MASTERARBEIT | MASTER'S THESIS

Titel | Title

Diachronic Linguistic Diversity in Canada

verfasst von | submitted by

Juliane Rabea Benson B.A.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 647
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Digital Humanities
Betreut von | Supervisor:	Univ.-Prof. Mag. Hannes Fellner M.A. PhD

Acknowledgment

I would like to thank my friends and family who supported me throughout my master's studies and while writing my master's thesis.

Thanks to Carla, Robin, and Max, who accompanied me on my journey through the master's thesis with joint study sessions. Thanks to Jona for the exchanges of ideas about the master's thesis and for the distractions from the master's thesis. Many thanks to my proofreaders Katharina, Carla, Hannes, and Petra.

I would also like to express my special thanks to everyone who made my research stay in Canada possible and wonderful as part of my master's thesis: Charles and Sigwan. As well as the University of Vienna, which provided financial support for my research stay through the KWA scholarship. Many thanks also to Petra for the spontaneous road trip we took together.

I would also like to thank Hannes Fellner and Andreas Baumann for your support with my research and for your guidance. And lastly the whole DigiLingDiv project team for the nice working environment.

Abstract

More than two-thirds of Indigenous languages were classified as endangered or extinct in Canada in 2018. This reflects the global trend of languages becoming endangered and at risk of disappearing. As one of the largest countries in the world, Canada has a rich linguistic diversity, which includes two official languages alongside numerous Indigenous and immigrant languages.

This master's thesis specifically examines how Canada's linguistic diversity has evolved over the past 30 years.

The first goal of this master's thesis is to develop a reusable workflow for creating diachronic speaker datasets along with the actual dataset. Language data from *Ethnologue* for Canada and the Canadian census served as the foundation for this work. The diachronic analysis of Canada's linguistic diversity is based on quantitative methods, including measures of Linguistic richness, linguistic evenness, the Index of Linguistic Diversity by Harmon and Loh (2010), the Linguistic Diversity Index by Greenberg (1956), Shannon entropy and the Hill numbers framework, as well as an adapted Red List Index (Butchart et al., 2007).

The study shows that Canada's linguistic diversity has increased over the last 30 years, which can be linked in particular to the increase in immigrant languages and speaker numbers. There have been no major changes to the official languages regarding all measures. Indigenous languages have become more diverse in their own language group, but at the same time more endangered. Therefore, the thesis concludes that one should not be misled by the numbers; increased diversity is not necessarily linked to the preservation of Indigenous languages—it all depends on how you measure diversity.

Zusammenfassung

Mehr als zwei Drittel der indigenen Sprachen wurden 2018 in Kanada als gefährdet oder ausgestorben eingestuft. Dies spiegelt auch den globalen Trend wider, dem zufolge Sprachen gefährdet und vom Aussterben bedroht sind. Als eines der größten Länder der Welt verfügt Kanada über eine reiche sprachliche Vielfalt, zu der neben zwei Amtssprachen auch zahlreiche indigene Sprachen und Sprachen von Einwanderern gehören.

Diese Masterarbeit untersucht insbesondere, wie sich die sprachliche Vielfalt Kanadas in den letzten 30 Jahren verändert hat.

Das erste Ziel der Masterarbeit ist die Entwicklung des Datensatzes mit einem wiederverwendbaren Workflow für die Erstellung eines diachronen Datensatzes. Als Grundlage dienten Sprachdaten aus *Ethnologue* für Kanada und der kanadische Census. Die diachrone Analyse zur sprachlichen Vielfalt Kanadas basiert auf quantitativen Methoden, die verschiedene Maße umfassen: der Linguistic richness, dem linguistic evenness, dem Index of Linguistic Diversity von Harmon and Loh (2010), dem Linguistic Diversity Index von Greenberg (1956), der Shannon entropy und dem Hill numbers Framework, sowie einem adaptierten Red List Index (Butchart et al., 2007).

Die Studie zeigt, dass Kanadas Sprachdiversität über die letzten 30 Jahre zugenommen hat, was besonders mit dem Anstieg an Einwanderersprachen und dessen Sprecherzahlen verknüpft werden kann. Die offiziellen Sprachen zeigen keine großen Veränderungen. Die Indigenen Sprachen wurden diverser in ihrer Sprachgruppe, aber zeitgleich bedrohter. Daher kommt diese Masterarbeit zum Schluss, dass die Zahlen täuschen können: Die zunehmende Diversität ist nicht notwendigerweise mit dem Erhalt von Indigenen Sprachen verbunden — alles hängt davon ab, wie Diversität bemessen wird.

Contents

1 Introduction

The global trend of language endangerment is well known and is also present in Canada. According to Simons (2019), more than two-thirds of Indigenous languages were classified as endangered or extinct in Canada in 2018. As one of the largest countries in the world, Canada exhibits a rich and unique diversity of languages. The country has two official languages, rooted in its colonial history, alongside numerous Indigenous and immigrant languages. Canada was colonized by various European powers, and played a notable role in the rivalry of dominance of the English and the French Crowns overseas. Today, English and French are the two official languages that dominate the multilingual society of Canada, whereas Indigenous languages are facing the threat of extinction. In addition, modern migration influences the linguistic landscape of Canada and the country pursues a policy of multilingualism to foster their language diversity. Canada is an exciting area of research for linguistic diversity due to its history and current language dynamics, which also reflect global trends. At the same time, Canada's languages have been relatively well-researched and the size of the country favors the availability of comprehensive data, making it an ideal case study for examining the evolution of linguistic diversity.

This master's thesis focuses on the following question: "How has linguistic diversity in Canada changed in the last 30 years?"

Previous research has explored linguistic diversity and developed diversity measures. Linguists have adapted methods originally used in biodiversity studies to examine linguistic diversity. However, there is no comprehensive diachronic study of linguistic diversity in Canada using the evolutionary linguistics approach. Existing studies are either global in scope or narrowly focused, while diachronic studies do not cover the geographic area in question.

To answer the research question, a quantitative analysis of linguistic diversity in Canada is performed on a dataset created in this study that is based on the diachronic data of *Ethnologue* (editions 13 to 22 and 27) and the Canadian census (years 1966 to 2021). The time frame of this study (the last 30 years) was chosen for reasons of data availability; *Ethnologue* editions since 1996 are available online. The quantitative analysis encompasses a descriptive analysis of the speaker number and status situation of the languages, the use of measurements of linguistic

diversity based on speaker numbers — linguistic richness, linguistic evenness, Index of Linguistic Diversity (ILD) by Harmon and Loh (2010), the Linguistic Diversity Index (LDI) by Greenberg (1956), Shannon Entropy and the Hill numbers Framework — and a measurement on the threat status of the linguistic diversity, the adapted Red List Index (RLI) (Butchart et al., 2007).

Additionally, the aim is to create a workflow that, similar to this work, enables researchers to use the diachronic data from *Ethnologue* for further research, such as analyzing the linguistic diversity of other countries.

In the end of this study the following subquestions should be answered: "How has the linguistic diversity changed for the sum of all languages and the different groups (official languages, Indigenous languages and immigrant languages)", "what is the trend of the status of these languages" and "do the different methods show same results?". In addition, the following hypothesis should be tested: Linguistic diversity is declining in Canada, which is reflected by a decrease in the LDI and Shannon entropy, while more languages become endangered, captured through a decrease of the RLI; Indigenous languages are particularly affected by this decline.

1.1 Road map

The first chapter (chapter 2) of the master's thesis examines the context including Canadian history and the country's language policy. This is followed by a presentation of the data sources in chapter 3. Chapter 4 describes the theoretical background of linguistic diversity and previous research. Chapter 5 introduces the methods, and chapter 6 explains the workflow for creating the dataset. With all these preparations the question of the Canadian linguistic diversity can be pursued and the analysis can be found in chapter 7, and the following chapter 8 summarizes the results.

2 Context

This study begins with a brief historical context of Canada and an overview of its language policy. The presentation of the data sources used in the analysis follows.

2.1 History of Canada

The colonization of Canada began in the 16th century on its east coast. Newfoundland was claimed by England, and Jacques Cartier claimed the land north of the St. Lawrence River for France and founded the colony of New France on parts of what is now the province of Quebec. The permanent settlement of Canada by European colonists had a profound impact on the Indigenous people of Canada. On the one hand, trade, especially in furs, was established and alliances were formed, but on the other hand, they brought diseases and weapons with them (Belshaw, 2016; Minister of Citizenship and Immigration Canada, 2021; The Canadian Encyclopedia, 2025b).

There were armed conflicts in various constellations between and among the Indigenous population and the European settlers. In the Beaver Wars, for example, the Algonquians, allied with France, waged war against the Iroquois, who were supported by the English. European powers thus fought proxy wars in the colonies. The territory claimed by England expanded, among other things, with the Hudson's Bay Company, which had trading privileges in northern Canada and replaced France's fur trade monopoly. During the Seven Years' War and after France's defeat by England at the Battle of the Plains of Abraham near Quebec City in 1759, France was forced to cede the colony of New France to the English Crown under the Treaty of Paris in 1763 (Belshaw, 2016; Minister of Citizenship and Immigration Canada, 2021; The Canadian Encyclopedia, 2025b).

In 1776, the *Declaration of Independence* created the United States, with the colonies in North America seceding from Great Britain. This led to the migration of Loyalists to Canada. In the War of 1812, the United States attempted to conquer the northern British colonies, with Canadian Indigenous people fighting on the side of the British. When slavery was abolished in Great Britain in 1833, this also applied to Canada as part of the British Empire. In 1840, the province

of Canada was established from Upper Canada (now Ontario) and Lower Canada (now Quebec). The *British North America Act* of 1867 formed the Dominion of Canada with the provinces of Ontario, Quebec, New Brunswick, and Nova Scotia. This marked the founding of the federalist state of Canada. The Canadian confederation introduced both the federal and provincial levels of government. Gradually, Canada grew, and other provinces and territories joined from former British colonies, and Canada expanded to the Pacific. As a result, the settlement of the West became a national priority. The Canadian Pacific Railway line was also built as part of this process (Belshaw, 2016; Minister of Citizenship and Immigration Canada, 2021; The Canadian Encyclopedia, 2025b).

As part of the British Empire, Canada participated in World War I in Europe. "The war strengthened both national and imperial pride" (Minister of Citizenship and Immigration Canada, 2021). Between 1916 and 1918, women's suffrage was introduced in Canada, with the exception of Quebec, where women have been allowed to vote since 1940. After the first world war, the British Commonwealth was founded with Canada as a member (Minister of Citizenship and Immigration Canada, 2021).

Like many other countries around the globe, Canada flourished in the 1920s and suffered during the Great Depression in the 1930s. Canada also fought alongside the Allies in World War II, for example, on D-Day when they landed at Juno Beach (Minister of Citizenship and Immigration Canada, 2021).

