991. "Part of speech tagging guidelines for the Penn Treebank Project". http://www.personal.psu.edu/xxl13/teaching/sp07/apling597e/resources/Tagset.pdf (28 January 2015).

Saraceni, Mario. 2008. "English as a lingua franca: between form and function". *English Today* 94(2), 20-26.

Sasse, Hans-Jürgen. 2001. "Scales between nouniness and verbiness". In Haspelmath, M. (ed.). *Language typology and language universals: an international handbook*. New York: de Gruyter, 495-509.

ScanDiaSyn. "Nordic Dialect Corpus and Syntax Database". http://www.tekstlab.uio.no/nota/scandiasyn/index.html (11 July 2014).

Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Schiller, Anne; Teufel, Simone; Stöckert, Christine. 1999. "Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset) [Guidelines for the tagging of German text corpora with STTS (small and large tagset)]". http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (19 August 2014).

Schmid, Helmut. 1994. "Probabilistic part-of-speech tagging using decision trees". Revised version of the paper presented at the International Conference on New Methods in Language Processing, Manchester, UK. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf (28 January 2015).

Schmid, Helmut. 2008. "Tokenization and part-of-speech tagging". In Lüdeling, Anke; Kytö, Meria (eds.). *Corpus linguistics: an international handbook. Volume 1*. Berlin: de Gruyter, 527-551.

Seidlhofer, Barbara. 2001. "Closing a conceptual gap: the case for a description of English as lingua franca". *International Journal of Applied Linguistics* 11(2), 133-158.

Seidlhofer, Barbara. 2002. "The shape of things to come? Some basic questions about English as lingua franca". In Knapp, Karlfried; Meierkord, Christiane (eds.). *Lingua franca communication*. Frankfurt/Main: Lang, 269-302.

Seidlhofer, Barbara. 2007. "English as a lingua franca and communities of practice". In Volk-Birke, Sabine; Lippert, Julia (eds.). *Anglistentag 2006 Halle Proceedings*. Trier: Wissenschaftlicher Verlag Trier, 307-318.

Seidlhofer, Barbara. 2009a. "Common ground and different realities: world Englishes and English as a lingua franca". *World Englishes* 28(2), 236-245.

Seidlhofer, Barbara. 2009b. "Orientations in ELF research: form and function". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 37-59.

Seidlhofer, Barbara. 2011a. "Conceptualizing 'English' for a multilingual Europe". In Houwer, Annick de; Wilton, Antje (eds.). *English in Europe today: sociocultural and educational perspectives*. Amsterdam: Benjamins, 133-146.

Seidlhofer, Barbara. 2011b. *Understanding English as a Lingua Franca*, Oxford: OUP.

Seidlhofer, Barbara. 2012. "Corpora and English as a lingua franca". In Hyland, Ken; Chau, Meng Huat; Handford, Michael (eds.). *Corpus applications in applied linguistics*. London: Continuum, 135-149.

Seidlhofer, Barbara; Breiteneder, Angelika; Pitzl, Marie-Luise. 2006. "English as a lingua franca in Europe: challenges for applied linguistics". *Annual Review of Applied Linguistics* 26, 3-34.

Seidlhofer, Barbara; Widdowson, Henry. 2009. "Conformity and creativity in ELF and learner English". In Albl-Mikasa, Michaela; Braun, Sabine; Kalina, Sylvia (eds.). *Dimensionen der Zweitsprachenforschung* [Dimensions of Second Language Research]. *(Festschrift for Kurt Kohn)*. Tübingen: Narr, 93-107.

Siegel, Jeff. 2004. "Morphological elaboration". *Journal of Pidgin and Creole Languages* 19(2), 333-362.

Siegel, Jeff. 2008. *The emergence of pidgin and creole languages*, Oxford: OUP.

Simpson, Rita C.; Lee, David Y.W; Leicher, Sheryl. 2007. "MICASE Manual: The Michigan Corpus of Academic Spoken English". http://micase.elicorpora.info/files/0000/0015/MICASE_MANUAL.pdf (5 February 2013).

Sinclair, John. 2004. "Corpus and text — basic principles". In Wynne, Martin (ed.). *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books, 1-16.

Smit, Ute. 2009. "Emic Evaluations and interactive processes in a classroom community of practice". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 200-224.

Smit, Ute. 2010a. "CLIL in an English as a lingua franca (ELF) classroom: on explaining terms and expressions interactively". In Dalton-Puffer, Christiane, Nikula, Tarja, Smit, Ute (ed.). *Language use and language learning in CLIL classrooms*. Amsterdam: Benjamins, 259-277.

Smit, Ute. 2010b. *English as a lingua franca in higher education: a longitudinal study of classroom discourse*, Berlin: de Gruyter.

Son, Rob J. J. H. van; Pols, Louis C. W. 2001. "Structure and access of the open source IFA-corpus". In Berkman, Ellen H. (ed.). *IFA Proceedings 24*. Amsterdam: University of Amsterdam, 15-25.

Štekauer, Pavol. 2002. "On the theory of neologisms and nonce-formations". *Australian Journal of Linguistics* 22(1), 97-112.

Stenström, Ana-Brita; Andersen, Gisle; Hasund, Ingrid Kristine. 2002. *Trends in teenage talk: corpus compilation, analysis and findings*. Amsterdam: Benjamins.

Stenström, Anna-Brita; Andersen, Gisle; Hasund, Kristine; Monstad, Kristine; Aas, Hanne. 1998. "Users' manual to accompany the Bergen Corpus of London Teenage Language (COLT)". http://clu.uni.no/icame/colt/cd/COLT.PDF (28 January 2015).

Stenström, Anna-Brita; Svartvik, Jan. 1994. "Imparsable speech: repeats and other non-disfluencies in spoken English". In Oostdijk, Nelleke; Haan, Pieter de (eds.). *Corpus-based research into language: in honour of Jan Aarts*. Amsterdam: Rodopi, 241-254.

Strangert, Eva; Ejerhed, Eva; Huber, Dieter. 1993. "Clause structure and prosodic segmentation". In *FONETIK-93 Papers from the 7th Swedish Phonetics Conference.* Uppsala, May 12-14 1993, 81-84.

TalkBank. "Child Language Data Exchange System". http://childes.talkbank.org/ (10 July 2014).

Taylor, Lita J.; Knowles, Gerry. 1988. "Manual of information to accompany the SEC Corpus", University of Lancaster. http://clu.uni.no/icame/manuals/SEC/INDEX.HTM (28 January 2015).

TELF Project. "Tübingen English as a Lingua Franca Corpus. Project website". http://projects.ael.uni-tuebingen.de/telf/ (16 July 2014).

Titania. "Bank of English user guide". University of Birmingham. http://www.titania.bham.ac.uk/docs/svenguide.html (8 July 2014).

Trosterud, Trond. 2009. "A constraint grammar for Faroese". In Bick, Eckhard; Hagen, Kristin; Müürisep, Kaili; Trosterud, Trond (eds.). *NODALIDA 2009 workshop Constraint Grammar and robust parsing.* Tartu: NEALT, 1-7.

Trudgill, Peter. 2011 [revised version from 1999]. "Standard English: what it isn't". In Bex, Tony; Watts, Richard J. (eds.). *Standard English: the widening debate*. London: Routledge, 117-128.

Trudgill, Peter. 2014. "Before ELF: GLF from Samarkand to Sfakia". *Journal of English as a Lingua Franca* 3(2), 387-393.

UCREL. 1993a-2014. "CLAWS part-of-speech tagger for English". Lancaster: Lancaster University. http://ucrel.lancs.ac.uk/claws (17 August 2014).

UCREL. 1993b-2014. "UCREL CLAWS1 (LOB) tagset", Lancaster: Lancaster University. http://ucrel.lancs.ac.uk/claws1tags.html (17 August 2014).

UCREL. 1993c-2014. "UCREL CLAWS2 tagset", Lancaster: Lancaster University. http://ucrel.lancs.ac.uk/claws2tags.html (18 August 2014).

UCREL. 1995-2004. "Claws tagset C8", Lancaster: Lancaster University. http://ucrel.lancs.ac.uk/claws8tags.pdf (18 August 2014).

Uí Dhonnchadha, Elaine; Frenda, Alessio; Vaughan, Brian. 2012. "Issues in designing a corpus of spoken Irish". In Pauw, Guy de; Schryver, Gilles-Maurice de; Forcada, Mikel L.; Sarasola, Kepa; Tyers, Francis; M.; Wagacha, PeterWaiganjo (eds.). *Proceedings of the workshop on language technology for normalisation of less-resourced languages*. Istanbul: ELRA, 1-5.

Vertovec, Steven. 2007. "Super-diversity and its implications". *Ethnic and Racial Studies* 30(6), 1024-1054.

Vertovec, Steven. 2010. "Towards post-multiculturalism? Changing communities, conditions and contexts of diversity". *International Social Science Journal* 61(199), 83-95.

VOICE. 2011. "The Vienna-Oxford International Corpus of English (version 1.0 XML)". *Director: Barbara Seidlhofer; Researchers: Angelika Breiteneder, Theresa Klimpfinger, Stefan Majewski, Marie-Luise Pitzl.*

VOICE. 2013. "The Vienna-Oxford International Corpus of English (version POS Online 2.0)". http://voice.univie.ac.at/pos/ (28 January 2015).

VOICE Project. 2007a. "Spelling conventions". http://www.univie.ac.at/voice/documents/VOICE_spelling_conventions_v2-1.pdf (28 January 2015).

VOICE Project. 2007b. "VOICE transcription conventions [2.1]". http://www.univie.ac.at/voice/voice.php?page=transcription_general_information (28 January 2015).

VOICE Project. 2009a. "Output styles". Using VOICE Online. http://www.univie.ac.at/voice/help/output_formats (28 January 2015).

VOICE Project. 2013a. "VOICE POS Online.". http://univie.ac.at/voice/help/pos (28 January 2015).

VOICE Project. 2014a. "VOICE Part-of-Speech tagging and Lemmatization Manual [1st revised version]". http://www.univie.ac.at/voice/documents/VOICE_tagging_manual.pdf (28 January 2015).

VOICE Project. 2007c. "Mark-up conventions. VOICE Transcription Conventions [2.1]". http://www.univie.ac.at/voice/documents/VOICE_mark-up_conventions_v2-1.pdf (28 January 2015).

VOICE Project. 2007d. "VOICE C1 manual". (2nd version). Unpublished PDF file.

VOICE Project. 2008. "VOICE transcription set [2.1]". *Unpublished internal transcription conventions*.

VOICE Project. 2009b. "Using VOICE Online". http://www.univie.ac.at/voice/help (28 January 2015).

VOICE Project. 2013b. "Annotation differences between VOICE Online and VOICE POS Online". Using VOICE Online. http://univie.ac.at/voice/help/pos (28 January 2015).

VOICE Project. 2013c. "Corpus description". http://www.univie.ac.at/voice/page/corpus_description (28 January 2015).

VOICE Project. 2013d. "Corpus information". http://www.univie.ac.at/voice/page/corpus_information (28 January 2015).

VOICE Project. 2014b. "VOICE Project website". http://www.univie.ac.at/voice/page/corpus_information (28 January 2015).

Voutilainen, Atro. 1995. "Morphological disambiguation". In Karlsson, Fred; Voutilainen, Atro; Heikkilä, Juha; Anttila, Atro (eds.). *Constraint grammar: a language-independent system for parsing unrestricted text*. Berlin: de Gruyter, 165-284.

Voutilainen, Atro. 1999a. "A short history of tagging". In Halteren, Hans van (ed.). *Syntactic wordclass tagging*. Dordrecht: Kluwer, 9-21.

Voutilainen, Atro. 1999b. "Orientation". In Halteren, Hans van (ed.). *Syntactic wordclass tagging*. Dordrecht: Kluwer, 3-7.

Voutilainen, Atro. 2003. "Part-of-speech tagging". In Mitkov, Ruslan (ed.). *The Oxford handbook of computational linguistics*. Oxford: OUP, 219-231.

Vukovics, Iris. 2013. *Of 's and ofs: German and Finnish speakers' choice of possessive construction in English*. MA thesis, University of Vienna.

Wacker, Elisabeth. 2011. *English as a Lingua Franca and the third person –s*. BA thesis, University of Vienna.

Walker, James A. 2010. *Variation in linguistic systems*. New York, NY: Routledge.

Watterson, Matthew. 2008. "Repair of non-understanding in English in international communication". *World Englishes* 27(3-4), 378-406.

Widdowson, Henry G. 1997. "EIL, ESL, EFL: global issues and local interests". *World Englishes* 16(1), 135-146.

Widdowson, Henry G. 2003. *Defining issues in English language teaching*. Oxford: OUP.

Widdowson, Henry G. 2012. "ELF and the inconvenience of established concepts". *Journal of English as a Lingua Franca* 1(1), 5-26.

Widdowson, Henry G. 2014. Personal conversation, 10 Nov.

Wierzbicka, Anna. 2004. "Prototypes save". In Aarts, Bas; Denison, David; Keizer, Evelien; Popova, Gergana (eds.). *Fuzzy grammar: a reader*. Oxford: OUP, 461-478.

Wiese, Heike. 2013. "From feature pool to pond: the ecology of new urban vernaculars". *Working Papers in Urban Language & Literacies* (Paper 104), 1-29.

Wu, Dekai; Ngai, Grace; Carpuat, Marine. 2003. "A stacked, voted, stacked model for named entity recognition". In Daelemans, Walter; Osborne, Miles (eds.). *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. Stroudsburg, PA: ACL, 200-203.

Xiao, Richard Z. 2008. "Well-known and influential corpora". In Lüdeling, Anke; Kytö, Meria (eds.). *Corpus Linguistics: an international handbook. Volume 1*. Berlin: de Gruyter, 383-457.

APPENDICES

Appendix A: Short list of tags in VOICE

I) The VOICE Tagset, sorted alphabetically according to word class categories	
Adjective, comparative	JJR
Adjective, positive	JJ
Adjective, superlative	JJS
Adverb, comparative	RBR
Adverb, positive	RB
Adverb, superlative	RBS
Breathing	BR
Cardinal Number	CD
Coordinating conjunction	CC
Determiner	DT
Discourse Marker	DM
Existential there	EX
Foreign word (Non-English speech)	FW
Formulaic Item	FI
Generic Noun Tag	N
Generic Verb Tag	V
Interjection	UH
Laughter	LA
List Item Marker	LS
Modal Verb	MD
Noun, plural	NNS
Noun, singular or mass	NN
Onomatopoeic noise	ONO
Partial word	XX
Particle	RP
Pause	PA
Possessive Ending	POS
Predeterminer	PDT
Preposition	IN
Pronoun, personal	PP
Pronoun, possessive	PP$
Pronoun, relative	PRE
Pronunciation Variations and Coinages	PVC
Proper Noun, plural	NPS
Proper Noun, singular	NP
Response particle	RE
Spelling out	SP
Subordinating conjunction	IN
Symbol	SYM
to, infinitive use	TO
Unintelligible speech	UNI
Unknown	UNK
Verb be, base form	VB
Verb be, contracted form	VBS

Verb be, gerund or present participle	VBG
Verb be, past participle	VBN
Verb be, past tense	VBD
Verb be, present 3rd person singular	VBZ
Verb be, present non-3rd person singular	VBP
Verb do, contracted form (meaning 'does')	DOS
Verb have, base form	VH
Verb have, contracted form	VHS
Verb have, gerund or present participle	VHG
verb have, past participle	VHN
Verb have, past tense	VHD
Verb have, present 3rd person singular	VHZ
Verb have, present non-3rd person singular	VHP
Verb other than be and have, base form	VV
Verb other than be and have, gerund or present participle	VVG
Verb other than be and have, past participle	VVN
Verb other than be and have, past tense	VVD
Verb other than be and have, present 3rd person singular	VVZ
Verb other than be and have, present non-3rd person singular	VVP
Wh-adverb	WRB
Wh-determiner	WDT
Wh-pronoun	WP

II) The VOICE Tagset, sorted alphabetically according to tags

BR	Breathing
CC	Coordinating conjunction
CD	Cardinal Number
DM	Discourse Marker
DOS	Verb do, contracted form (meaning 'does')
DT	Determiner
EX	Existential there
FI	Formulaic Item
FW	Foreign word (Non-English speech)
IN	Preposition or subordinating conjunction
JJ	Adjective, positive
JJR	Adjective, comparative
JJS	Adjective, superlative
LA	Laughter
LS	List Item Marker
MD	Modal Verb
N	Generic Noun Tag
NN	Noun, singular or mass
NNS	Noun, plural