Since 1975, Canada is part of the Group of Seven (G7). In 1982 the British legislative jurisdiction over Canada ended with the *Constitution Act*, which also encompasses the Charter of Rights and Freedoms (The Canadian Encyclopedia, 2025b).

Even though the francophone nationalist movement in Quebec emerged, which is also reflected in the language policy, a referendum on the separation question failed in 1980 and 1995. However, this did lead to the recognition of Quebec as a nation within Canada. Canada reached its present territorial form in 1999 when Nunavut became part of Canada (The Canadian Encyclopedia, 2025b).

2.1.1 History of Indigenous People in Canada

The history of Indigenous people in Canada can be traced back more than 10000 years (McGhee, 2025).

Canada recognizes three groups of Indigenous people. The first, the Inuit, are people living mainly in the arctic region. The term *First Nations* refers to the original inhabitants mostly south of the Arctic who form the second group. More than 600 First Nations are recognized today.

The third group is the Métis, people with Indigenous and European ancestry, who are forming a nation as well. (Parrott, 2023)

With the colonization, the lives of Indigenous people changed. Europeans interfered with the environment and, in turn, with the livelihoods of Indigenous communities, as described above. Seeing the relevance of alliances with Indigenous peoples for military actions, the British Indian Department was created in 1755. For the same reason, the *Treaties of Peace and Friendship* were concluded with First Nations. At the same time, Indigenous and African people were enslaved by the British empire. With the end of the Seven Years' War and the definitive victory of Great Britain over France in the question of colonies in North America and the *Royal Proclamation* in 1763, the Crown presented itself as protector by recognizing that Indigenous title and rights can only be erased through a treaty with the Crown. The same indigenous territory could only be transferred to the Crown and by treaty. After the War of 1812, the treatment of Indigenous people changed, as they were now seen more as a threat to the British Government that should be minimized by assimilating the Indigenous people into the Euro-Canadian society. This was branded as "civilization" and a justification for colonialism. In addition, more and more land was ceded, and Indigenous people were displaced into reserves. This "civilization" policy was also expressed by law, i.e., the *Gradual Civilization Act* in 1857. Enfranchisement was encouraged as a form of assimilation, thereby allowing the Indigenous people to gain certain new rights by abandoning other current rights and status. (Belshaw, 2016)

After the Confederation, the first *Indian Act* was passed in 1876, which follows this policy:

> "The *Indian Act* attempted to generalize a vast and varied population of people and assimilate them into non-Indigenous society. It forbade First Nations peoples and communities from expressing their identities through governance and culture. The Act replaced traditional structures of governance with band council elections. Hereditary chiefs — leaders who acquire power through descent rather than election — are not recognized by the *Indian Act*" (The Canadian Encyclopedia, 2022).

It defined an Indigenous person with *Indian Status* as someone who is officially registered. Not all Indigenous people are included in this term. The status defines certain rights and benefits, e.g., regarding taxes. Many types of ceremonies or gatherings were now prohibited. The act also manages the reserves. Reserves are land assigned to First Nations people to "live securely, obtain education and practise agriculture" (Irwin, 2022), although the reserves are just a fraction of the traditional land, and restricted and changed the livelihood of First Nations. Furthermore, the management of First Nations is controlled by the government. Around that time, the *Numbered Treaties* were concluded between First Nations and the Crown about land; Canada needed the land mainly for industrial purposes and to drive their assimilation policy

forward. First Nations were promised rights and payments. "Many First Nation leaders sought to use the treaties as a means of coping with the destruction of traditional economies — notably, the decimation of the buffalo on the Prairies" (Filice, 2016). These treaties, together with the *Indian Act*, are the basis of the alliance of First Nations and Canada (Belshaw, 2016; Irwin, 2022; Parrott, 2023; The Canadian Encyclopedia, 2022).

Following the creation of the Dominion Canada, the expansion to the West took place as Canada purchased territories from the Hudson's Bay Company for the Dominion without consulting the local population: the Métis and First Nations. This led to the Red River Rebellion. The rebellion was suppressed, and the province of Manitoba and the Northwest Territories were created, forcing the Métis to migrate westward. A second rebellion, in which First Nations people and Métis again experienced a loss of land rights and the disappearance of their livelihoods, was the Northwest Resistance. This resistance was also suppressed, and as consequence, the pass system was installed. This restricted, controlled and monitored the movement of Indigenous people, even if it never had a legal basis (Belshaw, 2016; Minister of Citizenship and Immigration Canada, 2021; Nestor, 2018; The Canadian Encyclopedia, 2025b).

Residential schools were boarding schools for Indigenous children funded by the government and co-managed by the church. It was also a tool of assimilation by separating the children from their home and culture. The first residential school was opened in 1831 in Ontario. In their treaties, First Nations asked for support in schooling, hoping it would be beneficial for their children to learn the skills needed in the more industrialized world. With the *Indian Act*, the government ensured schooling through law, but under the disguise of education, the purpose of assimilation was pursued to make the First Nations financially independent. As a result, "industrial" schools as well as the cheaper boarding schools with less of a specific focus on trade were created. The schools became obligatory, which enabled the forced separation of children from their parents, and changed their purpose "from integration to segregation" (Belshaw, 2016). The peak of residential schools was in 1930 with 80 at the same time. In the residential schools, children received education, but were also forced to work to finance and maintain the schools. Besides the poor education, more than 150,000 students had to suffer in the residential schools.

> "Indigenous children were taken from their homes, separated from their communities and families, and were forbidden to speak their languages or practice their culture. Students were poorly fed; subjected to medical experiments; forced to undertake manual labour; and were subjected to physical, sexual, and mental or spiritual abuse" (Belshaw, 2016).

In 1996, the last residential school closed. More than 6000 Indigenous children died in resi-

dential schools, and the trauma is still present. As a result of the *Indian Residential Schools Settlement Agreement*, the *Truth and Reconciliation Commission* was created in 2008 as part of the investigation to uncover the history of residential schools, and to compensate and support survivors. The final report of the commission (2015) concludes:

> "The establishment and operation of residential schools were a central element of this [assimilation and elimination] policy, which can best be described as 'cultural genocide.'"

Since 2013 the Orange Shirt Day remembers the residential schools and that "Every Child Matters". The report on residential schools was only the beginning, and the investigations continue still. In 2021, for example, possible unmarked graves were found on the grounds of residential schools. Also in 2021, the final *Report of the National Inquiry into Missing and Murdered Indigenous Women and Girls* was published (Belshaw, 2016; Miller, 2024; The Canadian Encyclopedia, 2025b).

Together with the assimilation policy, enfranchisement was encouraged. "By enfranchising, a person was supposed to be consenting to abandon Indigenous identity and communal society (with its artificial legal disabilities) in order to merge with the 'free', individualistic and non-Indigenous majority" (McCardle, 2021). In other words, by meeting certain criteria, First Nations people got the right to vote by losing the *Indian Status*. It was after the Second World War, between 1949 and 1969, that Indigenous people got voting rights on a provincial level, and in 1960, they were allowed to partake in federal elections without losing their status. Also, in 1951, women got the right to vote in Band Councils. When organizations were allowed, new ones were established to represent the interests of the Indigenous population. Following initiatives by Indigenous people, Indigenous rights were enshrined into the *Constitution Act* of 1982. In the 1990s, modern agreements were made with Indigenous people that concerned rights and land claims. The *United Nations Declaration on the Rights of Indigenous Peoples* was presented in 2007, "an agreement that recognizes Indigenous rights to self-government, land, equality and language, as well as basic human rights" (Henderson & Bell, 2019). But only in 2016, the Canadian government signed the international law, and in 2021 it was incorporated into Canadian law (Belshaw, 2016; Canadian Museum of Immigration at Pier 21, 2025; The Canadian Encyclopedia, 2025b).

2.1.2 History of Immigration in Canada

In the history of immigration, colonization plays a major role in Canada; as European settlers came to Canada permanently. At this time, immigration was dominated by French and by

British people, but also people from other European countries settled in Canada. In addition, at the same time, enslaved Africans were brought to North America and the colonies. Following the American Revolution, British Loyalists came to Canada from the south, resulting in a wave of immigration. This also included Black Loyalists, but due to discrimination, many left for Sierra Leone. In the first half of the 19th century, more than one million British migrants arrived. In addition, the abolition of slavery during this period made Canada a destination for refugees. Bad circumstances in Europe made people flee to North America. One significant example of this was the Irish potato famine (Canadian Museum of Immigration at Pier 21, 2025; Minister of Citizenship and Immigration Canada, 2021; The Canadian Encyclopedia, 2025b).

Shortly after the confederation, an immigration policy was defined, showing the national importance of the immigration question. The *Immigration Act* of 1869 set rules for safe passage and defined who could enter Canada, while the *Dominion Lands Act* of 1872 provided land grants to attract settlers to western Canada. In particular, the industrialization asked for laborers and promoted immigration. For example, the construction of the Canadian Pacific Railway, which connected the east and the west of Canada, was also carried out by a significant number of Chinese immigrants in 1885 under bad conditions. Shortly after the completion of the railway, the *Chinese Immigration Act* was passed, making it clear that immigration from China was no longer welcome. The Canadian immigration policy was highly racist and discriminatory, preferring white British, US-American, or European migrants who were considered easy to assimilate. The national priority of settling the West was further encouraged by programs and facilitated by the railway. Furthermore, the land situation in the US shifted immigration to the north. Canada experienced an immigration boom between 1896 and 1913. However, with the second *Immigration Act* of 1906, the government's power was extended to admit, refuse, or deport newcomers. The *Naturalization Act* of 1914 defines that immigrants who want to become Canadian citizens are required to have lived in Canada for five years and speak English or French (Belshaw, 2016; Canadian Museum of Immigration at Pier 21, 2025; Dirks, 2024; Minister of Citizenship and Immigration Canada, 2021; The Canadian Encyclopedia, 2025b; Troper, 2024).

During World War I, immigration was further restricted and discouraged. Moreover, immigrants who came from enemy countries were arrested and interned in labor camps. After the war, immigration was further restricted and could be denied on ideological grounds. Particularly strict measures were taken against immigration from Asia. In the aftermath of the Great Depression, thousands of immigrants were deported under questionable circumstances, and the most restrictive immigration policy was pursued, allowing very few people to enter Canada. Consequently, "Immigration dropped and many refugees were turned away, including Jews

trying to flee Nazi Germany in 1939." (Minister of Citizenship and Immigration Canada, 2021) During World War II, only highly educated refugees and children were admitted, as well as war brides, while *enemy aliens* were subjected to internment. Canada has since apologized for the arrests (Belshaw, 2016; Canadian Museum of Immigration at Pier 21, 2025; Dirks, 2024; Minister of Citizenship and Immigration Canada, 2021; Troper, 2024).