NP	Proper Noun, singular
NPS	Proper Noun, plural
ONO	Onomatopoeic noise
PA	Pause
PDT	Predeterminer
POS	Possessive Ending
PP	Pronoun, personal
PP$	Pronoun, possessive
PRE	Pronoun, relative
PVC	Pronunciation Variations and Coinages
RB	Adverb, positive
RBR	Adverb, comparative
RBS	Adverb, superlative
RE	Response particle
RP	Particle
SP	Spelling out
SYM	Symbol
TO	to, infinitive use
UH	Interjection
UNI	Unintelligible speech
UNK	Unknown
V	Generic Verb Tag
VB	Verb be, base form
VBD	Verb be, past tense
VBG	Verb be, gerund or present participle
VBN	Verb be, past participle
VBP	Verb be, present non-3rd person singular
VBS	Verb be, contracted form
VBZ	Verb be, present 3rd person singular
VH	Verb have, base form
VHD	Verb have, past tense
VHG	Verb have, gerund or present participle
VHN	verb have, past participle
VHP	Verb have, present non-3rd person singular
VHS	Verb have, contracted form
VHZ	Verb have, present 3rd person singular
VV	Verb other than be and have, base form
VVD	Verb other than be and have, past tense
VVG	Verb other than be and have, gerund or present participle
VVN	Verb other than be and have, past participle
VVP	Verb other than be and have, present non-3rd person singular
VVZ	Verb other than be and have, present 3rd person singular
WDT	Wh-determiner
WP	Wh-pronoun
WRB	Wh-adverb
XX	Partial word

Appendix B: List of all ambiguous tags in VOICE

CC(CC)/DT(DT)
CD(CD)/NN(NN)
DM(DM)/IN(IN)
DM(DM)/JJ(JJ)
DM(DM)/NN(NN)
DM(DM)/RB(RB)
DM(DM)/VV(VV)
DM(DM)/VVP(VVP)
DM(DM)/WDT(WDT)
DM(DM)/WP(WP)
DT(DT)/CC(CC)
DT(DT)/DT(NN)
DT(DT)/IN(IN)
DT(DT)/JJ(JJ)
DT(DT)/NN(NN)
DT(DT)/PDT(PDT)
DT(DT)/PP(PP)
DT(DT)/PRE(PRE)
DT(DT)/RB(RB)
DT(DT)/RE(RE)
DT(DT)/RE(UNK)
EX(EX)/RB(RB)
FI(FI)/JJ(JJ)
FI(FI)/NN(NN)
FI(FI)/PP(PP)
FI(FI)/VV(VV)
FI(FI)/VVP(VVP)
IN(IN)/PRE(PRE)
IN(IN)/RB(RB)
IN(IN)/RP(RP)
IN(IN)/TO(TO)
IN(IN)/VVG(VVG)
IN(IN)/VVN(VVN)
IN(IN)/VVP(VVP)
JJ(JJ)/JJ(CD)
JJ(JJ)/JJ(NN)
JJ(JJ)/JJ(NPS)
JJ(JJ)/JJ(RB)
JJ(JJ)/JJ(UNK)
JJ(JJ)/JJ(VVN)
JJ(JJ)/NN(JJ)
JJ(JJ)/NN(N)
JJ(JJ)/NN(NN)
JJ(JJ)/NN(NNS)
JJ(JJ)/NNS(NNS)
JJ(JJ)/NP(N)
JJ(JJ)/NP(NP)
JJ(JJ)/NP(NPS)
JJ(JJ)/NP(UNK)
JJ(JJ)/NPS(NPS)
JJ(JJ)/RB(RB)
JJ(JJ)/RE(RE)
JJ(JJ)/V(NN)
JJ(JJ)/V(V)
JJ(JJ)/VV(VV)
JJ(JJ)/VVD(VVD)
JJ(JJ)/VVD(VVP)
JJ(JJ)/VVG(NN)
JJ(JJ)/VVG(UNK)
JJ(JJ)/VVG(VVG)
JJ(JJ)/VVN(VVN)
JJ(N)/NP(N)
JJ(NNS)/NN(NNS)
JJ(RB)/RB(RB)
JJ(VV)/RB(VV)
JJR(JJR)/RBR(RBR)
JJS(JJS)/NN(NN)
JJS(JJS)/NNS(NNS)
JJS(JJS)/RBS(RBS)
MD(MD)/NN(NN)
MD(MD)/VHD(VHD)
NN(JJ)/V(JJ)
NN(NN)/CD(CD)
NN(NN)/DM(DM)
NN(NN)/JJ(JJ)
NN(NN)/NN(JJ)
NN(NN)/NN(NNS)
NN(NN)/NN(RB)
NN(NN)/NN(UNK)
NN(NN)/NN(V)
NN(NN)/NN(VV)
NN(NN)/NN(VVG)
NN(NN)/NN(VVP)
NN(NN)/NN(VVZ)

NN(NN)/NNS(NNS)
NN(NN)/NP(NP)
NN(NN)/RB(RB)
NN(NN)/RE(RE)
NN(NN)/SP(SP)
NN(NN)/UNK(VV)
NN(NN)/UNK(VVN)
NN(NN)/V(JJ)
NN(NN)/V(UNK)
NN(NN)/V(V)
NN(NN)/V(VVD)
NN(NN)/V(VVG)
NN(NN)/V(VVN)
NN(NN)/V(VVZ)
NN(NN)/VV(VV)
NN(NN)/VVD(VVD)
NN(NN)/VVN(VVN)
NN(NN)/VVP(VVP)
NN(NN)/WP(WP)
NN(NNS)/V(VVG)
NN(NNS)/VV(VV)
NN(NNS)/VVP(VVP)
NNS(NNS)/JJ(JJ)
NNS(NNS)/NNS(CD)
NNS(NNS)/NNS(JJ)
NNS(NNS)/NNS(NN)
NNS(NNS)/NPS(NPS)
NNS(NNS)/VVZ(VVP)
NNS(NNS)/VVZ(VVZ)
NNS(UNK)/VVZ(VVZ)
NP(NP)/JJ(JJ)
NP(NP)/NN(NN)
NP(NP)/NP(JJ)
NP(NP)/NP(NPS)
NP(NP)/NPS(NP)
NP(NP)/NPS(NPS)
NP(NP)/VVD(VVD)
NPS(NPS)/JJ(JJ)
NPS(NPS)/N(N)
NPS(NPS)/NPS(JJ)
PDT(PDT)/DT(DT)
PDT(PDT)/RB(RB)
POS(POS)/VBS(VBS)
PP$(PP$)/PP(PP)
PP(DM)/PP(PP)
PP(PP)/NN(NN)
PP(PP)/PP$(PP$)
PP(PP)/PP(DM)
PP(PP)/PP(DT)
PP(PP)/RB(RB)
PRE(PRE)/RB(RB)
PRE(PRE)/WDT(WDT)
PRE(PRE)/WP(WP)
PVC(JJ)/PVC(NN)
PVC(JJ)/PVC(NNS)
PVC(JJ)/PVC(NP)
PVC(JJ)/PVC(RB)
PVC(JJ)/PVC(VVG)
PVC(JJ)/PVC(VVN)
PVC(JJR)/PVC(NNS)
PVC(NN)/PVC(NNS)
PVC(NN)/PVC(NP)
PVC(NN)/PVC(RB)
PVC(NN)/PVC(VVG)
PVC(NNS)/PVC(RB)
PVC(NNS)/PVC(VVZ)
PVC(NPS)/PVC(JJ)
PVC(VV)/PVC(VVD)
PVC(VV)/PVC(VVP)
PVC(VVD)/PVC(VVN)
PVC(VVD)/PVC(VVP)
PVC(VVD)/PVC(VVZ)
RB(RB)/CC(CC)
RB(RB)/DM(DM)
RB(RB)/EX(EX)
RB(RB)/IN(IN)
RB(RB)/JJ(JJ)
RB(RB)/JJR(JJR)
RB(RB)/NN(NN)
RB(RB)/RB(JJ)
RB(RB)/RB(UNK)
RB(RB)/RE(RE)
RB(RB)/RP(RP)
RB(RB)/UNK(VV)
RBR(RBR)/JJR(JJR)
RE(RE)/DT(DT)
RE(RE)/JJ(JJ)
RE(RE)/NN(NN)
RE(RE)/RB(RB)
RE(RE)/RE(RB)
RP(RP)/IN(IN)
RP(RP)/RB(RB)
UNK(JJ)/V(V)

V(V)/JJ(JJ)
V(V)/N(N)
V(V)/NN(NN)
V(V)/NN(NNS)
V(V)/V(NN)
V(V)/V(VVD)
V(V)/V(VVG)
V(V)/V(VVN)
V(V)/V(VVZ)
VBP(VBP)/VBP(VBZ)
VBS(VBS)/VBS(VBP)
VBS(VBS)/VHS(VHS)
VBZ(VBZ)/VBZ(VBP)
VH(VH)/VHP(VHP)
VHD(VHD)/VHD(VH)
VHD(VHD)/VHN(VHN)
VHP(VHP)/V(V)
VHP(VHP)/VH(VH)
VHP(VHP)/VHP(VHZ)
VHZ(VHZ)/VHZ(VHP)
VV(VV)/JJ(JJ)
VV(VV)/JJ(RB)
VV(VV)/NN(NN)
VV(VV)/V(JJ)
VV(VV)/V(NN)
VV(VV)/V(V)
VV(VV)/V(VVD)
VV(VV)/V(VVG)
VV(VV)/V(VVN)
VV(VV)/V(VVZ)
VV(VV)/VVD(VVD)
VV(VV)/VVN(VVN)
VV(VV)/VVP(VVP)
VVD(VVD)/V(VV)
VVD(VVD)/V(VVP)
VVD(VVD)/V(VVZ)
VVD(VVD)/VVD(VV)
VVD(VVD)/VVD(VVN)
VVD(VVD)/VVD(VVP)
VVD(VVD)/VVN(VVN)
VVD(VVD)/VVP(VVP)
VVD(VVD)/VVZ(VVZ)
VVG(JJ)/VVG(VVG)
VVG(NN)/VVG(NNS)
VVG(NN)/VVG(V)
VVG(NN)/VVG(VV)
VVG(NN)/VVG(VVG)
VVG(NN)/VVG(VVN)
VVG(NNS)/VVG(VVG)
VVG(VVG)/V(V)
VVG(VVG)/VVG(V)
VVG(VVG)/VVG(VV)
VVG(VVG)/VVG(VVN)
VVG(VVG)/VVG(VVP)
VVG(VVG)/VVG(VVZ)
VVN(VVN)/V(JJ)
VVN(VVN)/V(V)
VVN(VVN)/V(VV)
VVP(DM)/V(VVG)
VVP(VVP)/V(N)
VVP(VVP)/V(V)
VVP(VVP)/V(VVD)
VVP(VVP)/V(VVG)
VVP(VVP)/V(VVN)
VVP(VVP)/V(VVZ)
VVP(VVP)/VV(VV)
VVP(VVP)/VVG(VVG)
VVP(VVP)/VVP(DM)
VVZ(VVZ)/VVZ(NN)
VVZ(VVZ)/VVZ(VVG)
VVZ(VVZ)/VVZ(VVP)
WDT(WDT)/WP(WP)
WP(WP)/WRB(WRB)

Appendix C: List of all tokens with differing form and function tags in VOICE (only non-ambiguous tags combinations)

CD(DT)
CD(JJ)
CD(NN)
CD(NNS)
DT(NN)
DT(RB)
FI(JJ)
IN(NN)
JJ(CD)
JJ(JJR)
JJ(N)
JJ(NN)
JJ(NP)
JJ(RB)
JJ(UNK)
JJ(VV)
JJ(VVN)
JJ(VVP)
JJR(JJ)
JJS(RBS)
NN(JJ)
NN(NNS)
NN(RB)
NN(UNK)
NN(VV)
NN(VVP)
NNS(NN)
NNS(UNK)
NP(JJ)
NP(NPS)
NPS(NP)
PP(DM)
PP(DT)
PP(PP$)
PP(UNK)
PVC(CD)
PVC(FI)
PVC(JJ)
PVC(JJR)
PVC(JJS)
PVC(NN)
PVC(NNS)
PVC(NP)
PVC(NPS)
PVC(POS)
PVC(PP)
PVC(RB)
PVC(RBR)
PVC(UNK)
PVC(VBG)
PVC(VV)
PVC(VVD)
PVC(VVG)
PVC(VVN)
PVC(VVP)
PVC(VVZ)
RB(DT)
RB(JJ)
RB(NN)
RBR(RBS)
UNK(JJ)
UNK(JJR)
UNK(MD)
UNK(N)
UNK(NN)
UNK(NNS)
UNK(PRE)
UNK(RB)
UNK(V)
UNK(VV)
V(JJ)
V(N)
V(NN)
V(NP)
V(UNK)
V(VH)
V(VV)
V(VVD)
V(VVG)
V(VVN)
V(VVP)
V(VVZ)
VB(VBP)
VBD(VBP)
VBG(NN)
VBG(VB)
VBN(VB)

VBP(VBZ)
VBP(VHP)
VBS(VBP)
VBZ(VBP)
VHD(VH)
VHG(VH)
VHP(V)
VHP(VBP)
VHP(VHD)
VHP(VHZ)
VHZ(V)
VHZ(VBD)
VHZ(VH)
VHZ(VHP)
VVD(V)
VVD(VV)
VVD(VVG)
VVD(VVN)
VVD(VVZ)
VVG(JJ)
VVG(NN)
VVG(NNS)
VVG(UNK)
VVG(V)
VVG(VV)
VVG(VVD)
VVG(VVN)
VVG(VVP)
VVG(VVZ)
VVN(JJ)
VVN(V)
VVN(VV)
VVN(VVD)
VVN(VVG)
VVN(VVP)
VVP(DM)
VVP(VVD)
VVP(VVN)
VVP(VVZ)
VVZ(JJ)
VVZ(N)
VVZ(UNK)
VVZ(V)
VVZ(VV)
VVZ(VVN)
VVZ(VVP)

Appendix D: List of first language abbreviations occurring in this thesis according to the alpha-3/ISO 639-2 Code

Source: http://www.loc.gov/standards/iso639-2/php/code_list.php

Alpha-3/ISO 639-2 Code	Language, Country of origin
alb	Albanian
ara-YE	Arabic, Yemen
bos-BA	Bosnian, Bosnia and Herzegovina
cze-CZ	Czech, Czech Republic
dan-DK	Danish, Denmark
dut-NL	Dutch, Netherlands
eng-GB	English, Great Britain
eng-IE	English, Ireland
eng-MT	English, Malta
eng-US	English, United States of America
est-EE	Estonian, Estonia
fin-FI	Finnish, Finland
fre-FR	French, France
fre-LU	French, Luxembourg
ger	German
ger-AT	German, Austria
ger-CH	German, Switzerland
ger-DE	German, Germany
gre-CY	Greek, Cyprus
hun-HU	Hungarian, Hungary
ind-ID	Indonesian, Indonesia
ita-IT	Italian, Italy
kir-KG	Kirghiz, Kyrgyzstan
kor	Korean
kor-KR	Korean, the Republic of Korea
lav-LV	Latvian, Latvia

Alpha-3/ISO 639-2 Code	Language, Country of origin
mac-MK	Macedonian, Macedonia
mlt-MT	Maltese, Malta
nor-NO	Norwegian, Norway
pol	Polish
pol-PL	Polish, Poland
rum-RO	Rumanian, Rumania
rus-RU	Russian, Russia
scc-RS	Serbian, Serbia (the ISO code of *scc* was changed to *srp*)
scr-HR	Croatian, Croatia (the ISO code of scr was changed to *hrv*)
slo-SK	Slovak, Slovakia
slv	Slovenian
slv-SI	Slovenian, Slovenia
spa-AR	Spanish, Argentina
spa-ES	Spanish, Spain
swe-SE	Swedish, Sweden

Appendix E: Details on speech events occurring in this thesis

Domains

ED	Educational
LE	Leisure
PB	Professional Business
PO	Professional Organizational
PR	Professional Research and Science

Speech event types

con	conversation
int	interview
mtg	meeting
pan	panel
prc	press conference
qas	question-answer session
sed	seminar discussion
sve	service encounter
wgd	working group discussion
wsd	workshop discussion

Title	Duration	Words	Number of speakers
EDcon250	00:47:59	6379	7
EDcon4	00:47:04	5420	9
EDcon496	00:32:19	5075	5
EDcon521	00:57:26	11637	15
EDint330	00:51:01	8703	5
EDsed301	00:50:28	8292	8
EDsed31	02:15:32	22198	22
EDsed362	00:55:21	7164	20
EDsed364	00:25:19	4173	14
EDsve423	00:21:38	2611	5
EDwgd241	00:53:49	10834	9
EDwgd6	00:31:49	4421	11
EDwsd242	00:28:05	5398	15
EDwsd302	01:24:51	14698	23
EDwsd303	00:59:48	10516	12
Title	**Duration**	**Words**	**Number of speakers**

EDwsd304	01:25:31	13952	20
EDwsd499	01:27:44	14804	16
LEcon229	00:21:47	3285	4
LEcon352	00:09:55	2126	7
LEcon405	00:13:43	1933	3
LEcon545	00:37:20	7413	7
LEcon560	02:21:43	21867	10
LEcon565	00:28:23	3979	2
LEcon566	00:49:00	5634	2
LEcon573	00:09:15	1731	2
LEcon8	00:30:42	3114	6
LEint554	00:03:45	847	6
PBmtg269	02:33:51	22585	7
PBmtg27	01:17:35	15068	7
PBmtg280	00:27:22	4768	6
PBmtg3	03:28:06	24601	6
PBmtg300	03:08:25	35277	9
PBmtg414	01:56:00	21195	6
PBmtg463	01:32:03	15484	8
PBpan25	01:09:34	10873	11
PBpan581	00:51:00	8128	8
POmtg314	02:00:48	14767	12
POmtg316	00:39:30	5910	11
POmtg403	01:16:04	12848	7
POmtg404	01:26:29	14641	11
POmtg439	01:06:15	11922	11
POmtg444	00:29:45	6692	10
POmtg447	00:19:12	4177	10
POmtg541	01:08:52	11345	7
Title	**Duration**	**Words**	**Number of speakers**

POprc559	00:34:57	5421	11
POwgd524	00:53:51	10880	8
POwgd12	01:13:39	13175	12
POwgd14	02:44:00	20130	14
POwgd243	01:12:30	11210	11
POwgd26	00:48:44	8122	9
POwgd317	01:09:49	12251	7
POwgd524	00:53:51	10880	8
POwsd266	01:08:55	11217	13
POwsd372	01:15:15	10846	13
PRcon29	00:21:59	3162	4
PRcon534	00:12:26	2116	2
PRcon550	00:03:55	873	2
PRint30	00:08:00	1521	3
PRint595	00:15:12	3069	4
PRpan1	01:55:58	14834	14
PRpan13	02:56:37	10963	17
PRpan225	00:44:12	6469	16
PRpan294	01:13:26	11500	13
PRqas19	00:16:52	1666	4

Appendix F: VOICE Part-of-Speech Tagging and Lemmatization Manual

Part-of-Speech Tagging and Lemmatization Manual

(1st revised version May 2014)

Contents

1 Introductory remarks[1]

Part-of-speech tagging (POS tagging), i.e. the assignment of word class categories to tokens in a corpus, has become a standard feature in corpus annotation. The obvious advantage of POS tagging for corpus users is that it enhances the searchability of a corpus, since it provides additional information about the (corpus) data which corpus users would otherwise have to laboriously work out for themselves.