After the war, Canadian citizenship was officially defined for the first time. Furthermore, the restriction on Chinese immigration was lifted, enabling citizenship with voting rights. In 1952, the *Immigration Act* was renewed, and an immigration appeal board was created. This allows rejected and deported immigrants to challenge the decision. The postwar years were marked by a new wave of immigration landing on the east coast, primarily on Pier 21 in Halifax, from where they found their settlement all over Canada. Due to the Cold War, immigration from eastern Europe was restricted. However, southern Europeans, for example Italians, immigrated to Canada by ship and through Pier 21. In 1967, a points system was established to regulate immigration in more objective terms to decrease discrimination and arbitrariness. Admission was granted on the basis of merits. Two years later, Canada signed the *United Nations Refugee Convention* with a delay of 18 years. In the 1970s, an official multiculturalism policy was introduced that should protect the individuals' diversity and dignity. These years also marked a change, as the majority of immigrants were no longer of European origin. An even clearer definition and the distinction of refugees and immigrants was included in the *Immigration Act* of 1976; leading to the reform of the immigration appeal board to the *Immigration and Refugee Board* in 1989. Refugees came from European countries, but also from Africa, like Uganda, from Latin Americas, e.g. Chile, or Asia, e.g. Vietnam, fleeing from oppression, war and persecution and seeking protection. Canada welcomed the immigration of people wanting to support the Canadian economy through substantial monetary investment, forming a bypass for asylum-seeking people (Belshaw, 2016; Bishop, 2015; Canadian Museum of Immigration at Pier 21, 2025; Troper, 2024).

In the 20th century, the immigration shifted from settlements in rural areas to urban centers, which was promoted through industrialization and is still the norm today (Belshaw, 2016; Canadian Museum of Immigration at Pier 21, 2025; Dirks, 2024; Troper, 2024). The museum Canadian Museum of Immigration at Pier 21 (2025) concludes: "In the 21st century, Canada's immigration system is designed to be responsive to global humanitarian crises, national security concerns and labour market needs".

2.2 Canadian language policy

2.2.1 Official Languages

The *Official Languages Act* of 1969 defines that the official languages in Canada are English and French. This is true for the federal level. However, the provinces and territories have individual regulations. Most of them have English as their official language. Quebec is the only province to have French as its sole official language. New Brunswick is officially bilingual with English and French. In all provinces, English is the majority language, with the exception of Quebec being historically francophone.

The official languages have their origins in the colonization and the British and French colonies. After the cession of New France to the English Crown, the *Quebec Act* in 1774 ensured French-language rights and Catholicism was allowed to accommodate French settlers. This was maintained and established with the Confederation at the federal level and in Quebec. The *British North America Act* states that English and French should be used complementary and linked this with denominational schooling, as anglophones are protestant and French-speaking people tend to be catholic. Nonetheless, there was no equality between these languages, and the Francophones were discriminated in the provinces through laws. In Ontario, for example, the *circular No. 17* banned French in schools. Although those laws are now abolished and "the right to use Canada's official languages; and the right of French or English minorities to an education in their language" (Foot, 2020) are written in the *Charter of Rights and Freedoms* that is part of the *Constitution Act* of 1982, the French situation outside of Quebec is still not equal to English on federal level (Mougeon, 2015; Office Of The Commissioner Of Official Languages & Burnaby, 2020; The Canadian Encyclopedia, 2025a).

With the rise of francophone nationalism in Quebec, language was also a major topic. The preservation of the French language was meant to guarantee the survival of the Quebec nation. As one of the first steps, *Bill 63* (*Act to Promote the French Language in Quebec*) was enacted in 1969, to give parents the freedom to choose the school language for their children, but French language classes are mandatory in English schools. The official language of the province was set to French in 1977 with the *Charte de la language française* also known as *Bill 101*. This stipulated that French should be the dominant language in public. This was criticized as it violated the freedom of speech. So, the law was adapted to allow bilingualism, while maintaining the predominance of French (Mougeon, 2015; Paquet & Levasseur, 2019).

2.2.2 Indigenous Languages

As an important act of colonization, places are given names; some places are renamed, thereby losing the connection between the place and its original inhabitants. The names have meanings that describe the land, and thus knowledge about the characteristics of the place disappears. Other names are adopted into the language of the colonizers, who, however, do not know their meaning. This is how Canada got its name: It originates from the Iroquois word *kanata*, which means village (Canadian Geographic, 2018a, 2018b; Minister of Citizenship and Immigration Canada, 2021).

"Broadly defined, an Indigenous language is any language that is 'native' to a particular area" (Walsh, 2005). In Canada there are more than 70 distinct Indigenous languages from 12 language families (Gallant, 2022; Statistics Canada, 2017a, 2023a):

- Algonquian languages
- Athapaskan languages
- Haida
- Inuktut (Inuit) languages
- Iroquoian languages
- Ktunaxa (Kutenai)
- Michif
- Salish languages
- Siouan languages
- Tlingit [1]
- Tsimshian languages
- Wakashan languages

Some of the Indigenous languages are creole languages. For example, Michif is a creole language based on Cree and French used by the Métis. Furthermore, there were Indigenous sign languages, but they have mostly been displaced by the American and the Quebec Sign Languages. (Gallant, 2022)

[1] the census 2021 integrates Tlingit in the Athapaskan family.

Because of the colonization and restrictive history in general towards Indigenous people, a lot of the Indigenous languages are endangered. Especially the residential schools were structured to lead to language loss. The children were not allowed to use their Indigenous first language, but were forced to use English and French respectively. The prohibition was also enforced with physical punishment. This created a fear among the children of using their language and was also reflected in the non-transmission of the language to the younger generation. Since language is closely related to identity, the ban on Indigenous languages would presumably support the assimilation policy. Today, the languages often have a very small speaker community. The languages are used more by the people living in reserves, while Indigenous people living in urban areas and outside the reserves tend to use them less. Being recognized as a language under threat helps to raise awareness and funding for revitalization programs. Those programs can take various forms and can be led by schools and universities, organizations, or the community itself. They can be based, among other things, on language learning to increase the speaker numbers, on public relations to increase the visibility, or on documentation to form a basis for the language education and to keep the knowledge on and of the language. This is also a precautionary measure to later theoretically be able to revitalize dormant languages and to prevent definite extinction. The language revitalization initiatives are quite diverse, using different tools and methods, e.g., digital tools such as websites, meeting different purposes and needs. One organization is for example the First Peoples' Cultural Council (FPCC), it fosters language through different programs and initiatives. They are a partner of the Endangered Languages Project (ELP) that collects and offers a pool of online resources. Every 4 years, since 2010, they also publish a report on the *Status of B.C. First Nations Language* (First Peoples' Cutural Council, 2010, 2014, 2018, 2022), which are valuable resources for this study. This efforts aim to prevent language loss and preserve languages, and therefore also culture, knowledge, and identity (Brinklow, Littell, Lothian, Pine, & Souter, 2019; Eaton & Turin, 2022; *First Peoples' Cultural Council*, 2025; Gallant, 2022; Khawaja, 2021; McIvor, 2018; Norris, 2007; Rice, 2023).

Indigenous languages are only official in Nunavut and the Northwest Territories. Nova Scotia has recognized Mi'kmaq as the province's first language in 2022. The Canadian government, with the signing of the UN *Declaration on Indigenous Rights*, commits to recognizing the languages of the Indigenous people. And with the *Indigenous Language Act* 2019, they committed furthermore to the protection and support in revitalizing the languages. Despite criticism, Inuit specific needs are not explicitly noted, and it does not obligate monetary support. Globally, there are efforts to support endangered languages and to raise awareness of the threat to Indigenous languages. In this context, the *International Decade of Indigenous Languages*

2022-2032 with the aim to “highlight the urgent need to preserve, revitalize, and promote Indigenous languages” (UNESCO, 2024) was declared by UNESCO (Gallant, 2022; Office Of The Commissioner Of Official Languages & Burnaby, 2020).

2.2.3 Immigrant Languages

Immigrant languages are languages that came to Canada through migration, with the exception of English and French. Immigrant languages do not have an official status in Canada, but are usually official languages elsewhere. The languages have origins all over the world reflecting the origins of the immigrants. Therefore, the language communities also reflect the waves of immigration, which can also be seen in the use of the language. If they came at a time when the official languages were favored, immigrants tended to shift from their heritage language to the official language, also to help their children to integrate. Today, however, Canada’s language and multiculturalism policy pursues the goal of integration, which is intended to preserve cultural identity and language. The value of transmitting the heritage language to the next generation is seen as more valuable and as a “form of human capital” (Arsenault Morin & Geloso, 2020). Nevertheless, it is still assumed that at least one of the official languages will be adopted. In the province of Quebec, with its francization policy, the immigrants are encouraged to learn French through courses organized by the Quebec government. Immigrant languages are preserved more through second language programs. Since migration mainly occurs to urban regions, more immigrant languages are spoken there (Nagy, 2021; Office Of The Commissioner Of Official Languages & Burnaby, 2020; Paquet & Levasseur, 2019; Saint-Jacques, Chambers, & Cooper, 2019).

3 Dataset

For this study, diachronic language and speaker data from *Ethnologue* is used. This data is enriched by the past six Canadian censuses to have a solid basis for the analysis.

3.1 *Ethnologue*

Ethnologue is a standard resource for language data. The first edition was published in 1951 and to date there are 28 editions. The publisher is SIL international (former Summer Institute of Linguistics). It is a global christian non-profit organization which is located in the United States. They work among other things on bible translation and linguistic research. (*SIL*, 2025) In this context *Ethnologue* as language database was created to gather information on this missionary work. Today, "the unique contribution of Ethnologue is to document the global language ecology" (Eberhard, Simons, & Fennig, 2025a). Since the 13th edition of *Ethnologue* from 1996, it is available online; earlier editions were only printed. Nowadays the *Ethnologue* main presence is the website which is behind a paywall. Since the 18th edition in 2015 a new edition is published annually on February 21st, the International Mother Language Day. English is the underlying language and the whole language catalog is in English, which is also important for terminology questions. Together with the International Standard Organization, SIL international developed the ISO 639-3 standard, a three-letter code for the identification of languages. The codes were created on the basis of the *Ethnologue* and serves since 2005 as the identifier within the *Ethnologue*. SIL global is the Language Coding Agency for ISO 639 Set 3 and defines a distinct language following three principles (Eberhard et al., 2025a):

- "Two related language varieties are normally considered to belong to the same individual language if speakers of each language variety have inherent understanding of the other language variety at a functional level (i.e., they can understand each other based on knowledge of their own language variety without needing to learn the other language variety). Where such mutual intelligibility does not exist, the two language varieties are generally seen to belong to different individual languages."