While there are, of course, large-scale corpora of L1 data whose spoken components are also part-of-speech tagged (e.g. BNC, COCA), there are to date no fully POS-tagged corpora of spoken L2 data, let alone English as a lingua franca (ELF) data. POS-tagging VOICE was in many aspects different from traditional POS tagging; in the absence of suitable models to refer to, the part-of-speech tagging of VOICE was a challenging and time-consuming process, carried out between 2009 and 2012. In fact, the tagging process itself raised a number of questions, e.g. about the (im)possibility of clear-cut categorization of intrinsically variable language. Given this particular condition, it might be helpful to make a few introductory remarks about the implications for POS tagging within an ELF corpus framework:

Firstly, it is important to stress that POS tagging is, just like any form of annotation, necessarily only an approximate process: the information it provides is always to some degree a function of subjective interpretation. Language use is of its very nature intrinsically variable, and could not function as a means of communication otherwise, so the idea that it can be definitively categorized into distinct parts of speech must always be understood as to some degree a convenient descriptive fiction, albeit a useful and widespread one which linguists and language professionals make use of in the description and teaching of language, and which is recorded/codified in grammar books. However, when criteria for categorization are specified, how far particular instances of actual use meet these criteria is often problematic. There are times when linguistic forms and/or their co-textual connections give sufficient evidence for their grammatical categories to be assigned with some degree of confidence. But there are also many cases when the evidence is inconclusive.

Secondly, most POS tagging has to date been carried out on corpora of native speaker data, predominantly written, and it is this kind of data that tagging procedures have been developed to deal with. Thus, written L1 (English) data can be annotated by direct reference to established grammars and tagging procedures. Where problematic cases occur, decisions to assign one part of speech tag or another to a linguistic form can be informed by familiarity with what 'normally' occurs in native speaker usage. There can be no such appeal to 'normality' in the POS tagging of VOICE. The data are quite different, consisting of spontaneous and, to a large extent, highly interactive speech events capturing the spoken usage of English not as a native language but as a lingua franca, where the usual conventions of seeming L1-normality do not apply. The speakers in VOICE interact with each other by exploiting the resources of English in varied and nonconventional ways. Not surprisingly, the occurrence of many non-canonical forms in the ELF data poses somewhat of a challenge when trying to apply conventionally codified word class categories in the process of POS

[1] In this 1st revised version of the *VOICE Part-of-Speech Tagging and Lemmatization Manual*, dated May 2014, a number of errata contained in the original version were corrected. The authors would like to thank Nora Dorn and Claudio Schekulin for their much appreciated help with the revised version of this document.

tagging. Essentially, relying fully on existing English language tagging practices for VOICE would have constituted an attempt to apply a system of annotation to data it was not designed to account for.

This naturally places a particular premium on interpretation. In POS tagging VOICE, we were thus faced with making decisions about how to tag forms which did not meet the criteria for conventionally codified categorization – forms, for example, that were morphologically marked as nouns but did not syntactically function as such, e.g. the word 'sticker' in *I cannot **sticker** this*. But since they needed to be tagged in one way or the other, operational decisions needed to be taken, e.g. by categorising items which did not easily fit into a particular category, and thus ruling out other, perhaps equally legitimate, options. Such operational decisions are, to some degree, always bound to be arbitrary, but unavoidable in the process of POS tagging, which necessarily involves the static categorization of what is intrinsically variable language use. In the tagging of spoken ELF data as recorded in VOICE, this becomes especially apparent.

And herein lies the particular significance of POS tagging for a corpus like VOICE. Our aim was to modify existing tagging categories and procedures to arrive at a POS-annotated version of VOICE so as to make it more user-friendly. In the process of producing VOICE POS, the problems of applying conventional categories to spoken ELF data made us aware of the essential features of any natural language use that are made particularly evident in the use of English as a lingua franca – features that are perhaps sometimes not perceived when dealing with native speaker usage because they are so familiar.

In short, the process of adding POS tags to VOICE was in many cases far from straightforward: the variation in the data and lack of a fully relevant precedent to follow posed many challenges. These, we dealt with as best we could in the tagging scheme described in this manual. All these challenges, and the way they were met, are further discussed in Osimk-Teasdale (in prep.) and Radeka (in prep.).

Vienna, January 2013, Barbara Seidlhofer, Ruth Osimk-Teasdale, Michael Radeka

2 Technical remarks on the tagging of VOICE

2.1 Tokenization

Tokenization is the process of dividing up a text into separate grammatical segments, e.g. into individual tokens, or sentential elements. This segmenting into tokens is a pre-processing step for other annotations such as part-of-speech tagging and lemmatization. The tokenization in VOICE included an extraction of 'pure' text (without mark-up), as well as pauses and laughter from VOICE XML, and the splitting of the text into individual parts, i.e. into tokens.

Verb contractions and the genitive *'s* are split into their individual components. Each component then receives a separate POS tag, as demonstrated below.

Text	Tokenization into individual parts	Part-of-speech Tagging
it's	it + 's	it_PP 's_VBS, it_PP 's_VHS, let_VV 's_PP, president_NN 's_POS
you're	you + 're	you_PP 're_VBP
we'll	we + 'll	we_PP 'll_MD
gonna	gon + na	gon_VVG na_TO
won't	wo + n't	wo_MD n't_RB
student's	student + 's	student_NN 's_POS
students'	students + '	students_NNS '_POS

2.2 Part-of-speech tagging

This section explains the technical procedures for part-of-speech tagging VOICE. The first operational decision anyone annotating a corpus with parts of speech has to take is the choice of a tagger and a tagging methodology. As VOICE is the first corpus of English as a lingua franca to be annotated with part-of-speech tags, there were no directly comparable models or easily transferrable tagging methodology for our data. In order for the workload to be feasible, we could not, however, start from scratch in designing our own guidelines and taggers either. We therefore had to rely on the adaptation of existing tagging models for the POS tagging of VOICE.

2.2.1 Adaptation of tagset

We chose to work with the tagset from the *Part of Speech Tagging Guidelines for the Penn Treebank Project* (Santorini 1991) because this tagset uses rather coarse-grained categories, which reduce the number of possible ambiguities. This suited the needs for tagging the varied, spoken nature of the ELF data in VOICE. As expected, an analysis of the tagger's output on larger portions of our data

revealed that we shared problems faced by previous approaches to POS tagging (e.g. ICE-GB, BNC) which applied and adapted taggers originally designed to deal with written language to typical features of spoken language. Such challenges are, amongst others, disfluencies, repetitions, re-starts, discourse markers and pauses. Additionally, tagger and tagset could not account for a number of features characteristic of our data, e.g. the input of multilingual speakers, including code-switches, non-canonical forms, and non-canonical form-function relationships. In order to account for these features we extended both the tagset of the Penn Treebank (cf. 3.3 VOICE Tagset, tags marked in green) and the tagging formats (cf. 3.2. Tagging formats in VOICE).

2.2.2 Creating the VOICE Tagging Lexicon

A POS tagger's lexicon lists word forms together with their possible tags according to a predefined tagset. Most of the available POS taggers are trained solely on the Wall Street Journal (Marcus, Marcinkiewicz & Santorini 1993). For the lexicon used for tagging VOICE, we drew on a number of different sources in order to build a lexicon which was as comprehensive as possible. We did not use the lexica which come with different taggers but chose to build our own in order to best cover the spoken, interactive and linguaculturally diverse features of VOICE data. In the VOICE lexicon, we included written, as well as spoken, Penn-Treebank corpora (Wall Street Journal, Brown, Switchboard, ATIS-3), large lexical databases like Wordnet (1995, Fellbaum 1998) and CELEX (Baayen, Piepenbrock & Gulikers 1995) and a number of word lists from our reference dictionary, which was the Oxford Advanced Learner's Dictionary, Edition 7 (OALD7, Hornby et al. 2007). Additionally, we consulted and added other resources to the VOICE tagging lexicon, e.g. those provided by the Natural Language Toolkit (NLTK) and GATE, as well as tokens from the frequency lists of the BNC (Kilgarriff 2006) and COCA (Davies n.d.). Finally, all of these resources were made compatible with the VOICE tagset, i.e. either by assigning tags where there were none, or adjusting the tag categories to those in the VOICE tagset. In a last step, the lexicon was completed with manually entered information: a few thousand words in the VOICE data remaining without an entry in the lexicon were added, e.g. many VOICE specifics, as the input of multilingual speakers, as well as markers of spoken language, and proper nouns. Where needed, additional tags were added to any tokens in the lexicon in order to account for non-canonical functions in VOICE. For example, for *partly* in the sequence *a partly answer*, we allowed for the tag JJ, in addition to RB. This last step was a dynamic process throughout the duration of the tagging of VOICE, in which new functions were added as we encountered them.

2.2.3 Choice of taggers and tagging procedure

After an initial trial run of a tagger on our spoken ELF data, in which we tested a number of utterances from VOICE on TreeTagger (Schmid 1994), achieving an accuracy of only 83-86% (Osimk-Teasdale 2013), we also tested other taggers, e.g. the Stanford tagger (Toutanova et al. 2003) and LTAG (Shen, Satta & Joshi 2007), achieving similar results; we furthermore attempted to improve accuracy by implementing hybrid systems (Brill & Wu 1998; van Halteren, Daelemans & Zavrel 2001; Wu, Ngai & Carpuat 2004; Radeka 2009) and by applying domain adaptation techniques (Daume III, Kumar & Saha 2010; Radeka in prep.). As expected, all of these achieved much lower accuracy rates on VOICE data (Radeka in prep.) than on the formal written native language they were trained on. In order to achieve higher tagging accuracies, we adapted both tagset and tagging strategy to the needs of our data. In order to make an informed decision for the part-of-speech tagging of VOICE, we took account of the various tagging procedures used by state-of-the-art taggers, i.e. Decision Trees (Schmid 1994), Maximum Entropy (Ratnaparkhi 1996; Toutanova et al. 2003), Support Vector

Machines (SVM) (Giménez & Marquez 2003), Transformation-based Learning (TBL) (Brill 1995), Memory-based Learning (Daelemans et al. 1996), Conditional-Random-Fields (Lafferty, McCallum & Pereira 2001), as well as Markov Models (Brants 2000). All these achieve similar accuracies on different types of data. In order to establish procedures relevant to our data, we chose to use a combination of three different types of taggers within a stacked-TBL framework, similar to N-fold Templated Piped Correction (Wu, Ngai & Carpuat 2004). This procedure of using three different taggers meant that the advantages of each individual tagger could be capitalized on. TBL is a system which corrects errors by formulating rules from the tagger output in comparison with a correctly annotated version. The advantage of these rules (in contrast to other tagging systems) is that they are transparent and therefore interpretable for a human annotator. Thus, TBL can also be used to detect regularities in forms of rules. These rules can be used as diagnosis tools for individual tagging systems, as well as data analysis tools, in that they can highlight interesting aspects of the data. TBL helps to analyse where and why tagging errors occur, and, as a consequence, allows the annotator to create manually crafted rules, which help to improve tagging accuracy (cf. Volk & Schneider 1998). Another important criterion for a suitable tagger for our data was a high degree of flexibility. TBL met this need as it allowed us to keep the lexicon separate from the disambiguation, and hence to adapt the lexicon, as well as the tagset and tagging format of the Penn Treebank, for our specific purposes.

The three taggers which we used in combination for the stacked TBL framework were TreeTagger (a Decision Tree tagger), Stanford Tagger (a bi-directional Maximum Entropy tagger), and LTAG (a bi-directional perceptron-like tagger). These taggers were first applied to 5 000, later to 10 000 words of data and subsequently compared to the manual annotation of these data. These 10 000 words of data had been manually tagged, the tagging results compared and, in diverging cases, discussed, by 2 project researchers. The manually and the automatically annotated data were then used as training sources for TBL. What TBL did was to generate rules based on the differences of these annotations in order to increase the tagging accuracy. In a second step, the remaining, not manually annotated, part of VOICE was tagged with the three taggers and the rules generated in the first step were applied to the taggers' output. Now the most frequently applied tags by the three taggers were chosen for each token, in a procedure commonly referred to as Voting (Brill & Wu 1998). From this, tag probabilities for each token in VOICE were calculated and added to the VOICE lexicon. This now complete lexicon, based on the estimated tag probabilities occurring in VOICE, was then used to initiate a parallel TBL procedure (Radeka 2009) in which replacement (Brill 1995) and reduction rules (Lager 2001) are learned. Through the analysis of these rules we gained systematic insights into the strengths and weaknesses of the taggers' output, which in turn highlighted which aspects a) could be fixed by rule modification or manual rule creation, and b) which needed to be annotated manually. In fact, this resulted in rather large parts of the corpus being annotated and disambiguated manually, e.g. discourse markers, multi-word items, long-distance dependencies, words standing alone or at the beginning or end of utterances. This manual annotation in combination with the stacked TBL procedure helped to further increase tagging reliability. We intend to carry out a more detailed analysis of the final tagging accuracy of the VOICE corpus in future research.

2.3 Lemmatization

Lemmatization is the process of grouping together word forms that belong to the same inflectional paradigm and assigning to each paradigm its corresponding uninflected form, called a 'lemma'. This form of annotation is related to the process of part-of-speech tagging, as the latter is a prerequisite

for lemmatization. Lemmatization is regarded as a highly useful type of corpus annotation because it provides additional search options.

The lemmatization of VOICE was carried out by an especially designed lemmatizer which was implemented in Python. This lemmatizer accessed the VOICE Lexicon in its final, completed version and retrieved the appropriate lemma from there. The lemmatizer did this by applying a number of manually implemented rules, which were related to the information contained in the POS tags (e.g. for regular verb forms: if tag for word X is VV (verb, base form), then lemma equals token), or, sometimes, also to the morphological information contained in the tokens (e.g. for regular adjective forms: if tag for word X is JJR (adjective, comparative), then lemma equals token without the suffix -*er*).

3 The VOICE part-of-speech tagging guidelines

3.1 Guiding principles

The main goal of POS tagging VOICE, apart from gaining insights into the data as such, was to develop a tagging procedure and a scheme as appropriate as possible to the ways in which English as a lingua franca is used in the interactions recorded and transcribed in the corpus. This meant it was important that tagging should be compatible with the variable character of English as a lingua franca, and a perspective in which the speakers are primarily viewed as language users in their own right, rather than 'language learners' (Osimk-Teasdale 2013).

The main principles that guided the process of part-of-speech tagging VOICE were the following:

a) We gave an **ELF perspective** priority at all times (see e.g. Seidlhofer 2011). This means that we generally asked ourselves first and foremost which tags it made most sense to assign to specific positions from an ELF point of view. This had priority over the second step, in which we asked how these decisions could be carried out practically, i.e. how to proceed in accordance with good practice. Our ELF guiding principle sometimes resulted in decisions that posed complex challenges to carrying out the manual and automatic part-of-speech annotation and the technical implementation of tagging decisions.

b) For tag categorization and technical implementation, we relied on **established procedures** as far as possible. This means that whenever established tagging procedures used for other corpora or available descriptions of English(es) were compatible with our ELF perspective, we adopted these. The *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* (Santorini 1991) served as a starting point for the VOICE Tagset. These we modified and extended in order to account as appropriately as possible for the ELF data. This included using the format FORM(FUNCTION) (cf. 3.2 Tagging formats in VOICE), as well as modifying the tagset originally used for the Penn Treebank Project, including tags for various characteristics of spoken language and a large number of discourse markers. In addition, the 7th edn. of the *Oxford Advanced Learner's Dictionary* (OALD7) was used as external reference dictionary for tagging decisions[2].

[2] cf. Breiteneder et al. (2006: 179ff.) and Pitzl, Breiteneder & Klimpfinger (2008: 25) for the reasoning behind using OALD7 for VOICE.

c) As a further guiding principle, we aimed at a tagging scheme which, while acknowledging the variable character of spoken ELF, would be **intuitively accessible** for researchers working with VOICE. We tried to achieve this by adhering to established points of reference, both **internal**, i.e. tagging guidelines in line with the already existing transcription conventions of VOICE (VOICE Project 2007b), as well as **external**, e.g. the *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* (Santorini 1991), the OALD7 (2005), and commonly acknowledged word class categories in general.

d) We tried to strike a **balance** between **inevitable interpretation**, i.e. **leaving options open** on the one hand, and avoiding **potentially excessive complexity** for corpus users, on the other. The former we did by making frequent use of **ambiguous tags**, in the format, and by introducing the basic tagging format **FORM(FUNCTION)** for all tokens (cf. 3.2 Tagging formats in VOICE for an explanation). In order to reduce complexity, however, we only allowed a maximum of **two tags per token,** e.g. *so_DM/RB.*[3] As this guideline was sometimes difficult to implement, especially with verb and noun forms, we introduced **'generic'** verb and noun tags (V and N, respectively), as in the example *xxx remark_NN/**V** anyway*. Here the token *remark* can be assigned the tag NN (noun), but the tags VVP (present tense verb) or VV (base form) are also possible, due to the lack of disambiguating co-text (*xxx* indicating unintelligible speech). This would, however, result in the assignment of more than two tags, and hence a generic verb tag (V) is given instead of two sub-specified verb tags (VVP,VV) (cf. also 3.3.2 The commented VOICE Tagset). In cases where more than two tags were possible for one token, and these could not be simplified with a generic verb or noun tag, the **tag 'unknown' (UNK)** was assigned. For example, in the sequence *xxx like xxx*, the tags VVP, VV, IN or DM would have been possible for 'like' (cf. 3.4.2 Tagging of individual elements), but no decision for one tag was possible due to the lack of disambiguating co-text, hence the tagging was *like_UNK*. Another way in which we tried to minimize the range of possibilities for interpretation was to use a **sequential tagging procedure** as consistently as possible. This means that wherever there was more than one way of interpreting a stretch of tokens, we assigned only those tags which conformed to a 'sequential', i.e. left to right, reading of the tokens. For example, in the stretch *this is the first slides,* the token *slides* was interpreted as noun with plural form and singular function (tagging: slides_NNS(NN)), due to the tokens *this is* preceding the token *slides*, indicating singular function. The (also plausible) reading of *is* having singular form and plural function instead, was not taken into account.

[3] NB: one "tag" always consists of a form-tag and a function-tag in VOICE POS Online (cf. 3.2. Tagging formats in VOICE).

3.2 Tagging formats in VOICE

Category	Format
Form and function tags **Format: TAG(TAG)**	For all tokens in the corpus, separate tags for paradigmatic form and syntagmatic function are assigned. The tag for form is indicated first, followed by a tag for function, given in brackets. Format: **FORM-tag(FUNCTION-tag)** There are **2 options** of this format: **OPTION 1:** form and function converge → identical form(function) tag is assigned, e.g. *a* ***house_NN(NN)*** **OPTION 2:** form and function do not converge → different tags for form and (function) are assigned, e.g. *two* ***house_NN(NNS)*** **NB:** The format FORM-tag(FUNCTION-tag) is relevant when working with VOICE POS Online, as users are able to search for form- and function-tags separately. The default search in VOICE POS Online always considers positions for both form- and function-tags. For example, the search *NNS* will yield all of the following results: • multicultural teams_**NNS(NNS)** • in one countries_**NNS**(NN) • three university_NN(**NNS**). For the sake of simplicity, for examples in this tagging manual only one tag will be indicated whenever it is implied that form-tag and function-tag converge. Hence, e.g. *the group* will be indicated as *the_DT group_NN*, not *the_DT(DT) group_NN(NN).* Both tags, i.e. FORM-tag(FUNCTION-tag), will be indicated in this manual only when form and function-tag do not converge. For any token, a **maximum number of two tags** is allowed, whereby a "tag" refers to a form and a function-component, as in e.g. *so_IN(IN)/RB(RB).*

2 TAGGING FORMATS

1) Non-ambiguous POS tag
(one word class category is assigned to a token)

Format: TAG	1 tag is assigned, e.g. *be_VB* → Format in VOICE POS Online: *be_VB(VB)*

2) Ambiguous POS tag
(two possible word class categories are assigned to a token)

Format: TAG/TAG	2 tags are assigned with tags in alphabetical order, separated by a slash e.g. *use a maltese word or* ***joke_NN/VVP*** *in maltese* → Format in VOICE POS Online: *joke_NN(NN)/VVP(VVP)*

3.3 VOICE Tagset

The list of part-of-speech tags used for annotating VOICE data with word class categories is called the VOICE Tagset. The VOICE Tagset consists of 69 different part of speech tags, which are largely based on the *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* (Santorini 1991). These were designed for written data and then extended for the SwitchBoard corpus (cf. Linguistic Data Consortium (LDC 1999). However, in the course of tagging VOICE, it became clear that both the tagset as well as the tagging format used for the Penn Treebank needed to be modified and extended for our kind of data. Some substantial changes were therefore made in order to make the tagset better suited for the character of spoken, interactive ELF data. All additions to the tagset used for the Penn Treebank, as well as all changes in a tag's categorization are marked in **green**. Section 3.3.1 provides an alphabetically sorted list with regard to categories. Examples and explanations for tags are provided in section 3.3.2.

3.3.1 The VOICE Tagset, sorted alphabetically according to categories

Category	Tag
Adjective	**JJ**
Adjective, comparative	**JJR**
Adjective, superlative	**JJS**
Adverb	**RB**
Adverb, comparative	**RBR**
Adverb, superlative	**RBS**
Anonymization	**NP**, additionally marked **a_** preceding the token
Breathing	**BR**
Cardinal Number	**CD**
Conjunction, coordinating	**CC**
Conjunction, subordinating	**IN**
Contracted 's	**DOS** = does **VBS** = is (=BES in Switchboard) **VHS** = has (=HVS in Switchboard) **POS** = possessive **PP** = personal pronoun ***us*** (PRP in Switchboard)
Determiner	**DT**
Discourse Marker	**DM** (single discourse markers) **FORM-tag(FUNCTION-tag:DM)** (multi-word discourse markers) (cf. 3.3.2)
Foreign Word (Non-English speech)	**FW,** additionally marked **f_** preceding the token
Formulaic Item	**FI**
Interjection	**UH**
Laughter	**LA**
List Item Marker	**LS**

Category	Tag
Noun, generic	**N**
Noun, plural	**NNS**
Noun, singular or mass	**NN**
Onomatopoeia	**ONO**
Partial Word	**XX**
Particle	**RP**
Pause	**PA**, annotated **_0, _1, _2,** … (numbers indicate pause duration)
Possessive Ending	**POS**
Predeterminer	**PDT**
Preposition	**IN**
Pronoun, personal	**PP**
Pronoun, possessive	**PP$**
Pronoun, relative	**PRE**
Pronunciation Variation and Coinages (PVC)	**FORM-tag:PVC(FUNCTION-tag),** additionally marked **p_** preceding the token (cf. 3.3.2)
Proper Noun, plural	**NPS**
Proper Noun, singular	**NP**
Spelt	**SP,** additionally marked **s_** preceding the token
Response Particle	**RE**
Symbol	**SYM**
***there*,** existential	**EX**
***to*,** infinitive use	**TO**
Unintelligible Speech	**UNI**
Unknown	**UNK**
Verb, base form	**VB** = verb be **VH** = verb have **VV** = all other verbs (all = VB in Penn Guidelines)
Verb, generic	**V**
Verb, gerund or present participle	**VBG** = verb be **VHG** = verb have **VVG** = all other verbs (all = VBG in Penn Guideline)
Verb, modal	**MD**

Category	Tag
Verb, past participle	**VBN** = verb be **VHN** = verb have **VVN** = all other verbs (all = VBN in Penn Guidelines)
Verb, past tense; includes the conditional form of the verb *be*	**VBD** = verb be **VHD** = verb have **VVD** = all other verbs (all = VBD in Penn Guidelines)
Verb, present, non-3rd person singular	**VBP** = verb be **VHP** = verb have **VVP** = all other verbs (all = VBP in Penn Guidelines)
Verb, present, third person singular	**VBZ** = verb be **VHZ** = verb have **VVZ** = all other verbs (all = VBZ in Penn Guidelines)
Wh-adverb	**WRB**
Wh-determiner	**WDT**
Wh-pronoun	**WP (VOICE:** only tagged **WP** when ***not*** used as a relative pronoun! → else **PRE tag)**

3.3.2 The commented VOICE Tagset, sorted alphabetically according to tags[4]

Tag	Explanation and examples
BR	**Breathing**, e.g. *hh, hhh, hhhh*
CC	**Coordinating conjunction**, e.g. *and, but, or*
CD	**Cardinal Number**, e.g. *one, twenty-eight* (**VOICE:** also including *zero*),
DM	**Discourse Marker**[5]. Discourse markers are words which have homonyms in other word class categories and can function as discourse markers. VOICE operates with a closed list. A distinction is made between SINGLE and MULTI-WORD Discourse Markers: **1) SINGLE WORD DISCOURSE MARKERS:** **Items:** *like, look, whatever, well, so, right* **Tag**: DM **2) MULTI-WORD DISCOURSE MARKERS** **Items**: *I mean, I see, mind you, you know, you see* **Tags:** Multi-word discourse markers are tagged with a conventional word class tag for FORM and the tag DM for (FUNCTION): *I_PP(DM) mean_VVP(DM)* *I_PP(DM) see_VVP(DM)* *mind_VVP(DM) you_PP(DM)* *you_PP(DM) know_VVP(DM)* *you_PP(DM) see_VVP(DM)*
DOS	for **contracted 's**, **DOS** = does, e.g. *Where**'s** she live?*
DT	**Determiner**, e.g. *a, the, that* Some items, such as *that*, are also tagged DT when occurring without a head noun (analogous to Santorini 1991: 8)
EX	***there***, existential
FI	**Formulaic Items**, includes all formulaic expressions which are in the closed list "VOICE Formulaic Expressions", e.g. greetings, farewells, thanks, apologies, wishes, miscellaneous expressions. (cf. 6.1. VOICE List of Formulaic Items)
FW	**Foreign word** (Non-English speech), e.g. *francais.* Additionally marked with the prefix f_ in VOICE POS XML and VOICE POS Online.
IN	**Preposition** or **subordinating conjunction**, e.g. *because, behind*
JJ	**Adjective**, e.g. *good*
JJR	**Adjective**, comparative, e.g. *better*
JJS	**Adjective**, superlative, e.g. *best*

[4] Since the *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* (Santorini 1991) served as a starting point for the VOICE Tagset, we are also using the explanations and partly also the wording used there. Note that these guidelines differ in a number of ways from later revised versions which include some tag changes which had to be made for the bracketing procedure (Santorini 1995). All changes to the original Penn Tagging Guidelines and adaptations made for VOICE are marked in green.

[5] Please note: 'mind you' and 'you see' are included in this list of discourse markers and in the Appendix (6.3.2 Multi-word discourse markers, p. 31 below). Unfortunately, due to an oversight the tagging of these two discourse markers does not appear in the published versions of VOICE POS. However, lists of these two discourse markers as they occur in VOICE and with the appropriate tags can be requested by sending an e-mail to <voice@univie.ac.at>.

Tag	Explanation and examples
LA	**Laughter**, e.g. *@, @@, @@@*
LS	**List Item Marker**, e.g. *section d_LS*
MD	**Modal**, e.g. *can, could, might, may*
N	**Generic Noun Tag**, used instead of ambiguous noun tags, e.g. NN/NNS or NP/NPS, primarily in tagging where there is a difference in form and function, e.g. *to: (.) register (.) to our lectures? and (.) stuffs_VVZ(N)* (cf. also Generic Verb Tag)
NN	**Noun**, singular or mass, e.g. *house, water*
NNS	**Noun**, plural, e.g. *houses*
NP	**Proper Noun**, singular, e.g. *european union*
NPS	**Proper Noun**, plural, e.g. *the netherlands_NPS*
ONO	**Onomatopoeic noises**, all onomatopoeia are represented in IPA-signs and are additionally marked with the prefix o_, e.g. *o_kr_IPA* in VOICE POS XML and VOICE POS Online.
PA	**Pause**, annotated with an underscore, followed by a number indicating the length of the pause in seconds (0 referring to up to approximately 0.5 seconds), e.g. *_0, _1, _2, ...*
PDT	**Predeterminer**, e.g. *all, both* when preceding a determiner
POS	**Possessive Ending**, e.g. for **contracted 's**, **POS** = possessive, e.g. *maria theresia's_POS eyes.*
PP	**contracted 's**, **personal pronoun *us***, e.g. *yeah let's_PP do something*, **possessive** and **reflexive pronouns** without case distinction, e.g. *they_PP knew that, do it yourself_PP*
PVC	**Pronunciation Variations and Coinages**, all items annotated <pvc> </pvc> in the transcription process were assigned the FORM-tag **PVC** and a suitable part-of-speech tag for function. Tokens given the tag PVCs are additionally marked with the prefix p_ preceding, e.g. *p_associational_PVC(JJ)* in VOICE POS XML and VOICE POS Online.
PP	**Pronoun,** personal, e.g. *I, me, you, he*
PP$	**Pronoun,** possessive, e.g. *my, your, mine, yours*
PRE	**Pronoun,** relative. Closed list: *that, which, who, whom,* and *whose.*
RB	**Adverb**, most words that end in *-ly* as well as degree words, e.g. *quite, too, very*
RBR	**Adverb**, comparative. Refers to adverbs with the comparative ending *-er*, with a strictly comparative meaning, e.g. *they are better_RBR recognized*
RBS	**Adverb**, superlative, e.g. *the most_RBR important education*
RP	**Particle,** e.g. *set up_RP support*
RE	**Response particle**, e.g. positive and negative minimal feedback, e.g. *no, yes, yeah, okay, yep, nah* (cf. 6.3.3 Interjections)
SP	**Spelling out,** referring to spelt items which could not be categorized further (cf. 3.1.2.11 Spelling out), additionally marked with the prefix s_ in VOICE POS XML and VOICE POS Online.
SYM	**Symbol**, used for mathematical, scientific or technical symbols, e.g. *x_SYM axis*
TO	***to***, infinitive use, e.g. *to_TO introduce*

Tag	Explanation and examples
UH	**Interjections** **CATEGORIZATION FOR VOICE:** these are markers of spoken discourse (e.g. hesitation markers) which do not have homonyms in other word class categories (as opposed to tokens tagged DM), e.g. *er, erm, yipee, whoohoo, mm:, haeh; a:h, wow* (cf. 6.3.3. Interjections).
UNI	**Unintelligible speech**, e.g. *x, xx, xxx*
UNK	**Unknown**, used for words which are ambiguous between more than two word class categories, e.g. due to lack of co-text.
V	**Generic Verb Tag**, used instead of ambiguous verb forms e.g. VV/VVP VVD/VVN, primarily in tagging where there is a difference in form and function, e.g. *will be communicate_V(VVG) with* (cf. also Generic Noun Tag).
VB/VH/VV (all = **VB** in Penn Guidelines)	**Verb, base form**, subsumes imperatives, infinitives and subjunctives VB = verb be VH = verb have VV = all other verbs
VBD/VHD/VVD (all = **VBD** in Penn Guidelines)	**Verb, past tense**; includes the conditional form of the verb *to be* VBD = verb be VHD = verb have VVD = all other verbs
VBG/VHG/VVG (all = **VBG** in Penn Guideline)	**Verb, gerund or present participle** VBG = verb be VHG = verb have VVG = all other verbs
VBN/VHN/VVN (all = **VBN** in Penn Guidelines)	**Verb, past participle** VBN = verb be VHN = verb have VVN = all other verbs
VBP/VHP/VVP (all = **VBP** in Penn Guidelines)	**Verb, present, non-3rd person singular** VBP = verb be VHP = verb have VVP = all other verbs
VBS	for **contracted 's**, **VBS** = be, e.g. *Tom's an excellent teacher.*
VBZ/VHZ/VVZ (all = **VBZ** in Penn Guidelines)	**Verb, present, 3rd person singular** VBZ = verb be VHZ = verb have VVZ = all other verbs
VHS	for **contracted 's**, **VHS** = have, e.g. *She's bought a nice dress.*
WDT	**Wh-Determiner,** e.g. *what, which, whatever* **VOICE**: Not used for relative pronouns. Used for e.g. *what_WDT kind* (vs. *what_WP do you like)*, also: *which, whichever, whatever*. Original Penn Treebank Guidelines: Wh-determiner e.g. *which*, and *that* when it is used as a relative pronoun.
WP	**Wh-pronoun**, e.g. *what, who, whom* **VOICE:** only tagged **WP** when ***not*** used as a relative pronoun, else tagged **PRE.**

Tag	Explanation and examples
WRB	**Wh-adverb**, e.g. *how, where, why, when* When used to introduce a relative or an interrogative clause.
XX	**Partial words,** e.g. *becau-* Corresponding to "Word fragments" in the VOICE Mark-up conventions (VOICE Project 2007c: 3), the absent part is indicated with a hyphen.