- "Where spoken intelligibility between language varieties is marginal, the existence of a common literature or of a common ethnolinguistic identity with a central language variety that both speaker communities understand is a strong indicator that they should nevertheless be considered language varieties of the same individual language."
- "Where there is enough intelligibility between language varieties to enable communication, they can nevertheless be treated as different individual languages when they have long-standing, distinctly named ethnolinguistic identities coupled with established linguistic normalization and literatures that are distinct."

On problems with the ISO code see Drude (2018). Since the question of delimitation of a language can be difficult between the criteria, the concept "macrolanguage" helps to fill the gap originating through the second principle, as Collin (2010) explains.

What information does *Ethnologue* collect?

"The purpose of *Ethnologue* is to provide a comprehensive listing of the known living languages of the world." (Eberhard et al., 2025a) There are more than 7000 known living languages. In 2025, *Ethnologue* is listing 7159 languages, which marks an increase from more than 6700 languages in 1996 (Eberhard et al., 2025a; Grimes, 1996). The listed languages are rather presented with a small summary than with exhaustive information on the language. In addition to the language catalog, there is a summary of the linguistic situation for each country of the world. Furthermore, *Ethnologue* provides language maps showing, for example, the location of languages. Also *Ethnologue* shares its own research in the form of articles or statistics. For this study, because of the geographical frame on Canada, the country pages are important. These provide a general summary, a list of the languages spoken there, as well as the languages sorted by their status (EGIDS, see chapters 4 and 5) in the country. Moreover, maps and graphs for the language situation in the country are shown. The content and representation changes overtime, but include especially the first two aspects: the summary and the language list. The table 3.1 describes the information available in the summary over time. A more precise explanation of the language lists is given in 6.2.

Ethnologue collects its data from different sources such as fieldwork and censuses. This means it provides more information than the individual sources itself, but as a consequence the data is not necessarily consistent and comparable because of the different sources which relay on different methods themselves. In general, the source is given together with the infor-

Information	1996	2000	2005	2009	2013	2015	2016	2017	2018	2019	2024
Population	X	X	X	X	X	X	X	X	X	X	X
Principal Languages		X	X	X	X	X	X	X	X	X	X
Immigrant Languages	X	X	X	X	X	X	X	X	X	X	
Literacy Rate	X	X	X	X	X	X	X	X	X	X	X
International Conventions											X
General References	X	X	X	X	X	X	X	X	X	X	X
Religion	X	X									
Blind Population	X	X	X	X							
Deaf Population		X	X	X	X	X		X	X	X	X
Deaf Institutions	X	X	X	X	X						
Recognized Nationalities									X	X	X
Data accuracy estimate	X	X									
Languages Counts	X	X	X	X	X	X	X	X	X	X	X
Diversity index		X									

Table 3.1: This table shows the data available in *Ethnologue* for Canada in the summary section

mation, but, as Hammarström (2015) points out, sometimes the source is missing and the data cannot be retraced. Another problem is that due to the status as a standard resource for linguistic data and the gathering process of *Ethnologue*, there is a risk for circular citation. (Paolillo & Das, 2006) Since *Ethnologue* relays on external information the topicality of the data is different for the languages and also for the regions.

Why *Ethnologue*?

The choice to use the *Ethnologue* as the data basis of this study has multiple reasons. First, the structure of the data is quite practical since the language information is organized by the geographical entity of a country, allowing Canada to serve as a frame for analysis. Second, diachronic data are available. For this study, because of accessibility constraints, the data from editions 13 to 22 and 27 are used (Eberhard, Simons, & Fennig, 2019, 2024; Gordon, 2005; Grimes, 1996, 2000; Lewis, 2009; Lewis, Simons, & Fennig, 2013, 2015, 2016; Simons & Fennig, 2017, 2018). This selection still covers the years 1996 to 2004 quite well; see table 3.2. Compared to other language catalogs, e.g., *Catalogue of Endangered Languages* (2024), *Ethnologue* contains more data and does not focus solely on endangered languages or Indigenous languages. Especially relevant are the data on speaker numbers that can not be found comprehensively in other catalogs like glottolog (Hammarström, Forkel, Haspelmath, & Bank, 2025). An addi-

tional argument is that the use of *Ethnologue* makes this study replicable for other countries. Moreover, as a collector, it bundles multiple sources and is therefore more exhaustive than only using census data for example.

Edition	13	14	15	16	17	18	19	20	21	22	27
Year	1996	2000	2005	2009	2013	2015	2016	2017	2018	2019	2024

Table 3.2: This table shows the publishing year for the *Ethnologue* editions used in this study.

3.2 Census

The Canadian census is used as a second data source. The purpose of this is to improve the quality within the dataset for analysis. Since *Ethnologue* uses the census as a data source, general compatibility is ensured. However, the integration of census data into *Ethnologue* is not comprehensive, and the census provides additional information on languages that are not included in *Ethnologue* for Canada. This is particularly relevant for immigrant languages, as these are listed separately from the language list in earlier editions. However, not all data on existing languages has been transferred by *Ethnologue*. The aim of complementing the data with the census is to close some data gaps and, due to the periodic nature of the census, to create a better basis for imputing the remaining data gaps. In this way, the quality of the data can also be improved, as incorrect or inaccurate data transfers from the census to *Ethnologue* are corrected. How the sources were merged in this work is described in chapter 6.5. Both *Ethnologue* and the census have served as data sources for linguistic diversity research in previous literature, but rarely in a complementary manner.

What is the Canadian Census?

The first known census on today's Canadian territory in 1666 counted European settlers in New France and resulted in the immigration project "filles du roi", which aimed to increase the population and the birth rate. In 1871, just after the confederation, a first national census was conducted in the Dominion Canada, which enumerated not only colonists but also Indigenous people. In 1918 the statistical agency *Dominion Bureau of Statistics* was created to centralize statistical tasks. Since 1971 this organization is called *Statistics Canada*.(Edmonston & Poulin, 2025; Kalbach, James-Abra, & Edmonston, 2021; Prévost, 2020; Statistics Canada, 2018c)

In Canada there are two sorts of census, the census of population and the census of agriculture, which are both based on a questionnaire. The census of environment does not use a questionnaire. The questions regarding the category language are posed in the census of population (forthcoming census). For the census there are two surveys, a short-form and a long-form, which are filled out by a bigger or smaller share of the population, respectively.

Starting with 1971 self-enumeration is used for the census and since 2006 there has been the option to fill it out online. However, in some places, the census is conducted through in-person interviews, but otherwise, one person of a household is responsible to complete the survey for all members. It is mandatory by law to participate in the census. Also, in 1971, Statistics Canada started to perform reverse record check to identify missing values and improve the data quality. Since 1991 the census includes also non-permanent residents. (Kalbach et al., 2021; Statistics Canada, 2018c)

The census has been conducted every five years since 1981, reduced from the previous interval of ten years. For this study the last six censuses from the years 1996, 2001, 2006, 2011, 2016 and 2021 are relevant. Whereby the time frame is defined parallel to the online available *Ethnologue* editions, and more historical data was neglected. The census is available in both official languages, English and French. In addition, assistant material in more languages is provided; as of 2021, the accessibility was increased with the addition of 13 Indigenous languages and 12 immigrant languages. The information on identity that was gathered through the ethnic origin since 1986 was replaced by a question on self-identification in 1996. (Statistics Canada, 2018c)

Questions

Questions and definitions aren't consistent, some change from year to year, in wording and/or appearance. However, the questions in relation to languages are the same for the period in question, and the definitions are similar enough to enable comparability (see 10.1). The French version shows only a small deviation from the English one, by adjusting the order of "English" and "French" by swapping the languages. Over time, the definitions, serving also as support in the process of filling out the questions, became more detailed to encourage a more precise answer. Other small options were changed in the response field by adding or removing lines.

For this study, the important questions are the one on mother tongue and the two on knowledge of languages. The question on mother tongue is the following:

> "What is the language that this person **first learned** at home **in childhood** and **still understands**?" (Statistics Canada, 2023c).

This question was always part of the short-form questionnaire and collected therefore the data

for the whole population. The question on the knowledge of languages is divided, one asking for official languages only and the other one on non-official languages:

> "Can this person speak English or French well enough to conduct a conversation?" (Statistics Canada, 2023c)

and

> "What language(s), **other than English or French**, can this person speak well enough to conduct a conversation?" (Statistics Canada, 2023c)

The question on official language competence was asked every year. Until 2006 it was part of the long-form encompassing a 20% sample of the population, from 2011 onwards the question appears in the short-form asked of the whole population. The question on knowledge of non-official languages, first asked in 1991, was the whole time in the long-form. Thus from 1996 to 2006 the question covered a fifth of the population and in 2016 and 2021 25% of the population. In 2011 the census was conducted solely on the short-form, hence the latter question was not included. (Statistics Canada, 2023c)

Because of online availability constraints, only the data on mother tongue is integrated for all six censuses in this study. Due to the missing long-form for 2011, knowledge of languages was not included for this year. However, 2001 and 2006 are not included because the data on knowledge of languages could not be found online, even though the census dictionaries of these years indicate that the questions were asked. Only summary data on this question is available, which is not detailed enough for this study's design. Since the data are just taken complementary to the *Ethnologue* data, and the focus in the analysis is on the first language (L1), this can be disregarded.

In general, it is important to take into account the political context in which census data is gathered (see also Duchêne and Humbert (2018)). But for Canada, the trust in the statistic agency is quite high and, therefore, it can be assumed that the information collected through the census is credible. Boissonneault and Hammarström (2025) evaluate diachronically the Canadian census data for speaker numbers with the *Ethnologue* for Indigenous languages. They found that in comparison to *Ethnologue* the census underestimates more commonly spoken languages and overestimates less commonly spoken languages.

4 Theoretical Background

This chapter provides background information on linguistic diversity and presents previous research on this topic.

4.1 From Diversity to Linguistic Diversity

> "the fact of many different types of things or people being included in something; a range of different things or people"

This definition is given by Cambridge University Press (2025). In general, it emphasizes that for diversity, differences between *things* are needed, while they should still be comparable in other characteristics. The second scope is that the *things* can form a group so that the group can contain a variety of those *things*. For the purpose of linguistic diversity, the *things* are languages. A group of people can speak different languages, and can therefore be considered diverse based on their languages. This definition also captures that diversity is not binary, but a spectrum. Something can be more or less diverse. Diversity as a research object is quite common, and especially known in connection with ecology studies, i.e., biodiversity. The Oxford English Dictionary (2025) complements the entry *diversity* with specific use cases of diversity: "Variation in plant and animal life, esp. as represented by the number of species, or dissimilarities between genomes." This clearly refers to biodiversity and already gives hints on how diversity is measured. To adapt the definition to linguistic diversity, "plant and animal life" must be replaced by *languages present* and "species" by *languages*. The last part can be ignored for the moment. The study of linguistic diversity in Canada is the study of the variety of (different) languages that are spoken in Canada. Therefore, one possibility to measure linguistic diversity is the number of languages. This measure is the most simple one, which counts the languages in a given area. Analogues to the measure that counts species, the species richness, the measure is called linguistic richness.