3.4 Further specifications on the tagging of VOICE

3.4.1 Tagging of individual categories

Category	Tagging practice
3.4.1.1 ANONYMIZATION	
	All anonymized tokens are labelled with the prefix **a_[...]** and are tagged **NP**, e.g. a_[S1]_NP, a_ [org4]_NP, a_[place5]_NP
3.4.1.2 COLLECTIVE NOUNS	
	Collective nouns are tagged singular or plural depending on whether the following verb is singular or plural. This is in line with the Penn Treebank Guidelines (cf. Santorini 1991: 18). For VOICE POS Tagging, we follow the rather broad definition of Carter & McCarthy (2006: 541), who state that a collective noun is "[a] type of noun referring to a group of people, animals or things", as well as the examples given in Carter & McCarthy (2006: 539) and Quirk (1997: 316f.). Not included in our definition of collective nouns are cases in which names of countries are used representatively for the population of a country, as in the example below. In these cases, the verb which follows is tagged with differing tags for FORM and (FUNCTION), e.g. *the rest of the country need_V(VVZ) it as well*
3.4.1.3 –ING CATEGORY	
	In dealing with ELF data, it was often extremely difficult to decide whether a word ending in *–ing* should be classified as verb, noun or adjective. Hence, it was decided that **all words** in VOICE ending in the morpheme *–ing* would be given a **uniform FORM**-tag, namely **VVG**, and a **(FUNCTION)** tag according to their **syntactic co-text**. For this category, the **FORM**-tag **VVG** stands for any word ending in the morpheme *–ing* (potentially followed by a plural *-s* morpheme). For the **(FUNCTION)**-tag, we only differentiated between either **VVG** and **NN** or **NNS**: The tag **VVG** was given when the word functioned as a present participle, and also when used as a participial adjective. The function-tags **NN** or **NNS**, respectively, were given when the word functioned as a singular or plural noun. **Tagging examples:** 1. Word ending in *–ing* functions as **verb** or a **participial adjective**: **TAG=VVG(VVG)**, e.g. *swimming_VVG(VVG)* man. 2. Word ending in *–ing* functions as **noun:** **TAG=VVG(NN),** e.g. the real *meaning_VVG(NN)* The only exception was made for words which end in *–ing* and are listed as **adjectives** in the **reference dictionary** (OALD7) and which we regard as lexicalised for the tagging of VOICE), and tagged with **JJ**, the tag for adjectives. 1. Word ending in *–ing* is an **adjective** in OALD7: **TAG=JJ**, e.g. *charming_JJ man*

–ING CATEGORY cont.

NB: **Compound nouns** with one of the parts ending in the morpheme *-ing* (e.g. *swimming pool*), were also tagged according to the procedure described above, e.g. deciding whether the individual parts of the compound functioned as nouns in their immediate co-text (tagging: *swimming_VVG(NN) pool_NN(NN)*). Thus, for these cases, we did not consult the reference dictionary OALD7, as we did for all other compounds listed in the VOICE List of Compound Nouns (Section 6.4). Hence, the compound noun combinations with one component ending in *-ing*, such as *swimming pool*, are not included in the VOICE List of Compound Nouns.

3.4.1.4 MULTI-WORD ITEMS

As **multi-word items** we understand sequences of tokens which, grammatically, seem to 'belong' together and thus, form a single unit.

All parts of a multi-word item are assigned **identical tags**, e.g. *per_RB se_RB; student_NN union_NN.* If the **head** of the multi-word item is marked plural, all parts are given a plural tag, e.g. *points_NNS of_NNS view_NNS, youth_NNS organizations_NNS*

There are **5 types of multi-word items**:

1. Compound Nouns (cf. 6.4 VOICE List of Compound Nouns)
2. Items in VOICE Multi-words (cf. 6.2 VOICE List of Multi-words)
3. Multi-word Discourse Marker (cf. 6.3.2 Multi-word discourse markers)
4. Multi-word Formulaic Items (cf. 6.1 VOICE List of Formulaic Items)
5. Proper Nouns and Names (cf. 3.4.1.7 PROPER NOUNS (NP,NPS) vs. COMMON NOUNS (NN,NNS))

3.4.1.5 PARTLY UNINTELLIGIBLE

Tokens of which parts are annotated as unintelligible (marked <un>x</un> in VOICE Online) are given the tag UNI. This means that the part that was intelligible to the transcriber is also assigned the tag UNI, e.g.
VOICE Online: ***super<un>x </un>***
VOICE POS Online: ***superx_UNI***

3.4.1.6 PRONOUNS

For the tagging of VOICE, a distinction is drawn between the following pronouns:

1. Personal pronouns (Tag: PP)
2. Possessive pronouns (Tag: PP$)
3. Relative pronouns (Tag: PRE)
4. Wh-pronouns (Tag: WP)

Other pronouns are not assigned an individual tag category but are subsumed under other part-of-speech categories. For example, **demonstrative pronouns** such as *this* in *it was very nice you let us do **this**,* are tagged DT, and **indefinite pronouns**, such as *someone* in ***someone** is waiting,* are tagged NN. The **reciprocal pronoun** *each other* is also tagged NN (cf. Biber 1999: 70f. for an overview of the different pronouns in English).

3.4.1.7 PROPER NOUNS (NP,NPS) vs. COMMON NOUNS (NN,NNS)

General guidelines:

1. The tag **NP** includes **Proper Nouns** (which belong to the category noun e.g. *America*) as well as **Proper Names** (i.e. a combination of a proper noun with other words as *United States of America*) if they refer to a single entity.
2. **External references**: In some cases we oriented ourselves towards our reference dictionary OALD7 and tagged as proper noun when it was capitalized there, e.g. with regard to **alcoholic drinks, festivals**. If necessary, other dictionaries and search engines were consulted.
3. **Multi-word tag for proper nouns:** For titles of films, books etc. we use a **multi-word tag**, i.e. every word is assigned the NP tag even if it is not a noun, e.g. *good_NP night_NP and_NP good_NP luck_NP.* This is an **open list** and is not included in the VOICE List of Multi-words (cf. 6.2).
4. **Compound nouns**: For proper and common compound nouns where the **head is plural**, the word preceding or following the head is tagged plural NNS or NPS respectively, e.g. *swimming_NNS pools_NNS, points_NNS of_NNS view_NNS.*
5. **Form(Function) tags:** For proper nouns and names we usually did *not* use OALD7 as an external reference for paradigmatic form and syntagmatic function, as OALD7 does not list the majority of proper nouns and names occurring in VOICE. Sometimes this would have also resulted in odd combinations of tags for form and function, e.g. *Goofy* is only listed as JJ in OALD7, but occurs in VOICE as the Disney character → we tagged *NP*, not *JJ(NP).*

Tag NP or **NPS** is used for:

- **Alcoholic drinks and brands** (if capitalized in OALD7), e.g. *desperados, beaujolais*
- **Car names** and **names for aeroplanes**, e.g. *audi, jumbolino, saabs*
- **Currencies**, e.g. *lek, rouble, lei, dinar*
- **Days of the week, months,** e.g. *tuesday, may*
- **Famous personalities, groups etc.**, e.g. *aristotle, the smiths*
- **Languages,** e.g. *finnish*
- **Names of people, places, institutions, companies, programmes**, e.g. *nato, erasmus*
- **Names of products,** e.g, *ajax*
- **Nationalities**: e.g. *dane*
- **Professional terminology**, such as terms for **mathematical concepts**, e.g. *cauchy fanapppi*
- **Recurrent festivities** and **public holidays**, e.g. *christmas, ramadan*
- **Religious and spiritual terms** e.g. *feng shui, catholicism*
- **Religious denominations**, e.g. *christian(s), muslim, jews, baha'is*
- **Titles of films, books, names of websites**, e.g. *guinness book of records, youtube*

Tag NN or **NNS** is used for:

- **Alcoholic drinks** (if not capitalized in OALD7), e.g. *tequila*
- **Chemical elements**, e.g. *lithium chloride*
- **Diseases**, e.g. *meningitis, flu*
- **Food and beverages**, e.g. *goulash, rooibush*
- **Ordinal numbers in dates,** *e.g. the first_NN of October*
- **Titles**, e.g. *doctor, missis* (unless occurring as part of a proper name, e.g. *queen_NP elizabeth_NP)*

3.4.1.8 RELATIVISERS	
	For relativisers, a distinction is made between **relative pronouns** (*that, which, who, whom, whose*), which are tagged **PRE** and **relative adverbs** (*how, where, why, when*), which are tagged **WRB**. In this, we follow the distinction between these two categories drawn by Biber et al. (1999: 608).
3.4.1.9 SPELLING OUT	
	Items which are spelt are tagged as if they were spelt out normally, e.g. *eu* = "European Union" = NP, *tv* = "television" = NN. This refers to English as well as non-English speech, e.g. *oebb* (Austrian federal railways, a company) is tagged NP, not FW. Items which are spelt are additionally marked with the prefix *s_* before the spelt item, e.g. *s_eu* The sub-categorization is as follows: ○ **CD** in place of a number, e.g. *if we have **s_x_CD** universities* ○ **SYM** for mathematical symbols ○ **LS** for list items ○ **NN** or **NNS** for spelt items which stand for nouns or function as nouns, e.g. if they can be pluralized. The same holds true for spelt items which function as Proper Nouns or Names (Tag **NP** or **NPS**), Verbs (corresponding verb-tag, e.g. **VVP**), etc. ○ **SP** in case of 'real spelling' or if the spelt item could not be identified further.
3.4.1.10 UNCERTAIN AND PARTLY UNCERTAIN SPEECH	
	Both uncertain and partly uncertain speech are not marked as such in VOICE POS, i.e. uncertain speech in VOICE Online marked with brackets '(...)' is treated as normal text in VOICE POS, and no longer indicated with brackets. These items are assigned a POS tag referring to the token without consideration of the brackets signalling uncertainty. e.g. **Uncertain speech:** VOICE Online: *yeah **(just about)*** VOICE POS Online: *yeah **just_RB about_IN*** e.g. **Partly uncertain speech:** VOICE Online: *a variety of **instrument(s)*** VOICE POS Online: *a variety of **instruments**_NNS*

3.4.2 Tagging of individual elements

As with the tagset, we used the *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* as starting point for the tagging of individual tokens (cf. 1991: 23ff.). The items listed below are cases we encountered in our data which did not have a corresponding guideline in Santorini (1991), or cases in which we found the guideline was not suitable for our data. In these cases, other external references (e.g. OALD7, other corpora, dictionaries and grammars) were consulted in order to decide on a suitable tagging scheme.

Individual token(s)	Tagging practice
ain't	*ai_VVZ n't_ RB* *ai_VVP n't_RB* (*ai_VHZ n't_RB*; *ai_VHP n't_RB* would also be possible in theory but do not occur in VOICE)
altogether	*altogether_RB*
and so on	*and_CC so_RB on_RB*
and that	**Meaning 'and similar':** e.g. *faked diamonds and_CC that_DT*
as regards	*as_IN regards_VVZ*
as such	*as_IN such_DT*
as well as	*as_RB well_RB as_IN*
get rid of, rid	*get_VV,VVP rid_VVN of*
gonna	*gon_VVG na_TO*
got	1. If clearly identifiable as participle → tag VVN, e.g. *have got_VVN*, 2. If simple past, or no or too little co-text to identify as past participle → tag VVD, e.g. *she got_VVD*
gotta	*got_VVD ta_TO*, or *got_VVN ta_TO* (see criteria for distinguishing between VVD and VVN cf. *got*)
how come	*how_WRB come_VV*, e.g. *how come the austrian are perceived as hh as being drunk all the time*
like	**Verb Present**: e.g. *I like_VVP it* **Verb Base Form**: e.g. *you don't like_VV the people* **Conjunction, Preposition**: e.g. *something like_IN that*; *it looks like_IN a sauce, I'll do it like_IN this* **Discourse Maker**: e.g. *they put like_DM erm poison all around*; *I was like_DM*
never mind	*never_RB mind_VV*
no	**Adverb**: e.g. *i am no_RB longer affiliated* **Determiner**: e.g. *there's no_DT money* **Response Marker**: e.g. *no_RE but i can make a chick break for you (.)*
okay	**Adjective**: e.g. *is this okay_JJ for everyone* **Adverb**: e.g. *we're doing okay_RB* **Response marker**: e.g. *okay_RE, I'll do it.*

Individual token(s)	Tagging practice
so	**Meaning "so that" and "therefore"**: Tag **IN**, e.g. *[first name1] will be there so_IN he will have the occasion to speak out there* **Adverbial use: Tag RB**, e.g. *so_RB good* **In certain fixed expressions: Tag RB**, e.g. *or so_RB, and so_RB on* **Clause-final** or **not related to main clause: Tag DM,** e.g. *pedagogical way so_DM _0*
so that	**Subordinating conjunction**: e.g. *so_IN that_IN WE do not go to the politicians* **Discourse Marker, followed by Determiner**: e.g. *so_DM that_DT's strong*
such	**Predeterminer:** e.g. *such_PDT a darling* **Determiner:** e.g. *such_DT documents*
that	**Determiner**: e.g. *put that_DT thing away*; *that_DT came after the mcsherry re-reform* **Relative pronoun**: e.g. *first thing that_PRE crosses your mind* **Subordinating conjunction**: e.g. *the problem was that_IN she was like RUNNING*
the -er the -er	*the_DT broader_JJR the_DT better_JJR*
the same	*the_DT same_NN* (if no noun is following)
though	**Conjuction**: e.g. *though_IN we are prepared* (*even_IN though_IN*, cf. 6.2. VOICE List of Multi-words) **Adverb**: e.g. *you know what's funny though_RB*
to	**Infinitive use**: e.g. *to_TO go* **Preposition:** e.g. *to_IN the market*
up to	*it's up_RB to_IN the labor market*, *up_RB to_IN fifteen minutes* (vs. *up_JJ to_JJ date_JJ*, cf. 6.2. VOICE List of Multi-words)
use(d) (to)	**Adjective**: e.g. *i'm used_JJ to it*; *a used_JJ car* **Verb**: e.g. *I used_VVD to do something*; *the designers have used_VVN that*

4 The VOICE lemmatization guidelines

This section provides general information on how the VOICE tagging formats were treated in the lemmatization process, followed by an explanation of the lemmatization rules for individual categories, listed in alphabetical order.

4.1 Lemmatization and VOICE Tagging formats

In VOICE, all tokens are assigned a maximum of two tag combinations in the basic format FORM(FUNCTION). However, due to this format, and the format for tag ambiguities, each basic tag can consist of more than one individual tag (cf. 3.2. Tagging formats in VOICE). Hence, each basic tag can be assigned more than one lemma. The rules for lemmatization are as follows: If the **lemmata** of the individual tags **converge**, only one lemma is assigned, e.g. token: *rules_NNS/VVZ*, lemma: *rule*. In those cases where the **lemmata** for the individual tags **do not converge**, more than one lemma is assigned. This was often the case for ambiguities, e.g. token: *including_IN/VVG*, lemmata: *including, include,* and with tokens that were assigned different tags for form and function, e.g. token: *feeling_VVG(NN)*, lemmata: *feel, feeling;* token: *preserved_V(JJ)*, lemmata: *preserve, preserved*.

4.2 Contracted forms

Contracted forms are assigned the corresponding lemma of their full forms, e.g. token: *'ve* (e.g. in *you've*), lemma: *have*; token: *'re* (e.g. in *you're*), lemma: *be*, token: *n't* (e.g. in *don't*), lemma: *not* etc.[6] The tokens *ta* (in *gotta*) and *na* (in *gonna*, *wanna)* are assigned the lemma *to*.

4.3 Interjections

The lemma for **interjections** (tagged with UH) is identical with the token itself, e.g. token: *yeah*, lemma: *yeah* (not e.g. *yes*).

4.4 Nouns

The lemmata for **nouns** are their respective singular forms, e.g. token: *languages*, lemma: *language*, the lemma for adjectives their positive form, e.g. token: *bigger,* lemma*: big*.

Pluralia tantum are not reduced to a non-existent singular form, e.g. token: *trousers,* lemma: *trousers*.

For all items in the closed list of **compound nouns for VOICE** (cf. 6.4), each part of the compound receives a separate lemma, e.g. *business cards*: token: *business, cards*, lemmata: *business, card*. All parts of these compound nouns are considered to function as a noun unit. Forms which are part of a noun compound but do not have a noun form, are not reduced to their base forms but lemmatized in their inflected form, although in other co-texts they might belong to another lemma, e.g. token: *added value*, lemmata: *added, value (not: add, value)*

4.5 Numbers

Ordinal numbers, e.g. *seventh* are lemmatized as such and not reduced to their cardinal form, e.g. token: *seventh,* lemma: *seventh*.