Nettle (1999) reflected linguistic diversity on different levels through time, the first, the *language diversity* refers to linguistic richness. Furthermore, phylogenetic diversity - "the number

of different lineages of languages found in an area" - and structural diversity - to what extent the "languages vary on [some] parameter"- are differentiated. Focusing on the first type, relevant for this study, Nettle identifies the ecological regime as most important factor for linguistic richness in subsistence economies; coming to the conclusion that language endangerment is connected to the modern economy.

For the quantitative analysis of linguistic diversity, linguistic richness is the most basic approach. But due to the limitations of this measurement, the size of the languages (abundance), mainly defined by speaker numbers, is taken into consideration as an additional factor; more precisely the relative abundance, which shows the linguistic evenness through the distribution of languages in an area. On this basis Harmon and Loh (2010) created the Index of Linguistic Diversity (ILD) with the purpose of measuring how linguistic diversity changed over time. In their study, they applied the measurement on data from 1970 to 2005 on a global scale and a macroregional level. Their results showed that linguistic diversity decreased by 20% globally. In the Americas, the biggest language loss was seen with a decline of 64%. Only in Eurasia a positive evolution of linguistic diversity is visible in this period. More information on the measures that are also applied in this study can be found in Chapter 5, including the mathematical expressions.

The presented Index of Linguistic Diversity is not to be confused with the Linguistic Diversity Index (LDI) defined by Greenberg (1956). The Linguistic Diversity Index is also computed on the basis of linguistic richness and linguistic evenness. Linguistic diversity is here defined as the probability that two people (of a group) do not speak the same language. Lieberson (1964) extended the Linguistic Diversity Index to measure the "linguistic diversity or communication between two or more spatially delineated populations or between socially defined subpopulations of a larger aggregate." This equals the cross-product of the relative abundances of the languages for two groups. It will give the probability that two people of different groups speak the same language and can therefore communicate. This would make it possible to see how good people can communicate within and across their groups. This measure of linguistic similarity can be transformed into a measure of linguistic diversity if the similarity is subtracted from one.

The critique of linguistic richness as an indicator for linguistic diversity is based on the overemphasis of languages with a small speaker community. At the same time, the Linguistic Diversity Index by Greenberg (1956) is criticized, since it favors dominant languages and small languages have quasi no effect. The alternative is to measure the Linguistic Diversity through Entropy, as it is based on the balance of languages. The more the languages are evenly distributed, the higher the entropy value and the linguistic diversity. Entropy was developed in physics and

was adapted by Shannon (1948) for information theory. Gullifer and Titone (2020) "proffer language entropy as a means to characterize individual differences in bilingual (and multilingual) language experience related to the social diversity of language use." Grin and Fürst (2022) are using the multilevel approach of alpha (α), beta (β) and gamma (γ) diversity inspired by biodiversity studies from Whittaker (1972). Alpha diversity is the diversity within a defined area and beta diversity refers to the diversity between the areas. Gamma diversity is then the total diversity resulting from alpha and beta. With this approach, the linguistic diversity can be compared within and among regions, as well as in combination. Grin and Fürst (2022) are computing the levels on the basis of Shannon's entropy, which was transformed for easier interpretation and better comparability, in the equivalent number following the Hill numbers framework (Hill, 1973).

In his study, Essfors (2025) quantified global linguistic diversity using the Leinster-Cobbold framework, thereby extending the established concept of hill numbers (number equivalents for linguistic richness, for Greenberg's index, and Shannon's entropy) to also incorporate similarity-aware metrics.

An introduction to the topic of linguistic diversity is also provided by Hammarström (2016) and Sallabank (2024). Other approaches to measure linguistic diversity, like the use of a generalized linear model by Paolillo and Das (2006), are not reviewed in this study.

4.2 Linguistic Diversity and Endangerment

Languages are globally endangered and at risk of being lost. According to Simons (2019), a language disappears on average every 40 days. As a result of language shift, when language is not passed on between generations or speakers "abandon their original vernacular language in favor of another" (Kandler & Unger, 2023), languages are threatened, and when there are no longer any native speakers, the language is considered dormant or extinct. If the last speaker of a language dies, the language dies with them. Before that, normally a phase of multilingualism is visible. Language death is not irreversible; languages can continue to be used as a second language. Nor are endangered languages automatically doomed to extinction. The size of the language community is also not a definitive indicator of language endangerment. There are small language communities whose language is stable and not endangered, and at the same time there are languages with comparatively many speakers that are nevertheless endangered because they are not being passed on to the next generation. Similarly, language contact, e.g. due to migration, does not lead automatically to language shift, but rather people shift to another language because of the advantages they perceive. (Hammarström, 2016;

Kandler & Unger, 2023; Paolillo & Das, 2006; Sallabank, 2024) To draw attention to the general threat to (indigenous) languages, UNESCO has proclaimed the aforementioned International Decade of Indigenous Languages 2022-2032.

In order to classify the degree of endangerment, scales were created that can be used to measure the threat, and thus the stability or vitality of languages. In its report *Language Vitality and Endangerment*, UNESCO Ad Hoc Expert Group on Endangered Languages (2003) has already emphasized the importance of linguistic diversity and presented a scale that ranks language vitality based on intergenerational language transmission. The scale has six levels, or here, including a subcategory, seven levels, ranging from safe to extinct. See also chapter 5 and table 5.1.

A second widely cited scale is the *Expanded Graded Intergenerational Disruption Scale* (EGIDS) by Lewis and Simons (2010), which is based on the *Graded Intergenerational Disruption Scale* (GIDS) by Fishman (1991). The GIDS comprises eight levels that range from stable to isolated speakers. The UNESCO scale was also consulted when creating the EGIDS. Lewis and Simons (2010) propose a scale that focuses on what the individual levels mean for institutions and possible revitalization. Languages can be assigned to one of the 13 levels using clearer definitions:

> "A language can be evaluated in terms of the EGIDS by answering five key questions regarding the identity function, vehicularity, state of intergenerational language transmission, literacy acquisition status, and a societal profile of generational language use" (Lewis & Simons, 2010).

The EGIDS has also made sure to be compatible with *Ethnologue's* five language vitality categories, although these are based more on the size of a language's speaker community and less on intergenerational transmission of the language. The categories are living, second language only, nearly extinct, dormant, and extinct. (Lewis & Simons, 2010) Since 2013, *Ethnologue* has been using the EGIDS to indicate the language status. Bickford, Lewis, and Simons (2015) adapted the EGIDS for sign languages. Zariquiey et al. (2022) compared different endangerment databases with their scales (including UNESCO and EGIDS) and showed that endangerment terminology is not used uniformly and therefore the degree of endangerment can differ.

Simons (2019) used the EGIDS to recreate the status of doomed or extinct languages of the past 200 years, predicting the percentage of lost languages by 2095 for each country. For Canada they predict that 67 to 80% will be extinct in 70 years, aligning with the number of endangered languages today. It can also be seen that the languages started to be doomed approximately in the last 100 years.

Loh and Harmon (2014) applied the IUCN Red List Scale (IUCN, 2012) to languages in order to compare language endangerment with species endangerment. Similarly, Sutherland

(2003) used this approach to compare global linguistic diversity with biodiversity. The number of languages increases with the size of the area, as well as in forest or mountain areas. A higher linguistic diversity can be observed closer to the equator. However, no correlation was found with islands, the season length, the GDP or television coverage. The time of settlement has, furthermore, no effect, suggesting that languages evolve more quickly and the "reasons for extinction risk differ between cultural and biological diversity". (Sutherland, 2003)

In ecology studies, Butchart et al. (2007, 2004) developed and improved the Red List Index (RLI), based on the IUCN Red List, as an indicator for the rate of biodiversity loss. Baumann, Lenzner, Fellner, and Essl (2024) and Zeh (2025) used this mathematical approach to quantify the change of language status applying the Red List Index method to the EGIDS. Baumann et al. (2024) showed that the duration of colonizations has a negative effect on the RLI and therefore promotes the threat on languages. Zeh (2025) did a correlation analysis with digitization indices.

Furthermore, Kandler and Unger (2023) are using a mathematical modeling approach on language shift. They replicated the historic spatial change in language shift in Scotland between the competing languages English and Gaelic. They showed that cultural and demographic factors, as well as the language status, are important and that bilingualism is a "temporary transition state". This approach could indicate which language revitalization programs are effective, emphasizing that stable bilingualism and social domains, where the endangered language is used, are important to preserve the endangered language, to not lose the benefits from the dominant language, and therefore stop a total language shift. Bouckaert et al. (2022) focused on phylogenetic diversity, and used historical knowledge on diversification of languages to build a global tree and to find *evolutionarily distinct globally endangered* languages.

4.3 Global Linguistic Diversity

Ethnologue is listing more than 7000 living languages. The most used languages are English, Mandarin and Hindi. The languages with the most first language speakers are Mandarin, Spanish and English. The speakers of the languages are unevenly distributed. Whereas Mandarin is mostly concentrated to China, English speakers can be found all over the world. (Eberhard et al., 2025a) Also the languages are unevenly distributed over the globe. Essfors (2025) shows that more than 80% of all languages originate in Africa, Papunesia and Eurasia. The most language families can be found in the macroregion Papunesia.

Linguistic diversity changes over time due to various factors. For example, power dynamics

influence the value of individual languages and lead to language shift (Sallabank, 2024). Gavin et al. (2013) gives an overview and review on linguistic diversity and patterns behind linguistic diversity. In general a negative relation of latitude is found for linguistic diversity worldwide, see, e.g., Gavin and Stepp (2014); Mace and Pagel (1995); Nettle (1999), at the same time, a negative correlation between language richness and language range area was observed. This means it is in conformity to the Rapoport's rule, which has the same patterns in biodiversity and species diversity context.

Axelsen and Manrubia (2014) analyzed linguistic diversity and environmental factors, finding that seclusion or connectivity are linked with linguistic diversity. The variables *river density* and *landscape roughness* are significant. Global predictors on language endangerment and therefore threat on linguistic diversity are examined by Bromham et al. (2022). They found a correlation with road density and education. Depending on the environment, the strategies of living are different, which has an influence on the linguistic diversity (Nettle, 1999).

Currie and Mace (2009) looked into cultural patterns and linguistic diversity. It appears that political complexity is the most significant variable, also associated with latitude and language area. This suggests that political complexity forms larger language areas and that environmental factors are less important. Furthermore, no direct effect of biological diversity on cultural diversity can be seen, even though biodiversity and linguistic diversity can overlap. Hua, Greenhill, Cardillo, Schneemann, and Bromham (2019) analyzed ecological patterns with linguistic diversity and found a correlation with climate, which is more important than landscape features. They trace the same behavior of biodiversity and linguistic diversity to the covariance with climate.

Other studies on linguistic diversity looked into the relationship between the spread of infectious diseases and linguistic diversity (Fincher & Thornhill, 2008) or discussed linguistic diversity and cultural heritage in the digital age (Hutson, Ellsworth, & Ellsworth, 2024).

4.4 Linguistic Diversity in Canada

Concerning linguistic diversity in Canada, there are mostly studies with a focus on the bilingual English-French situation and in the geographical context of the province of Quebec. For instance, Castonguay (2019) analyzes the efficiency of Bill 101 using diachronic census data from 1971 to 2016, finding that French declines drastically and that English has a small increase in the province. Especially Montreal is less francophone than the language policies of Quebec would like it to be, suggesting that the francization through Bill 101 did not succeed and French is still not in a stable position. Arsenault Morin and Geloso (2020), however, in their bilingual study

for Quebec show that the decline of French is not linked to an increase of English, but to an increase of multilingualism. Also, this study is based on the census data. The multilingualism can be seen positively, since it is in accordance with the national language policy.

Statistics Canada is analyzing the status of the languages in Canada through the census and publishing the results, e.g. Statistics Canada (2025a). Notable is the *Megatrend* from 2018 (Statistics Canada, 2018b) that analyzed the linguistic evenness of English, French, and non-official languages from 1901 to 2016. Furthermore, the abundance of immigrant languages shows that since 1971 the immigration diversified and with it also the languages present in Canada. Especially Chinese is spoken by a large population (1.3 million in 2016).

Another focus of linguistic diversity studies in Canada are immigrant languages. Pendakur (1990) did a quantitative analysis with census data using the *Kralt's index of language transfer*. He found that in Canada, language shift to the dominant language is a generational phenomenon. Furthermore, a comparison of the linguistic diversity of the three biggest cities - Montreal, Toronto and Vancouver - reveals that Montreal is the most overall diverse city, but Toronto is the most diverse city, if only immigrant languages are considered. Pendakur (1990) calculates the linguistic diversity based on the Lieberson's index, which is equivalent to the one by Greenberg described in chapter 5.

Studies related to linguistic diversity in Canada also address the endangerment of Indigenous languages. For instance, Boissonneault, Tallman, Gast, and Greenhill (2025) project dormancy risk for indigenous languages for the 21st century and predict a huge decrease of speaker numbers. In 2121, 99 percent of all speakers will likely covered by nine languages out of 27.

Canada served for case studies and as an example for the application of the method of linguistic diversity, i.e. entropy. Gullifer and Titone (2020) used Shannon's entropy to compare the linguistic diversity of Montreal, Quebec, and Canada. It shows that the city of Montreal is more diverse than the province of Quebec but less diverse than the country of Canada. Quebec reflects the linguistic diversity of Montreal, for the language groups, but with a larger speaker number of French. Canada, however, is more diverse showing a more balanced situation between English, French and immigrant languages plus other languages, ca. 10%, 62% and 22% vs. 55 %, 20%, 20%. Grin and Fürst (2022) are applying their multilevel approach to the real-life example of the three biggest Canadian cities. This approach too agrees that Montreal is linguistically most diverse. This is true for all measurements, but particularly visible in the gamma diversity (total diversity of the city including the diversity within the electoral districts and between them). This outcome is overall not surprising knowing of the bilingual history of the city and its importance in immigration history.

5 Method

All measurements are applied to the entirety of all languages in relation to the population. The population number is based on the population estimates published by Statistics Canada (2025b). Also, with the given data, the linguistic diversity is analyzed across the groups of official, Indigenous and immigrant languages, including the sum of all these groups and the sum of official and Indigenous languages as local languages. The official languages of Canada are English and French; Indigenous languages are all languages native to Canada according to *Ethnologue*. *Ethnologue* (Eberhard, Simons, & Fennig, 2025b) maps each language to a primary country, therefore Indigenous languages might not appear in the list if they were assigned to another country. Immigrant languages are all other languages that are not classified as official or Indigenous. This study focuses on the country of Canada as a whole; therefore, only methods used to analyze the linguistic diversity in this study are presented in more detail in this chapter. Methods intended for spatial comparison are not described. If not explicitly mentioned otherwise in the method (Index of Linguistic Diversity), the diachronic analysis is made by comparing the results for the time span.

5.1 Linguistic Richness and Evenness

Linguistic richness is the "number of different languages in a given geographical area" (Nettle, 1999). Therefore, it is the absolute frequency of languages in Canada and simply the count of those languages identified as distinct through the ISO code and *Ethnologue*. To analyze the richness solely individual languages, no macrolanguages, and only living languages, no extinct languages, are included.

As a measure of the abundance of the language groups, the speaker number is taken. This shows the size of the language through the number of individuals using the language.

To set those two absolute measures in relation, the "distribution of individual speakers among languages" is defined by Harmon and Loh (2010) as the linguistic evenness. This is important because even with the same number of languages, linguistic diversity can be quite different depending on whether a few languages are more dominant or whether speaker numbers are

quite equally distributed. In general, linguistic diversity is considered to be higher the more balanced the distribution of speakers is. For this study linguistic evenness is analyzed, only by the relative abundance of speakers for the language groups official, Indigenous and immigrant languages. In addition, on language level the years 1996/2001 and 2021 are compared.

5.2 Index of Linguistic Diversity

Harmon and Loh (2010) created based on a combination of linguistic richness and linguistic evenness the Index of Linguistic Diversity (ILD). The "goal of the index is to measure [..] changes in the relative distribution of first-language speakers among discrete languages within the total population" (Harmon & Loh, 2010).

Starting on a chosen year as a baseline, the change relative to this year is calculated through the cumulative product of the average change of the log-ratio of consecutive years of user fractions with the base year set to one.

So, first, the user fractions F_{ly} for each language per year had to be calculated with N_{ly}, the number of speakers of the language per year, and P_y, the population per year. The population is either the population estimate of Canada or the sum of the specific language group in question.

$$F_{ly} = \frac{(N_{ly} + 1)}{P_y} \tag{5.1}$$

Second, the log-ratio d_y of the fractions of the consecutive years F_{ly} and $F_{l(y-1)}$ is computed for each language:

$$d_y = \log_{10}\left(\frac{F_{ly}}{F_{l(y-1)}}\right) \tag{5.2}$$

The average change $\bar{d_y}$ is calculated by the arithmetic mean of the ratios. Finally, the Index of Linguistic Diversity I_y is the cumulative product of the anti-logged means and the starting year set to 1.

$$I_y = I_{y-1} 10^{\bar{d_y}} \tag{5.3}$$

In this measure, each language has an equal weight and the distance between languages is not taken into account. If the index declines, it means that languages are getting less evenly distributed, but more skewed. In other words, some languages become more dominant while for others the relative speaker number decreases. Therefore, the opposite happens when the index increases: linguistic evenness increases alongside linguistic diversity. The index of linguistic diversity is always relative to the starting year and just compares the years. It is not possible to

make predictions with this index but only to observe the trend. In any case, it should not be confused with the Linguistic Diversity Index by Greenberg described in the following section.

5.3 Linguistic Diversity Index

The Linguistic Diversity Index (LDI) by Greenberg (1956) was also inspired by diversity studies in ecology and applied their methods on linguistic diversity. In ecology, the Simpson Index (Simpson, 1949) describes as a proxy of diversity the probability that two individuals belong to the same species, or its variation, the Gini-Simpson Index, where the probability is measured that the two individuals chosen at random do not belong to the same species (Fedor & Zvaríková, 2019). Therefore, Greenberg's definition applied to the linguistic use case is the following:

> "If from a given area we choose two members of the population at random, the probability that these two individuals speak the same language can be considered a measure of its linguistic diversity. [...] Since we are measuring diversity rather than uniformity, this measure may be subtracted from 1" (Greenberg, 1956).

This is equivalent to the Gini-Simpson Index and Greenberg (1956) has formulated it as:

$$A = 1 - \sum_i i^2 \tag{5.4}$$

It is one minus the sum of all squared probabilities i. The probability for one person to speak a language is the relative abundance of that language, and for two people it is the product of these abundances. The index can have values between 0 and 1; where one means highest possible diversity and zero means no diversity. This measurement favors dominant languages, and small languages have quasi no effect on the diversity. In this study, the presented non-weighted method is used for simplicity. Greenberg also proposed a more complex, weighted method, which also takes interlinguistic distance into account. These methods were created for a monolingual context. However, Greenberg (1956) expands in his paper *The Measure of Linguistic Diversity* also to methods for polylingual contexts which are not further discussed here.

5.4 Shannon Entropy

Grin and Fürst (2022) are using Shannon's entropy as a diversity measure, as it "is less sensitive than richness, but more sensitive than the Greenberg index" (Grin & Fürst, 2022) regarding to

languages with small speaker numbers. For this study, the level approach (described earlier) is not followed as Canada is viewed as a unit without further geographical sections.

Shannon entropy S is the negative sum of the probability p of language i times the natural logarithm of the probability of the language and can be expressed as follows:

$$S = -\sum_{i=1}^{R} p_i \times \ln(p_i) \tag{5.5}$$

with R for linguistic richness. Shannon entropy can be used to measure how uncertain is the linguistic identity of a person in a community. This means that a higher entropy results in a higher uncertainty, and it is less predictable which language a person speaks. However, low entropy means that it is not very surprising which language a random person speaks because one language is highly dominant. The lowest possible entropy with the value zero is given in an area with only one language. The highest entropy, where all languages are perfectly distributed, is $logR$ (Gullifer & Titone, 2020).

5.5 Hill Numbers

It its not possible to compare the presented approaches to measure (linguistic) diversity. Hill (1973) therefore established a concept that makes it possible to compare the diversity measures of richness, Shannon's Entropy and Simpson's Index through an effective number of elements (here languages). "The numbers equivalent of a diversity index is the number of equally likely elements needed to produce the given value of the diversity index" (Jost, 2007).

The numbers equivalent (also "Hill Numbers") can be described through the following formula:

$$Z = \left(\sum_{i=1}^{R} p_i^q\right)^{\frac{1}{1-q}} \tag{5.6}$$

Z is the numbers equivalent, R the richness (number of languages), p the relative abundance of the language i. q represents the "*order* of the diversity" (Jost, 2006).

> "The order of a diversity indicates its sensitivity to common and rare species. The diversity of order zero (q = 0) is completely insensitive to species frequencies and is better known as species richness. All values of q less than unity give diversities that disproportionately favor rare species, while all values of q greater than unity disproportionately favor the most common species (Tsallis 2001, Keylock 2005).