[6] Exception for genitive –'s: lemma = 's

4.6 Pronouns

Pronouns, i.e. objective (e.g. *me, you, him*, ...), reflexive (e.g. *myself, yourself, himself*, ...) and possessive (e.g. mine, yours, his), as well as possessive determiners (*my, your, his*, ...) are assigned their nominative forms as lemma, e.g. token: *your*, lemma: *you*.

4.7 Quantifiers

For the **quantifiers** *much, many, more, most, less, lesser* and *least*, the lemmata are identical with their respective forms (no reduction to a positive form), e.g. token*: least*, lemma*: least* (not: *little* or *less*)

4.8 Verbs

For **verbs** the lemma is identical with the base form. For example, the tokens *go, goes, going, went, gone* and *gon* (in *gonna*) constitute a single inflectional paradigm and are all assigned the lemma *go*.

Modal verbs are lemmatized according to the same principle as verbs, i.e. the lemma is the respective base form. Examples: token: *can*, *ca* (in *can't*), lemma: *can*; token: *shall*, *should*, lemma: *shall*; token: *will, would, wo('nt)*, lemma: *will*.

5 Sources

Baayen, R. Harald; Piepenbrock, Richard; Gulikers, Leon. 1995. "The CELEX Lexical Database (CD-ROM)". Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Biber, Douglas (ed.). 1999. *Longman grammar of spoken and written English*. (1st edition). Harlow: Longman.

Brants, Thorsten. 2000. "TnT - A Statistical Part-of-Speech Tagger". In. *Proceedings of the Sixth Applied Natural Language Processing Conference*. Seattle, WA, 224-231.

Breiteneder, Angelika; Pitzl, Marie-Luise; Majewski,Stefan; Theresa Klimpfinger. 2006. "VOICE recording - Methodological challenges in the compilation of a corpus of spoken ELF". *Nordic Journal of English Studies* 5(2), 161–188. http://hdl.handle.net/2077/3153 (16 April 2014)

Brill, Eric. 1995. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging". *Computational Linguistics* 21(4), 543-565.

Brill, Eric; Wu, Jun. 1998. "Classifier Combination for Improved Lexical Disambiguation". In. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Volume 1. Association for Computational Linguistics. Stroudsburg, PA (1), 191-195.

Carter, Ronald; McCarthy, Michael. 2006. *Cambridge Grammar of English. CD-Rom. A comprehensive guide to spoken and written English usage*. Cambridge: CUP.

Daelemans, Walter; Zavrel, Jakub; Berck, Peter; Gillis, Steven. 1996. "MBT: A memory-based part of speech tagger generator". In. *Proceedings of the Fourth Workshop on Very Large Corpora*. Copenhagen, 14-27.

Daume III, Hal; Kumar, Abhishek; Saha, Avishek. 2010. "Frustratingly easy semi-supervised domain adaptation". In. *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*. ACL 2010. Uppsala, 15 July 2010, 53-59.

Davies, Mark. n.d. "Word frequency lists and dictionary from the Corpus of Contemporary American English". http://www.wordfrequency.info/free.asp (29 November 2012).

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Giménez, Jesús; Marquez, Lluis. 2003. "Fast and accurate part-of-speech tagging: The SVM approach revisited". In Nicolov, Nicolas (ed.). *Recent Advantages in Natural Language Processing III: Selected papers from RANLP 2003*. Samokov, Bulgaria. Amsterdam: Benjamins, 153-163.

Hornby, Albert Sydney; Wehmeier, Sally; McIntosh, Colin; Turnbull, Joanna; Ashby, Michael. 2007. *Oxford advanced learner's dictionary*. (7th edition). Oxford: Oxford University Press.

Kilgarriff, Adam. 2006. "BNC database and word frequency lists". http://www.kilgarriff.co.uk/bnc-readme.html (29 November 2012).

Lafferty, John; McCallum, Andrew; Pereira, Fernando. 2001. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In. *Proceedings of International Conference in Machine Learning (ICML-01)*. Williamstown, MA, 282-289.

Lager, Torbjörn. 2001. "Transformation-Based Learning of Rules for Constraint Grammar Tagging". In. *Proceedings of the 13th Nordic Conference in Computational Linguistics*. Uppsala.

Linguistic Data Consortium (LDC). 1999. "Addendum to the part-of-speech tagging guidelines for the Penn Treebank project (Modifications for the SwitchBoard corpus)". http://www.cis.upenn.edu/~bies/manuals/tagguid2.pdf (18 October 2012).

Marcus, M.P; Marcinkiewicz, M.A; Santorini, B. 1993. "Building a large annotated corpus of English: The Penn Treebank". *Computational Linguistics* 19(2), 313-330.

Miller, George A. 1995. "WordNet: A Lexical Database for English". *Communications of the ACM* 38(11), 39-41.

Nelson, Gerald. 2005. The ICE Tagging Manual. Revised version. http://ice-corpora.net/ice/taggingmanual.doc (16 April 2014)

Osimk-Teasdale, Ruth.2013. "Applying existing tagging practices to VOICE". In Mukherjee, Joybrato; Huber, Magnus (eds.). *Corpus linguistics and variation in English: Focus on Nonnative Englishes (Proceedings of ICAME 31)*. Helsinki: VARIENG.

Osimk-Teasdale, Ruth. in prep. *Parts of speech in English as a lingua franca: the POS tagging of VOICE.* PhD Thesis, University of Vienna.

Pitzl, Marie-Luise; Breiteneder, Angelika; Klimpfinger, Theresa. 2008. "A world of words: processes of lexical innovation in VOICE". *Views* 17, 21-46.

Quirk, Randolph (ed.). 1997. *A comprehensive grammar of the English language*. (14th edition). London [u.a.]: Longman.

Radeka, Michael. 2009. *Paralleles transformationsbasiertes Lernen: Kombination von Regelmengen mit einem korpusbasierten Selektionsverfahren*. Magisterarbeit, Ruprecht-Karls-Universität Heidelberg.

Radeka, Michael. in prep. *Tagging VOICE: A parallel stacked TBL approach.* PhD Thesis, University of Vienna.

Ratnaparkhi, Adwait. 1996. "A maximum entropy part-of-speech tagger". In. *Proceedings of the First Conference on Empirical Methods in NLP*. Philadelphia, PA, 133-142.

Santorini, Beatrice. 1991. "Part of Speech Tagging Guidelines for the Penn Treebank Project". http://www.personal.psu.edu/xxl13/teaching/sp07/apling597e/resources/Tagset.pdf (18 October 2012).

Santorini, Beatrice. 1995. "Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd Printing)". http://www.ling.helsinki.fi/kit/2010s/clt236/docs/PennTaggingGuide.pdf (3 December 2012).

Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In. *Proceedings of the International Conference on New Methods in Language Processing.* Manchester, UK, 44-49.

Seidlhofer, Barbara. 2011. *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Shen, Libin; Satta, Giorgio; Joshi, Aravind K. 2007. "Guided learning for bidirectional sequence classification". In. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*. Prague, 760-767.

Toutanova, Kristina; Klein, Dan; Manning, Christopher; Singer, Yoram. 2003. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network". In. *Proceedings of HLT-NAACL 2003*. Edmonton, Canada, 252-259.

van Halteren, Hans; Daelemans, Walter; Zavrel, Jakub. 2001. "Improving accuracy in word class tagging through the combination of machine learning systems". *Computational Linguistics* 27/2, 199-229.

VOICE Project. 2007a. "Spelling conventions". http://www.univie.ac.at/voice/documents/VOICE_spelling_conventions_v2-1.pdf (19 May 2011).

VOICE Project. 2007b. "VOICE Transcription Conventions [2.1]". http://www.univie.ac.at/voice/voice.php?page=transcription_general_information (6 December 2012).

VOICE Project. 2007c. "Mark-up conventions. VOICE Transcription Conventions [2.1]". http://www.univie.ac.at/voice/documents/VOICE_mark-up_conventions_v2-1.pdf (18 July 2011).

Volk, Martin; Schneider, Gerold. 1998. "Adding Manual Constraints and Lexical Look-up to a Brill-Tagger for German". In. *Proceedings of the ESSLLI-Workshop on Recent Advances in Corpus Annotation*. Saarbrücken.

Wu, Dekan; Ngai, Grace; Carpuat, Marine. 2004. "Raising the Bar: Stacked Conservative Error Correction Beyond Boosting". In. *Proceedings of the fourth International Conference on Language Resources and Evaluation (LREC-2004)*. Lisbon, 21-24.

6 Appendix

6.1 VOICE List of Formulaic Items

DESCRIPTION: Contains formulaic expressions which are in the closed list of **VOICE Formulaic Items**, including greetings, farewells, please, thanks, apologies, wishes and miscellaneous expressions. It is based on the exclamations listed in OALD7 and the categorization of formulaic expressions in ICE (Nelson 2005: 13).

TAGGING: **Tag:** FI

The list of formulaic items generally only contains very short formulaic chunks. In slightly longer, syntactically analysable stretches, the syntactic co-text was not tagged with FI but with the 'conventional' POS tag. e.g. *thank_FI you_FI very_RB much_RB*.

greetings & farewells
bye
bye-bye
ciao
good afternoon
good day
good evening
good morning
goodbye
goodnight
hello
hey
hi
see you
welcome

please & thanks
thanks
thank you
please
you're welcome

apologies
pardon
pardon me
sorry
excuse me

expletives
christ
damn
dammit
dear
dude
fuck
gee
gosh
heck
jesus
(my) god
(my) goodness
shit
shoot
boy
man

wishes
congratulation(s)
happy birthday
happy ramadan
merry christmas

miscellaneous expressions
attention
bingo
bravo
cheers (not used to mean '*thanks*' in VOICE)
encore
viva

6.2 VOICE List of Multi-words

DESCRIPTION: Included in this **closed list** are the most frequent multi-word chunks from the word classes **Adverb, Adjective, Conjunction** and **Preposition** in VOICE. This list is based on Multi-word items in the OALD7 (reference dictionary), those used in the BNC and those which appear on the VOICE bi- and trigrams. Also included are foreign Multi-word items, mostly of Latin or Greek origin (e.g. *ad hoc*).

TAGGING: Each part of the multi-word item is assigned the **same tag**, e.g. *a_RB lot_RB*
NB: Many of the items listed here have counterparts which do not function as a multi-word and receive 'conventional' POS tags. These cases have been disambiguated, e.g. *so_IN that_IN when we go back* vs. *so_DM that_DT 's why*; *kind_RB of_RB good* vs. *that kind_NN of_DT law*.

a) ADVERBS (Tags: RB)

a bit
a cappella
a little
a little bit
a lot
a priori
ad hoc
ad nauseam
all right
any more
as well
at all
at least
de facto
et cetera
for example
for instance
for sure
in general
kind of
more or less
of course
over there
per capita
per cent
per se
sort of
sui generis
vice versa

b) ADJECTIVES (Tags: JJ)

a cappella
a priori
ad hoc
ad nauseam
all right
brand new
de facto
far eastern
fed up
fully fledged
in vitro
middle eastern
new age
next door
number one
out of date
per capita
per cent
per cent
politically correct
roman catholic
social democratic
sold out
sui generis
up to date
upper class
well balanced
well built
well defined
well developed
well disposed
well done
well informed
well known
well paid
well used
worked up

c) CONJUNCTIONS (Tags: IN)

even if
even though
now that
so that

d) PREPOSITIONS (Tags: IN)

according to
because of
depending on
in order to
in terms of
instead of
next to
out of
such as
vis-à-vis

6.3 VOICE List of Discourse Markers and Interjections

6.3.1 Single word discourse markers

Items: *like, look, whatever, well, so, right*
Tag: DM

6.3.2 Multi-word discourse markers

Items: *I mean, I see, mind you, you know, you see*
Tags: Multi-word discourse markers are tagged with a conventional tag for form, and the tag DM for function, for all parts of the discourse marker:

I_PP(DM) mean_VVP(DM)
I_PP(DM) see_VVP(DM)
mind_VVP(DM) you_PP(DM)
you_PP(DM) know_VVP(DM)
you_PP(DM) see_VVP(DM)

6.3.3 Interjections

DESCRIPTION: These are the items listed as discourse markers in the VOICE Mark-up conventions (VOICE Project 2007c: 4). They do not have a homonym in a different word class category but fulfil the following discourse functions. The items in green have been added for VOICE POS.

TAGGING: **Tag:** UH (NB: Non-English discourse markers are tagged **FW**.)

Interjection	Function
er, erm	**Hesitation**/filler
huh	tag-question
yay, yipee, whoohoo, mm:	**Exclamations** joy/enthusiasm
haeh	questioning/doubt/disbelief
a:h, o:h, wow, poah	astonishment/surprise
oops	apology
ooph	exhaustion
ts, pf	disregard/dismissal/contempt
ouch, ow	pain
sh, psh	requesting silence
oh-oh:, u:h	anticipating trouble
ur, yuck	disapproval/disgust
oow	pity/disappointment
blah	**expressing predictability or lack of interest for something**

6.4 VOICE List of Compound Nouns

DESCRIPTION: The VOICE list of Compound Nouns is a closed list, consisting of 1) items listed as multiword noun units in our **reference dictionary**, e.g. *public service*, but also e.g. *master of ceremony* and 2) multiword noun units occurring **most frequently** (30 times or more) in the **n-gram lists** for our data, e.g. *joint program*. This list includes all 359 noun combinations tagged as compound nouns in VOICE, however, not including those where the first word ends in the morpheme *–ing* (e.g. *swimming pool*) (see 3.4.1.3.)

TAGGING: Format for **singular** compound nouns: e.g. *academic_NN year_NN*
Format for **plural** compound nouns: e.g. *business_NNS cards_NNS*

academic year
added value
adult education
affirmative action
age group
age limit
alarm clock
amusement park
armed forces
art gallery
art history
art nouveau
artificial intelligence
artificial language
assistant professor
associate professor
au pair
auxiliary language
back door
balance sheet
bank holiday
banoffi pie
bar code
best practice
big bang
big toe
birth rate
black box
black market
black sheep
blister pack
blood pressure
body language
bonded warehouse
bonfire night
bottom line
brain drain
brand name
bullet point
bus stop
business administration
business card
business school
calendar year
capital city
carbon copy
cash and carry
cash flow
catchment area
central bank
central government
chain reaction
checks and balances
chief executive
chip card
christmas tree
civil rights
civil servant
civil war
clean up
coat of arms
code of practice
coffee break
coffee shop
common denominator
common ground
common law
common market
common room
common sense
community service
computer science
conceptual art
condensed milk
consumer goods
contact person
continental shelf
convenience store
court of appeal
credit card
critical mass
culture shock
current account
dance floor
day off
day out
dead end
department store
differential equation
digestive system
direct action
double agent
double room
dress code
dress rehearsal
due date

duty-free shop
end product
end result
education system
exchange rate
exclamation mark
extreme sports
fairy tale
family name
feta cheese
field trip
financial aid
fire alarm
fire station
first floor
first language
first name
fish and chips
flat rate
flip chart
flow chart
focal point
focus group
frame of reference
free trade
free will
front desk
front page
front runner
full professor
further education
general assembly
general knowledge
general practice
general public
generation x
giant slalom
global village
golden retriever
golden rule
good faith
good practice
good sense
grand master
grass roots
green light
grocery store
gross domestic product
ground floor
ground rule
group work
half day
hard copy
head office
health care
heart attack
high five
high heels
high school
high season
high water
higher education
home economics
home page
human nature
human resources
ice cream
income support
information technology
intelligence test
internal market
inverted commas
irish coffee
ivory tower
jet lag
job description
joint degree
joint master
joint program
joint venture
labor force
labor market
last name
lead time
legal action
legal tender
letter of intent
life expectancy
life insurance
lingua franca
local government
loose end
low point
low tide
lower house
lump sum
main street
market share
master of ceremonies
master plan
maternity leave
media studies
member state
membership criterion
middle ages
middle class
middle name
middle school
military service
mineral water
minimum wage
minority government
mission statement
mobile phone
money-back guarantee
monopoly money
mother tongue
movie theater
mutual recognition
nation state
national anthem
national curriculum
native speaker
natural gas
natural science
news agency
normal distribution
nuclear physics
nuclear power
office hours
old age
open market

open season
order of magnitude
organic chemistry
paper cutter
peace process
pencil case
pension scheme
petrol station
phone call
phone number
point of view
point of sale
political correctness
political science
political scientist
population explosion
position paper
post office
present day
press agency
press conference
press release
press secretary
pressure cooker
price controls
price tag
price war
primary school
prime minister
private company
private law
private school
private sector
production line
public access
public opinion
public relations
public school
public sector
public service
public transport
public transportation
purchase price
quality assurance
quality control
question mark
race car
raw material
red carpet
red wine
red-light district
research and development
response time
right wing
road map
rock music
role model
round trip
sales rep
science fiction
score sheet
seat belt
second language
second name
secondary school
security council
seed money
senior citizen
service provider
short cut
short time
side street
sim card
sine qua non
ski lift
small talk
social fund
social inclusion
social science
social security
social studies
social worker
split second
stainless steel
star sign
state university
status quo
stock exchange
stock market
student union
stream of consciousnesssuccess story
summer school
supply and demand
suspension bridge
swiss cheese
tape recorder
target language
task force
telephone number
terms of reference
three quarters
time bomb
time frame
time limit
time span
time zone
top ten
town hall
track record
trade union
trash can
travel agency
trust fund
tuition fees
upper class
vested interest
video camera
voluntary service
voluntary work
way out
web page
welfare state
white fish
white wine
wine gum
work experience
work permit
world cup
youth hostel
youth organization

Appendix G: VOICE Transcription Conventions (Mark-up and Spelling)

TRANSCRIPTION CONVENTIONS [2.1]

Mark-up conventions

VOICE
Vienna-Oxford International Corpus of English

Mark-up conventions

Version 2.1 June 2007

1. SPEAKER IDS	
S1: S2: …	Speakers are generally numbered in the order they first speak. The speaker ID is given at the beginning of each turn.
SS:	Utterances assigned to more than one speaker (e.g. an audience), spoken either in unison or staggered, are marked with a collective speaker ID **SS**.
SX:	Utterances that cannot be assigned to a particular speaker are marked **SX**.
SX-f: SX-m:	Utterances that cannot be assigned to a particular speaker, but where the gender can be identified, are marked **SX-f** or **SX-m**.
SX-1: SX-2: …	If it is likely but not certain that a particular speaker produced the utterance in question, this is marked **SX-1**, **SX-2**, etc.
2. INTONATION	
Example: S1: that's what my next er slide**?** does	Words spoken with rising intonation are followed by a question mark "**?**" .
Example: S7: that's point two**.** absolutely yes**.**	Words spoken with falling intonation are followed by a full stop "**.**" .
3. EMPHASIS	
Example: S7: er internationalization is a very IMPORTANT issue Example: S3: toMORrow we have to work on the presentation already	If a speaker gives a syllable, word or phrase particular prominence, this is written in capital letters.
4. PAUSES	
Example: SX-f: because they all give me different **(.)** different **(.)** points of view	Every brief pause in speech (up to a good half second) is marked with a full stop in parentheses.
Example: S1: aha **(2)** so finally arrival on monday evening is still valid	Longer pauses are timed to the nearest second and marked with the number of seconds in parentheses, e.g. (1) = 1 second, (3) = 3 seconds.