> The critical point that weighs all species by their frequency, without favoring either common or rare species, occurs when q = 1" (Jost, 2006).

For richness, the *True diversity*, as it is called by Jost (2006), is represented by the numbers equivalent of order zero and is just the count of languages.

$$^{R}Z = R = \sum_{i=1}^{R} p_i^0 \tag{5.7}$$

For Greenberg's Linguistic Diversity Index and Shannon's entropy, however, calculations are needed to compute the Hill numbers. As described above, the Linguistic Diversity Index describes the concentration and is quite insensitive to small languages. This is represented by an order 2 diversity $q = 2$:

$$^{G}Z = \frac{1}{1-G} = \frac{1}{\sum_{i=1}^{R} p_i^q} \tag{5.8}$$

with G as Greenberg's Linguistic Diversity Index. This is also equivalent to Inverse-Simpson.

Shannon's entropy S, as a balanced measure between richness and the Linguistic Diversity Index, uses the order one ($q = 1$) for the transformation to the Hill number:

$$^{S}Z = \exp(S) = \exp\left(-\sum_{i=1}^{R} p_i \times \ln(p_i)\right) \tag{5.9}$$

All expressions in this chapter are the formulations of Grin and Fürst (2022). They are using for the case of linguistic diversity, the numbers equivalent of order one, derived through Shannon entropy, to have a more intuitive way to calculate linguistic diversity.

5.6 Red List Index

To quantify the extinction risk and see trends in the status of biodiversity the Red List Index (RLI) was developed and improved by Butchart et al. (2007; 2004).

The RLI is one minus the normalized average threat score (Baumann et al., 2024; Butchart et al., 2007):

$$RLI_t = 1 - \frac{\sum_S W_{c(t,S)}}{W_{EX} \cdot N} \tag{5.10}$$

W is the the threat level and W_{EX} the maximum score, in other words, for this case the level for extinct languages. N is therefore here the number of all languages, including extinct languages. The Red List Index is between one and zero with one indicating that all languages are stable and zero that all languages are extinct.

The RLI was originally applied on the Red List, which primarily covers in more detail the endangered status. The scale developed by UNESCO is quite similar and does not differentiate the status in general for living languages that are not considered threatened. The EGIDS, however, is much more precise also for living languages. As it is the language status scale used in *Ethnologue* starting in 2013, the EGIDS can just serve as the underlying basis for the calculation of the RLI for the second half of the period analyzed in this study. Before this, *Ethnologue* had provided much simpler information on the language status as it is also described by Lewis and Simons (2010) and in chapter 4. However, for Canada only the categories living, nearly extinct and extinct are present in the editions before the establishment of the EGIDS. For the purpose of this study and to be able to analyze the Red List Index for the whole time frame, the status were clustered into the following categories: 'living', 'endangered', and 'extinct'; as well as 'unestablished' and 'unknown'. 'Unestablished' just occurs in the edition of 2024 in this study. 'Unknown' is applied to all languages present in the edition with no further information on the status. This is mainly the case for immigrant languages that do not appear on the language list in *Ethnologue*, but are listed in the summary. The table 5.1 shows the different scales mentioned as well as the cluster composition for this study.

For the computation of the RLI the status scales which are ordinal have to be transformed to be quantitative and are therefore assigned with weights. According to the definition that the most stable status has the weight zero, the category 'living' corresponds to zero. Since 'unestablished' and 'unknown' can turn into any status, but most likely not extinct, as these status were assigned to immigrant languages by *Ethnologue*, they have the weight one in this study. This results in the weights two for 'endangered' and three for 'extinct'. (See table **??**.)

Cluster	Weight
'living'	0
'unestablished'	1
'unknown'	1
'endangered'	2
'extinct'	3

Table 5.2: This table shows the weights needed for the RLI assigned to the Cluster.

For the more detailed RLI based on the EGIDS starting in 2013, the levels of the EGIDS were ranked and assigned to the weights 0 to 12. 'unestablished' and 'unknown', for the same reasons as above, are set between the 'living' and the 'endangered' levels,not as an additional

weight, but at six; see table 5.3. This also represents the median of the weights and can therefore be seen as the middle. The languages of ’unknown’ or ’unestablished’ status are not super stable, but more likely to be living. They can take from there potentially every threat level.

EGIDS Level	Weight
’0’	0
’1’	1
’2’	2
’3’	3
’4’	4
’5’	5
’6a’	6
’6b’	7
’7’	8
’8a’	9
’8b’	10
’9’	11
’10’	12
’unestablished’	6
’unknown’	6

Table 5.3: This table shows the weights needed for the RLI assigned to the EGIDS.

IUCN Red List	EGIDS	UNESCO	Ethnologue	Cluster
Least Concern (LC)	0: International	Safe	Living	Living
	1: National			
	2: Regional			
	3: Trade			
	4: Educational			
	5: Written			
	6a: Vigorous			
Near Threatened (NT)	6b: Threatened	Stable yet Threatened		Endangered
Vulnerable (VU)		Vulnerable		
Endangered (EN)	7: Shifting	Definitely endangered		
	8a: Moribund	Severely endangered		
Critically Endangered (CR)	8b: Nearly Extinct	Critically endangered	Nearly Extinct	
Extinct in the Wild (EW)	9: Dormant	Extinct	Dormant	Extinct
Extinct (EX)	10: Extinct		Extinct	

Table 5.1: This Table shows the endangerment Scales: IUCN Red List (used for Biodiversity), EGIDS (used by *Ethnologue* since 2013), UNESCO (Framework provided by UNESCO to assess language status and vitality), *Ethnologue* (status categories used before the EGIDS) and the clustered status used for this study for the whole time frame. Note that *Ethnologue* also had the Category 'Second Language Only', which is not present for Canada, and cannot be assigned to the EGIDS or the other Scales, as status is in general based on intergenerational transmission and and on the first language. 'Only Second Language' would refer to dormant. Table adapted from (Loh & Harmon, 2014).

6 Workflow

This chapter is the documentation of the workflow. It encompasses data gathering, cleaning, and preparation for the analysis. The analysis itself is described in chapter 7. An overview of the workflow can also be seen in the figures 6.1 and 6.2.

6.1 Scraping

The quantitative component of my study is grounded in language speaker data for Canada, drawn from the *Ethnologue* editions 13 to 22 and 27 (corresponding to publishing years: 1996, 2000, 2005, 2009, 2013, 2015, 2016, 2017, 2018, 2019 and 2024). This selection is based on accessibility constraints: the data from editions 13 to 22 are available on the Internet Archive and can be accessed via the Wayback Machine. The most recent 27th edition (at the time of data collection) is available through institutional access on the official *Ethnologue* website. [1] To extract the relevant data, web scraping techniques were employed. This scraping workflow was implemented by using Python 3.13 with its build-in-functions and the following libraries: `pandas` (for data manipulation, pandas development team (2025)), `pycountry` (for standardized country codes, Theune (2024)), `requests` and `bs4` (for HTTP requests and HTML parsing, Reitz (2025); Richardson (2007)), `selenium` (for handling dynamic content, ThoughtWorks (2025)), `waybackpy` (for interfacing with the Wayback Machine, Mahanty (2022)), as well as `time` and `typing` (for timing operations and type safety, respectively). The scraping pipeline was designed in a modular way, ensuring reusability for different countries and editions. This ensures not only transparency and reproducibility but also scalability of the research framework.

The final module of the data collection pipeline is the function `scrape_country_year (country, year)` where the year represents the year of the edition, which is more relevant for the interpretation of the data than the edition number itself since the editions were not published at regular intervals. This approach also makes sense for those unfamiliar with *Eth-*

[1] Now all editions (13 to 28) are available directly in the archive on the *Ethnologue* website. To scrape the data from there, the methods can be easily adapted. Mainly the URLs change, but not the HTML code, also the years have to be dealt with to be able to scrape not only the editions until 2019.

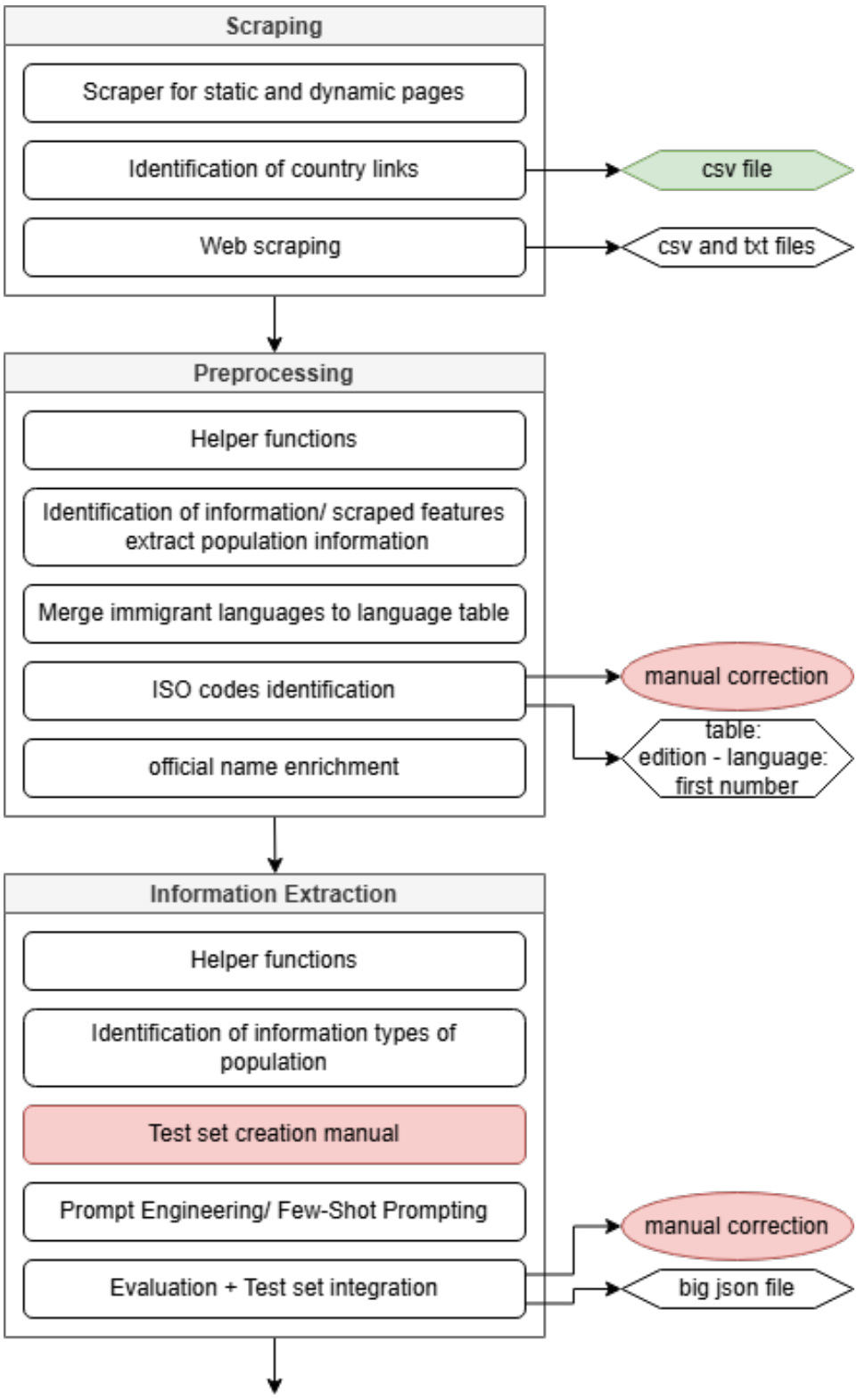

Figure 6.1: This figure shows the first part of the workflow for the preparation of the dataset. The flowchart continues on the next page. Green indicates final data files, red marked fields are manual tasks.