5. OVERLAPS	
Example: S1: it is your best **<1> case </1>** scenario (.) S2: **<1> yeah </1>** S1: okay	Whenever two or more utterances happen at the same time, the overlaps are marked with numbered tags: **<1> </1>**, **<2> </2>**,… Everything that is simultaneous gets the same number. All overlaps are marked in blue.
Example: S9: it it is (.) to identify **some<1>thing** </1> where (.) S3: <1> mhm </1>	All overlaps are approximate and words may be split up if appropriate. In this case, the tag is placed within the split-up word.
6. OTHER-CONTINUATION	
Example: S1: what up till (.) till twelve? S2: **yes=** S1: **=really**. so it's it's quite a lot of time.	Whenever a speaker continues, completes or supports another speaker's turn immediately (i.e. without a pause), this is marked by "=".
7. LENGTHENING	
Example: S1: you can run faster but they have much **mo:re** technique with the ball	Lengthened sounds are marked with a colon "**:**".
Example: S5: personally that's my opinion the: **er::m**	Exceptionally long sounds (i.e. approximating 2 seconds or more) are marked with a double colon "**::**".
8. REPETITION	
Example: S11: e:r i'd like to go t- t- **to to** this type of course	All repetitions of words and phrases (including self-interruptions and false starts) are transcribed.
9. WORD FRAGMENTS	
Example: S6: with a minimum of (.) of **participa-** S1: mhm S6: -**pation** from french universities to say we have er (.) a joint doctorate or a **joi-** joint master	With word fragments, a hyphen marks where a part of the word is missing.
10. LAUGHTER	
Example: S1: in denmark well who knows. **@@** S2: <**@**> yeah </**@**> **@@** that's right	All laughter and laughter-like sounds are transcribed with the **@** symbol, approximating syllable number (e.g. ha ha ha = @@@). Utterances spoken laughingly are put between <**@**> </**@**> tags.

11. UNCERTAIN TRANSCRIPTION	
Example: S3: i've a lot of very **(generous)** friends Example: SX-4: they will do whatever they want because they are a compan**(ies)**	Word fragments, words or phrases which cannot be reliably identified are put in parentheses **( )**.
12. PRONUNCIATION VARIATIONS & COINAGES	
Example: S4: i also: (.) e:r played (.) tennis e:r **<pvc> bices </pvc>** e:r we rent? went?	Striking variations on the levels of phonology, morphology and lexis as well as 'invented' words are marked **<pvc> </pvc>**.
Example: S9: how you were controlling such a thing and how you **<pvc> (avrivate) </pvc>** (it)	What you hear is represented in spelling according to general principles of English orthography. Uncertain transcription is put in parentheses **( )** .
Example: S6: what we try to explain here is the foreign direct investment growth (2) in a certain industry (.) and a certain **<pvc>** compy **{company} </pvc>**	If a corresponding existing word can be identified, this existing word is added between curly brackets **{ }**.
Example: S2: anyway i make you an a total (.) **<pvc>** summamary {summary} **<ipa>** sʌməˈmærɪ **</ipa> </pvc>** of destinations	Particularly when it comes to salient variations on the level of phonology, e.g. sound substitution or addition, a phonetic representation should be added between **<ipa> </ipa>** tags.
13. ONOMATOPOEIC NOISES	
Example: S1: it may be quite HARMLESS and at the end of the day you (.) **<ono>** dəʃ dəʃ dəʃ **</ono>** (.) somebody	When speakers produce noises in order to imitate something instead of using words, these onomatopoeic noises are rendered in IPA symbols between **<ono> </ono>** tags.
14. NON-ENGLISH SPEECH	
Example: S5: **<L1de> bei firmen </L1de>** or wherever	Utterances in a participant's first language (L1) are put between tags indicating the speaker's **L1**.
Example: S7: er this is **<LNde> die seite? (welche) </LNde>** is	Utterances in languages which are neither English nor the speaker's first language are marked **LN** with the language indicated.
Example: S4: it depends in in in **<LQit> roma </LQit>**	Non-English utterances where it cannot be ascertained whether the language is the speaker's first language or a foreign language are marked **LQ** with the language indicated.
Example: S2: erm we want to go t- to **<LNvi> xx xxx </LNvi>** island first of all	Unintelligible utterances in a participant's L1, LN or in an LQ are represented by **x's** approximating syllable number.
Example: S4: and now we do the boat trip (1) **<L1xx> xxxxx </L1xx>** S3: mhm	Utterances in a language one cannot recognize are marked **L1xx**, **LNxx** or **LQxx**.

Example: S3: <L1fr> oui un grand carre **{yes like a big square}** </L1fr> (.) i <fast> think it would </fast> be better if we put the tables a <soft> different way </soft>	If possible, translations into English are provided between curly brackets **{ }** immediately after the non-English speech.
15. SPELLING OUT	
Example: S1: and they (3) created some (1) some er (2) JARGON. do you know? the word JARGON? (.) **<spel> j a r- </spel> <spel> j a r g o n? </spel>** jargon	The **<spel> </spel>** tag is used to mark words or abbreviations which are spelled out by the speaker, i.e. words whose constituents are pronounced as individual letters.
16. SPEAKING MODES	
Example: S2: because as i explained before is that we have in the **<fast>** universities of cyprus we have **</fast>** a specific e:rm procedure <fast> </fast> <slow> </slow> <loud> </loud> <soft> </soft> <whispering> </whispering> <sighing> </sighing> <reading> </reading> <reading aloud> </reading aloud> <on phone> </on phone> <imitating> </imitating> <singing> </singing> <yawning> </yawning>	Utterances which are spoken in a particular mode (fast, soft, whispered, read, etc.) and are notably different from the speaker's normal speaking style are marked accordingly. The list of speaking modes is an open one.
17. BREATH	
Example: S1: so it's always **hh** (.) going around (2) yeah	Noticeable breathing in or out is represented by two or three **h's** (hh = relatively short; hhh= relatively long).
18. SPEAKER NOISES	
<coughs> <clears throat> <sniffs> <sneezes> <snorts> <applauds> <smacks lips> <yawns> <whistles> <swallows>	Noises produced by the current speaker are always transcribed. Noises produced by other speakers are only transcribed if they seem relevant (e.g. because they make speech unintelligible or influence the interaction). The list of speaker noises is an open one.
Example: S1: yeah <1> what </1> i think in in doctor levels	These noises are transcribed as part of the running text and put between pointed brackets **< >.**

S7: <1> **<clears throat>** </1>	
Example: SX-m: but you NEVER KNOW when it's popping up you never kno:w S3: **<coughs (6)>**	If it is deemed important to indicate the length of the noise (e.g. if a coughing fit disrupts the interaction), this is done by adding the number of seconds in parentheses after the descriptor.
19. NON-VERBAL FEEDBACK	
<nods> <shakes head>	Whenever information about it is available, non-verbal feedback is transcribed as part of the running text and put between pointed brackets < >.
Example: S3: but i think if you structure corporate governance appropriately you can have everything (1) S7: <soft> mhm </soft> **<nods (2)>**	If it is deemed important to indicate the length of the non-verbal feedback, this is done by adding the number of seconds in parentheses.
20. ANONYMIZATION	
	A guiding principle of VOICE is sensitivity to the appropriate extent of anonymization. As a general rule, names of people, companies, organizations, institutions, locations, etc. are replaced by aliases and these aliases are put into square brackets **[ ]**. The aliases are numbered consecutively, starting with 1.
Example: S9: that's one of the things (.) that i (1) just wanted to clear out. (2) **[S13]**? Example: S6: so: (1) ei:ther MYself or mister **[S2/last]** or even boss (.) should be there every year Example: S8: so my name is **[S8] [S8/last]** from vienna	Whenever speakers who are involved in the interaction are addressed or referred to, their names are replaced by their respective speaker IDs. A speaker's first name is represented by the plain speaker ID in square brackets **[S1]**, etc. A speaker's last name is marked **[S1/last]**, etc. If a speaker's full name is pronounced, the two tags are combined to **[S1] [S1/last]**, etc.
Example: S2: that division is headed by (1) **[first name3] [last name3]** (1)	Names of people who are not part of the ongoing interaction are substituted by **[first name1]**, etc. or **[last name1]**, etc. or a combination of both.
Example: S5: erm she is currently head of marketing (and) with the **[org2]** (1)	Companies and other organizations need to be anonymized as well. Their names are replaced by **[org1]**, etc.
Example: S1: i: i really don't wanna have a: a joint degree e:r with the university of **[place12]** (.)	Names of places, cities, countries, etc. are anonymized when this is deemed relevant in order to protect the speakers' identities and their environment. They are replaced by **[place1]**, etc.

Example: S8: he get the <L1cs> diplom {diploma} </L1cs> of **[name1]** university (.) and french university can give him also the <L1cs> diplom {diploma} </L1cs>	Other names or descriptors may be anonymized by **[name1]**, etc., as in e.g. Charles University.
Example: S3: erm i- in the **[thing1]** is very well explained. so <2> i can </2> pa- <3> er pass you this </3> th- the definitions. S4: <2> aha </2> S4: <3> okay <@> okay </@> </3>	Products or other objects may be anonymized by **[thing1]**, etc.
21. CONTEXTUAL EVENTS	
{mobile rings} {S7 enters room} {S2 points at S5} {S4 starts writing on blackboard} {S4 stops writing on blackboard} {S2 gets up and walks to blackboard (7)} {S3 pours coffee (3)} {SS reading quietly (30)} … Example: S3: one dollar you get (.) (at) one euro you get one dollar twenty-seven. (.) S4: right. **{S5 gets up to pour some drinks}** S3: right now at this time (3) S1: er page five is the er (4) **{S5 places some cups and glasses on the desk (4)}** S1: i think is the descritip- e:r part of what i have just explained (.)	Contextual information is added between curly brackets **{ }** only if it is relevant to the understanding of the interaction or to the interaction as such. If it is deemed important to indicate the length of the event, this can be done by adding the number of seconds in parentheses. Explanation: The pause in the conversation occurs because of the contextual event.
22. PARALLEL CONVERSATIONS	
Example: S1: four billion <spel> u s </spel> dollars. (.) S4: quite impressive (.) S1: er **<to S2>** not quite isn't it **</to S2>** (.) i understand some other countries we handle	To indicate that a speaker is addressing not the whole group but one speaker in particular, the stretch of speech is marked with (e.g.) **<to S1> </to S1>**, choosing the speaker ID of the addressee.
Example: S7: i've i've found the people very stressed SS: @@@ S7: that's (.) i don't know how many of you study here but it's VERY important to push the close the door button in that elevator. this is something i've never <3> seen in sweden </3> **{parallel conversation between S1 and S3 starts}** or anywhere else <4> but it's very	Wherever two or more conversational threads emerge which are too difficult to transcribe, as a general rule only the main thread of conversation is transcribed. The threads which are not transcribed are treated like a contextual event and indicated between curly brackets **{ }**.

important to push this button </4> SS: <3> @@@@ </3> SS: <4> @@@@@@@@@ </4> @@ S7: <5> i never even saw this button in another el- elevator </5> SS: <5> @@@@@@@@@@@ </5> **{parallel conversation between S1 and S3 ends}** @@@	
23. UNINTELLIGIBLE SPEECH	
Example: S4: we **<un> xxx </un>** for the <7> supreme (.) three </7> possibilities S1: <7> next yeah </7>	Unintelligible speech is represented by **x's** approximating syllable number and placed between **<un> </un>** tags.
Example: S7: obviously the the PROCESS will **<un>** x **<ipa>** θeŋ **</ipa> </un>** (.) w- w- will (.) will take (.) at least de- decade	If it is possible to make out some of the sounds uttered, a phonetic transcription of the x's is added between **<ipa> </ipa>** tags.
24. TRANSCRIPTION BORDERS	
<**beg** CD1_4_00:35>	The **beg**inning of the transcript is noted by indicating the CD number, the track number and the exact position of the respective track in minutes and seconds.
<**end** CD1_21_01:27>	The **end** of the transcript is noted in the same way.
<end CD1_19_01:27> **(gap 00:06:36) {multiple parallel conversations, hardly intelligible}** <beg CD1_21_02:03>	A **gap** in the **transcription** is indicated in parentheses, including its length in hh:mm:ss. Curly brackets **{ }** are used in order to specify the reasons for or the circumstances of the gap.
<end CD1_24_3:02> **(nrec 00:00:45) {change of minidisk}** <beg CD2_1_00:00>	An **interruption** in the **recording** is indicated in the same way, but abbreviated as "**nrec**" (i.e. non-recorded). The length you indicate will normally be a guess.

In addition to the regular mark-up, transcribers supplement the transcripts with Transcriber's Notes in which they provide additional contextual information and observations about other features of the interaction not accounted for in the transcript.

For a detailed discussion of specific aspects of the transcription conventions cf. Breiteneder, Pitzl, Majewski, Klimpfinger. (2006). "VOICE recording – Methodological challenges in the compilation of a corpus of spoken ELF". *Nordic Journal of English Studies*, 5/2, 161-188.

TRANSCRIPTION CONVENTIONS [2.1]

Spelling conventions

VOICE
Vienna Oxford International Corpus of English

Spelling conventions

Version 2.1 June 2007

1. CHARACTERS	
a b c d e f g h i j k l m n o p q r s t u v w x y z	Only alphabetic **Roman characters** are used in the transcript. No diacritics, umlauts or non-roman characters are permitted in the running text.
2. DECAPITALIZATION	
Example: S8: so you really can <@> control my **e**nglish </@>	**No capital letters** are used except for marking emphasis (cf. mark-up conventions).
3. BRITISH SPELLING	
British spelling	British English spelling is used to represent naturally occurring ELF speech. The Oxford Advanced Learner's Dictionary (**OALD**), 7th **edition**, is used as the primary source of reference. If an entry gives more than one spelling variant of a word, the first variant is chosen. If there are two separate entries for British and American spelling, the British entry is selected.
4. SPELLING EXCEPTIONS	
cent**er**, theat**er** behavi**or**, col**or**, fav**or**, lab**or**, neighb**or** defens**e**, offens**e** dis**k** progra**m** travel (-l-: trave**l**ed, trave**l**er, trave**l**ing)	The 12 words listed on the left **and all their derivatives** are spelled according to American English conventions (e.g. colors, colorful, colored, to color, favorite, favorable, to favor, in favor of, etc.).
Example: S2: we are NOT quite sure if it will REALLY be (.) privati**z**ed next year	In addition, **all words** which can be spelled using either an *-is* or an ***-iz*** morpheme are spelled with ***-iz*** (e.g. to emphasize, organizations, realization, recognized, etc.).
5. NON-ENGLISH WORDS	
Example: S1: <L1de> wieso **oe**sterreich? {why austria} </L1de> Example: S3: <LNfr> c'est ferm**e**? {is it closed} </LNfr>	Non-English words are rendered in the standard variant of the original language (i.e. no non-standard dialect). The roman alphabet is always used, also in the case of languages like Arabic or Japanese. **No umlauts** (e.g. NOT österreich), **no diacritics** (e.g. NOT fermé) and **no non-roman characters** are permitted.

6. FULL REPRESENTATION OF WORDS	
Example: S7: the students that (.) decide freely to enter (.) this kind of master knows (.) for example that **he** can (.) at the end achieve (.) sixty credits	Although words may not be fully pronounced or may be pronounced with a foreign accent, they are generally represented in standard orthographic form. Explanation: S7 is Italian and pronounces the *he* in *he can* as /ɪ/, swallowing the initial h. Nevertheless, this is regarded as a minor instance of L1 accent and therefore represented in standard orthography (*he*).
7. FULL REPRESENTATION OF NUMBERS, TITLES & ABBREVIATIONS	
oh/zero, two, three, … one hundred, nineteen ten, eighteen twenty-seven, …	Numbers are fully spelled out as whole words. British English hyphenation rules apply.
missis (for *Mrs*), mister, miss, mis (for *Ms*), doctor, professor, …	Titles and terms of address are fully spelled out.
et cetera, saint thomas, okay,…	Forms that are usually abbreviated in writing, but spoken as complete words are fully spelled out.
8. LEXICALIZED REDUCED FORMS	
cos gonna, gotta, wanna	Lexicalized phonological reductions are limited to the four on the left. All other non-standard forms are fully spelled out (e.g. /hæftə/ = *have to*).
9. CONTRACTIONS	
i'm, there're, how's peter, running's fun, … i've, they've, it's got, we'd been, … tom'll be there, he'd go for the first, … we aren't, i won't, he doesn't, … what's it mean, where's she live, how's that sound … let's	Whenever they are uttered, all standard contractions are rendered. This refers to verb contractions with *be* (*am*, *is* *are*), *have* (*have*, *has*, *had*), *will* and *would* as well as *not*-contractions. Additionally, *'s* is used to represent *does* when reduced and attached to a *wh*-word. It is also used to represent the pronoun *us* in the contracted form *let's*.
10. HYPHENS	
Example: S3: more than thirteen years of experience er working in (.) er (.) design and development (.) er of (1) **real-time** software (.) er for industrial (.) implications	Hyphens are used according to British English hyphenation rules. The OALD, 7th edition, is used as the primary source of reference.

Example: S2: we would allow that within er an international **cooperation** (.)	If an entry gives more than one spelling variant of a word, the first variant is chosen.
11. ACRONYMS	
Example: S10: for the development of joint programmes within the **unica** networks.	Acronyms (i.e. abbreviations spoken as one word) are transcribed like words. They are not highlighted in any way.
12. DISCOURSE MARKERS	
	All discourse markers are represented in orthography as shown below. The lists provided are closed lists. The items in the lists are standardized and may not represent the exact sound patterns of the actual discourse markers uttered.
yes, yeah, yah okay, okey-dokey mhm, hm aha, uhu	**Backchannels and positive minimal feedback** (closed sound-acknowledgement token) (open sound-acknowledgement token)
no n-n, uh-uh	**Negative minimal feedback**
er, erm	**Hesitation**/filler
huh	tag-question
yay, yipee, whoohoo, mm:	**Exclamations** joy/enthusiasm
haeh	questioning/doubt/disbelief
a:h, o:h, wow, poah	astonishment/surprise
oops	apology
ooph	exhaustion
ts, pf	disregard/dismissal/contempt
ouch, ow	pain
sh, psh	requesting silence
oh-oh:, u:h	anticipating trouble
ur	disapproval/disgust
oow	pity, disappointment
Example: S3: <L1ja> **he:** </L1ja> Example: SX-m: <L1de> **ach ja {oh yes}** </L1de>	What are clearly L1-specific discourse markers are marked as foreign words. Due to the wide range of these phenomena in different languages, the L1-list is open-ended. A translation is added whenever this is possible.

For a detailed discussion of specific aspects of the transcription conventions cf. Breiteneder, Pitzl, Majewski, Klimpfinger. (2006). "VOICE recording – Methodological challenges in the compilation of a corpus of spoken ELF". *Nordic Journal of English Studies*, 5/2, 161-188.

Appendix H: English summary

This PhD thesis addresses the conceptual and practical challenges involved in the part-of-speech tagging (POS tagging) of the Vienna-Oxford International Corpus of English (VOICE). Part-of-speech tagging is the assignment of word class categories to tokens of a corpus. There has not been an attempt ever before to POS tag naturally occurring, transcribed spoken interactions of English as a lingua franca (ELF) as in VOICE, which is not only characterised by features of spoken language but also by heightened variability and plurilingual language use.

In chapters 1 and 2, a general introduction to relevant issues to this thesis, namely the categorisation of language and its challenges, the most common part-of-speech tagging methods, as well as ELF research and VOICE is given. Chapter 3 provides a review of previous approaches to POS tagging spoken language. It is shown that these do not sufficiently address the issues involved in tagging VOICE. Rather, existing tagging methods tend to gloss over issues which are relevant to dynamic spoken discourse and, hence, needed to be extended and, in some cases, discarded entirely, for the tagging of VOICE. The issues which arose in the POS tagging of VOICE and the, sometimes unpreceded, ways of handling of these are discussed in chapter 4, which concludes with the main guidelines developed for tagging VOICE.

Chapter 5 then illustrates the advantages of this unique approach to POS tagging by applying it to the specific issue of word class variation, as in the POS tagging of VOICE, many cases were encountered where ELF users go beyond the commonly assumed boundaries of word class categories in Standard English. An investigation of the most frequent types of word class variation in VOICE in a case study with regard to the forms which are shifted, the directionality of these shifts, as well as the environment in which the shifts occur shows that the word class shifts in VOICE interactions follow clear tendencies. This, it is argued, calls for a reconceptualization of the conventional concepts of conversion and multifunctionality with regard to ELF. Moreover, it is demonstrated that the tagging scheme developed for VOICE can yield insights into the nature of variable, spoken language data by investigating those cases in which the ELF users in VOICE go beyond conventional word class boundaries.

In chapter 6 is argued that the strength of the ‘linguistic’ approach chosen for tagging VOICE is that it views the annotation of such data as a discovery procedure and welcomes rather than glosses over, any challenging issues. As such, it brings out clearly complex aspects of spoken, plurilingual language in use and facilitates their investigation. Moreover, it links the practical

task of POS tagging to conceptual issues regarding the categorisation of spoken discourse and can serve as a starting point for the investigation of similar types of data.

Appendix I: Deutsche Zusammenfassung (German summary)

Diese Arbeit befasst sich mit den konzeptionellen sowie praktischen Fragestellungen welche sich aus der Wortklassenannotation (POS Tagging) des Vienna-Oxford International Corpus of English (VOICE) ergeben. Es gab bis dahin keine Versuche, Transkripte von unter natürlichen Bedingungen aufgenommenen Englisch als Lingua Franca (ELF)-Interaktionen mit POS Tagging zu annotieren. Solche Interaktionen beinhalten nicht nur typische Charakteristika gesprochener Sprache sondern auch vermehrt variablen sowie plurilingualen Sprachgebrauch.

Kapitel 1 und 2 dienen einer allgemeinen Einführung in für diese Arbeit relevante Themengebiete: Es wird einen Überblick zu Themen und Fragestellungen bezüglich der Kategorisierung von Sprachdaten und zu Methoden des POS Taggings gegeben. Darüber hinaus werden die wichtigsten Eckpunkte aktueller ELF Forschung und des Korpus VOICE besprochen. Darauf folgt eine Rezension der wichtigsten Literatur zu POS Tagging von gesprochener Sprache in Kapitel 3. Es wird gezeigt, dass bisherige Taggingmethoden die für das Tagging von VOICE relevanten Themen nicht ausreichend behandeln, sondern eher dazu tendieren, diese auszuklammern. Solche Themen betreffen beispielsweise typische Merkmale von dynamischem, gesprochenem Diskurs. Daher ergibt sich die Notwendigkeit, gängige Methoden für VOICE zu adaptieren oder, in manchen Fällen, neue, unkonventionelle Lösungen zu finden.

Die Fragestellungen, welche sich aus dem Tagging von VOICE ergaben und die dafür entwickelten Lösungen werden in Kapitel 4 erläutert. Dieses schließt mit den Leitlinien, die aus dem Tagging von VOICE resultierten. Am konkreten Beispiel der Wortklassenvariation (oder Wortklassenverschiebung) wird in Kapitel 5 schließlich der Nutzen dieses Taggingansatzes demonstriert.

In diesem Kapitel werden jene Fälle untersucht, in denen ELF SprecherInnen allgemein angenommene Grenzen konventioneller Wortklassen des Standard Englisch überschreiten, da dies ein prominentes Thema beim POS Tagging von VOICE darstellte. In einer Fallstudie werden die häufigsten Typen von Wortklassenvariation in VOICE bezüglich der Formen, welche variiert werden, der Richtung der Verschiebungen, sowie des Kontexts in welchem diese Variationen vorkommen, beleuchtet. Es wird gezeigt, dass die Wortklassenverschiebungen in VOICE-Interaktionen klaren Tendenzen folgen. Dies zeigt, dass eine Rekonzeptionalisierung konventioneller Konzepte (Konversion und Multifunktionalität) zur Erklärung von Wortklassenvariation im Bezug auf ELF nötig ist.

Darüber hinaus wird gezeigt, dass das für VOICE entwickelte Taggingschema Einblicke in die Eigenschaften von variabler, gesprochener Sprache gewähren kann.

Es wird argumentiert, dass die Stärke eines ‚linguistischen' Ansatzes welcher für das Tagging von VOICE gewählt wurde, darin liegt, dass der Vorgang einer Daten-Annotation als Entdeckungsreise betrachtet wird, auf der vermeintliche ‚Hindernisse' als untersuchenswerte Herausforderungen gesehen werden, die weiter verfolgt und nicht übergangen werden. Dadurch werden komplexe Aspekte von gesprochenem, plurilingualem Sprachgebrauch sichtbar und können untersucht werden. Darüber hinaus bringt dieser Ansatz die praktische Aufgabe des POS Taggings mit konzeptionellen Fragestellungen bezüglich der linguistischen Kategorisierung von gesprochenem Diskurs in Verbindung, und kann als Ausgangspunkt für ähnliche Daten dienen.

Appendix J: Curriculum Vitae (academic)

PERSONAL INFORMATION

Ruth Osimk-Teasdale
Date of birth: 26 July 1982
Nationality: Austrian

EDUCATION

2009 – ongoing **University of Vienna, Austria**
Doctoral studies, English Department. Doctoral thesis entitled „*Parts of speech in English as a lingua franca: the POS tagging of VOICE”*.
Teaching Degree for English and Philosophy and Psychology.

2008 **University of Cambridge, ESOL Examinations**
Certificate in English Language Teaching to Adults (CELTA)

2001-2007 **University of Vienna, Austria**
Studies of General and Applied Linguistics, with modules English language and linguistics (48 weekly hours) and German as a foreign language (24 weekly hours). Successful completion in 2007 (Degree: Mag. phil.).

1988 – 2000 **Volksschule** (Primary school) and **Neusprachliches Gymnasium** (Secondary School with focus on modern languages), Vienna.

ACADEMIC POSITIONS

06/2009-08/2013 **VOICE Project, Department of English, University of Vienna** (**see** http://www.univie.ac.at/voice/)
Fulltime researcher within the FWF-TRP Project Englisch als internationale Lingua Franca.

06/2005-06/2009 **VOICE Project, Department of English, University of Vienna**
Freelance work with the project, e.g. data collection for VOICE, transcription and checking of ELF conversations according to the VOICE Transcription Conventions.

03-06/2007 **Department of English, University of Vienna**
Pronunciation Lab Tutor for **PPOCS** (Practical Phonetics and Oral Communication Skills).

LIST OF ACADEMIC PUBLICATIONS AND PRESENTATIONS

Monographs and publications in academic journals

in prep *Parts of speech in English as a lingua franca: the POS tagging of VOICE*. PhD Thesis, University of Vienna.

forthc. “Categorising the unconventional: Potential and limits of existing points of reference for POS-tagging VOICE” *International Journal of Corpus Linguistics.* (with N. Dorn)

2014 "'I just wanted to give a partly answer': Capturing and exploring word class variation in ELF data". *Journal of English as a Lingua Franca* 3/1, 109–143.

2013 "Applying existing tagging practices to VOICE". Mukherjee, Joybrato; Huber, Magnus (eds.). *Corpus linguistics and variation in English: Focus on Nonnative Englishes (Proceedings of ICAME 31)*. Helsinki: VARIENG.

2010 "Testing the intelligibility of ELF sounds". *Speak Out!* 42, 14-18.

2009 "Decoding sounds: an experimental approach to intelligibility in ELF." *Vienna English Working PaperS* 18/1, 64-89

Verständlichkeit in Englisch als Lingua Franca. Die Rolle von Aspiration, [θ]/[ð] und /r/. Saarbrücken: VDM-Verlag Müller.

2007 *Aspiration, [θ]/[ð] und /r/ in Englisch als Lingua Franca – eine psycholinguistische Pilotstudie zu drei Vorschlägen des Lingua Franca Core*. MA thesis, University of Vienna.

Presentations

2014 'POS-tagging the Vienna-Oxford International Corpus of English: strategies an rewards'. Individual Paper, Olomouc Linguistics Colloquium, Olomouc, 5-7 June 2014. (with N. Dorn and C. Schekulin)

2013 'Categorising plurilingual user data? Challenges and solutions for POS-tagging VOICE'. Individual paper, pre-conference workshop on 'Compilation and Annotation of Spoken Corpora: Towards Best Practice', convened by Gisle Andersen, John Kirk and Susan Lee Nacey. ICAME34, Santiago de Compostela, 22 May 2013.

2012 '"You can actually perform *good* in another language if you put your mind to it": Conversion in ELF'. Individual Paper, 5th International Conference of English as a Lingua Franca, Istanbul, 24-26 May 2012.

'Vienna-Oxford International Corpus of English (VOICE)'. Invited paper, CLARINAT - DARIAH-AT. Europäische Forschungsinfrastrukturen in den Geisteswissenschaften, Vienna, 21-22 February 2012. (with B. Seidlhofer, M. Radeka and S. Majewski)

2011 'Developing tagging practices for VOICE – the challenge of variation'. Poster Display, 4th International Conference of English as a Lingua Franca, Hong Kong, 26-28 May 2011.

'Multiple voices: Highlighting aspects of form and function in ELF'. Symposium on 'East-West dialogue on English as a Lingua Franca, Part Two: ELF in Europe', convened by Jennifer Jenkins and Anna Mauranen. AILA, Beijing, 23-28 August 2011. (with A. Breiteneder)

'VOICE and ELFiA: Challenges in the collection and processing of ELF data'. Symposium on 'The VOICEs of Europe and Asia: diversity of data but harmony in approach', convened by Barbara Seidlhofer and Andy Kirkpatrick. AILA, Beijing, 23-28 August 2011. (with J. Patkin)

‘Annotation by transformation: Advantages of TBL for tagging spoken ELF data’. Individual paper, ICAME 32, Oslo, 1-5 June 2011. (with M. Radeka)

2010 ‘Will VOICE go POS?’. Visual Presentation, 3rd International Conference of English as a Lingua Franca, Vienna, 22-25 May 2010.

‘Evaluating the applicability of existing POS-taggers for a corpus of English as a lingua franca (VOICE)’. Individual work-in-progress paper, ICAME 31, Giessen, 26-30 May 2010

2009 ‘Aspiration, [θ]/[ð] and /r/ in English as Lingua Franca – a psycholinguistic approach’, Individual paper, 2nd International Conference of English as a Lingua Franca, Southampton, 6-8 April 2009.

Publication of corpora and accompanying electronic material

2013 *VOICE Part-of-Speech Tagging and Lemmatization Manual.* http://www.univie.ac.at/voice/documents/VOICE_tagging_manual.pdf (with B. Seidlhofer and M. Radeka)

“VOICE POS Online”. *Using VOICE Online*. Vienna. http://univie.ac.at/voice/help/pos (with B. Seidlhofer and M. Radeka)

The Vienna-Oxford International Corpus of English (version 2.0 XML). Director: Barbara Seidlhofer; Researchers: Angelika Breiteneder, Theresa Klimpfinger, Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka.

The Vienna-Oxford International Corpus of English (version POS XML 2.0). Director: Barbara Seidlhofer; Researchers: Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka, Nora Dorn.

The Vienna-Oxford International Corpus of English (version 2.0 online). Director: Barbara Seidlhofer; Researchers: Angelika Breiteneder, Theresa Klimpfinger, Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka. http://voice.univie.ac.at.

The Vienna-Oxford International Corpus of English (version POS Online 2.0). Director: Barbara Seidlhofer; Researchers: Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka, Nora Dorn http://voice.univie.ac.at/pos.

AWARDS AND SCHOLARSHIPS

07-09/2014 Dissertation Completion Fellowship awarded by the University of Vienna.