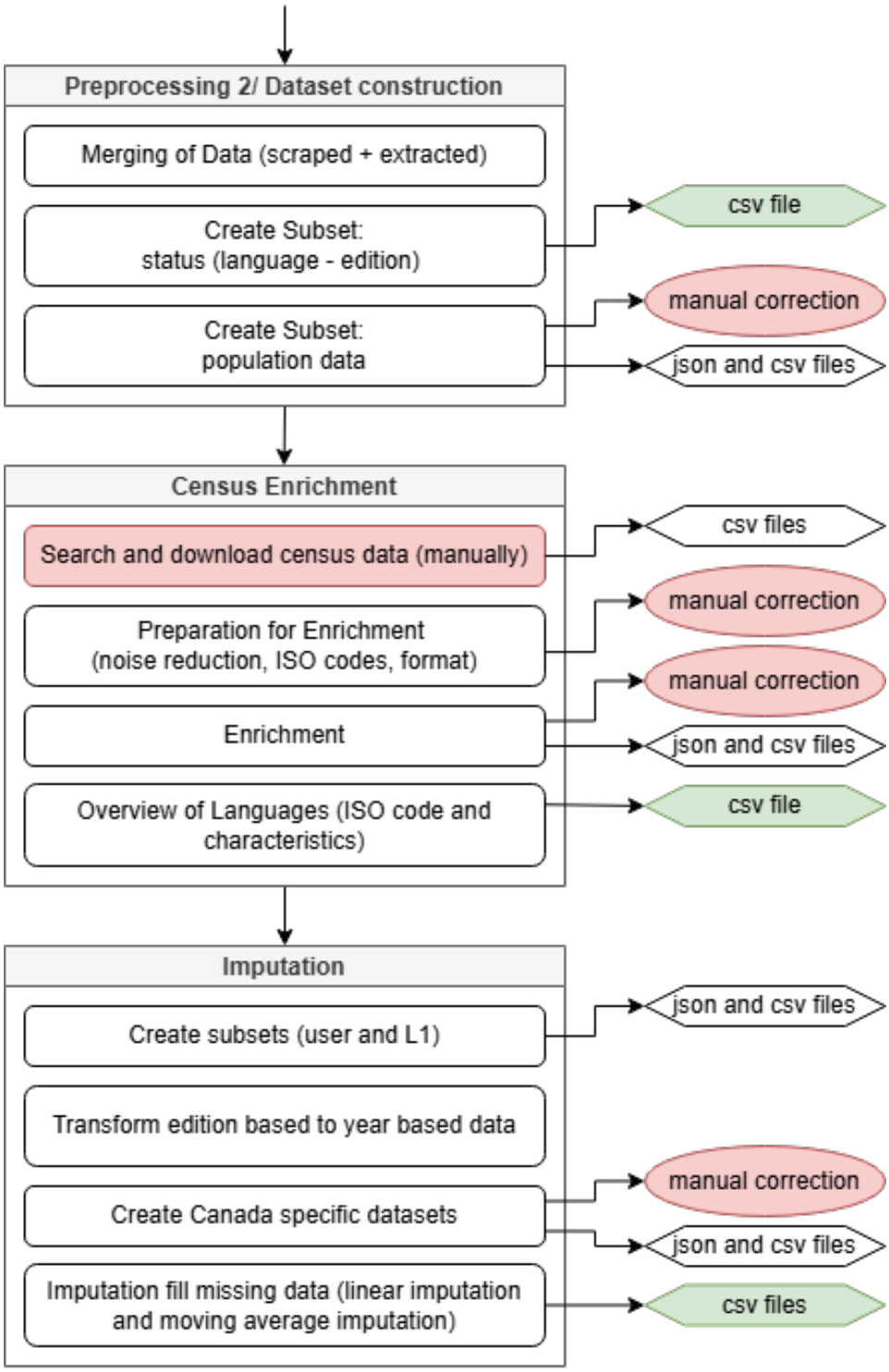

Figure 6.2: This is the continuation of the flowchart for the workflow for the preparation of the dataset. Green indicates final data files, red marked fields are manual tasks. The workflow is followed by the analysis.

nologue and its internal organization. The function integrates all necessary operations required for the automated retrieval and processing of language speaker data for a given country in a given year.

There are two kinds of websites, static and dynamic pages. All pages relying on the Wayback Machine are static; however, the *Ethnologue* webpage with the most recent edition is dynamic and behind a paywall, which can be accessed via login with the institutional account through the university library of Vienna. To accommodate this difference in content structure and access protocols, the scraping pipeline is divided into two dedicated modules for page retrieval. The first module, `get_static_page(url)`, is responsible for fetching static content. It sends a request to the target URL using the `requests` library and returns a parsed `BeautifulSoup` object containing the raw HTML. This approach is efficient and straightforward for archived pages hosted on the Internet Archive. The second module, `get_dynamic_page(url)`, is designed to handle dynamic content that requires JavaScript rendering and user authentication. It leverages the `webdriver_manager` from the `selenium` library to automate browser interaction. The `time` module is also used to introduce appropriate delays, ensuring that the JavaScript elements of the page are fully loaded before extraction. This setup allows the pipeline to simulate a logged-in user session and retrieve the complete rendered content as a `BeautifulSoup` object, making it suitable for processing alongside the static pages.

As a pre-step of the actual scraping process, a `.csv` file was created including information on which countries are listed in which editions and contain scrapable data. It also contains the necessary components for the creation of the URL for the scraping process. The file was created via web scraping of the overview pages of the editions and stored as a `.csv` file. This information was automatically retrieved using the two functions mentioned before, whereas the URLs were looked up manually. The required information was then extracted with the `BeautifulSoup` functions and stored and organized in a `pandas DataFrame`. With the library `pycountry`, the ISO 3166-1 alpha-2 code was retrieved for the countries found in the data. For cross-checking, the country codes were verified to see whether they correspond to the country name used by *Ethnologue*. To get a standardized name, another column was added with the country name retranslated from the ISO codes. Finally, the `.csv` file was corrected manually and missing names and ISO codes were added. This file is also important in case someone wants to look up the official country name that should be used as string input for the variable "country" in the final method.

With the country file, the URL can be generated (`get_url(year, country)`), which is the first step in the scraping process and, if necessary, the URL for the Wayback Machine is also created (`get_wayback_url(url, year)`). For this step, the `waybackpy` library is used to get

the timestamps necessary for the URL. This depends on the year of the edition in question. Once the required URL is identified, the website can be scraped with the `get_static_page` and `get_dynamic_page` functions, respectively, and the `BeautifulSoup` object(s) can then be further processed.

There are four different HTML website structures in the corpus of editions. At the HTML level, the editions can be grouped into 1996; 2000 to 2009; 2013 to 2019 and the current edition, here 2024. The HTML structure should not be confused with the appearance of the website, which would result in different groupings of editions. Each of the four groups has therefore its own function that combines scraping and information retrieval with an output of two files: one for the summary of the country data (`.txt` or `.csv`) and one for the languages (`.csv`). The files represent the HTML structure.

In general, the HTML structure of the edition of 1996 is quite simple, consisting of paragraphs of texts; therefore, the information on the language is not further structured through the HTML. The Editions 2000 to 2009 are using blockquotes for the country summary and a table structure for the language information, in which a language is represented as a row. For the editions 2013 to 2019 the information can be found on two websites, one containing the country summary and the other one information on the languages. It uses divisions for structuring the code. The edition of 2024 uses mainly sections, descriptions, and unordered lists for the structure. To get a more detailed overview on the HTML structures, see chapter 10.3. The available information varies slightly between the editions, which will be addressed in the next step: the preprocessing.

6.2 Preprocessing

Once the data are scraped from the website, a first preprocessing step has to be done to prepare for the information extraction. In this step, in addition to the already introduced libraries such as `pandas` and `pycountry`, the following libraries were used: `re` (for regular expression operations), `langcodes` (for ISO code identification of languages, Speer (2024)), and `unicodedata` (to get access to the Unicode Character Database for character normalization). Some base functions like `save_csv` and `read_csv` are created that are customized for the characteristics of the data, specifically the encoding and indexing. Likewise, a function `save_json` was needed. All of them work with a `pandas DataFrame`.

In a first step, it is helpful to extract a subset of the data frame with the scraped data, which contains the relevant information. The relevant information is the language name, the abbreviation or ISO code, the details of the language, and the status. From this point on, the ISO code

should serve as the identifier for the language. The subsets are slightly different depending on the edition, and the columns are renamed during the same process to create consistency in the naming. The column "details" contains information on the population using the language, and further information like the location where a language is spoken or located in Canada. With simple regular expressions, the column "details" can be split according to the type of information into two columns "population" and "details_info". The expression depends on the edition and its text structure. Since in this point the text structure is consistent in each edition, the splitting can be done this way. See table 6.1 for an overview of all possible information for each language entry of the language list.

Feature	Description
name	Primary name used by *Ethnologue*
iso-code	ISO 639-3 code
abbreviation	Abbreviation given by *Ethnologue* (editions 1996 and 200)
details	Editions 2000-2018: information on population and other information like location Editions 2019-2024: other information like location
population	Information on speaker numbers Edition 2024: renamed from users
details-info	Other information like location Editions 2019-2024 see details
status	Status of language Until 2009: living, endangered, extinct After 2013: EGIDS
alternate-names	Other names for the language
autonym	Autonyms of speakers for the language
classification	Linguistic classification (language family branch)
comments	Further information (like specifics for missionary work)
dialects	Dialects of the language

Table 6.1: This table shows the information available in *Ethnologue* for the languages in the country specific language list. Not all features are available in all editions or for all languages.

For a first overview of the data and to get a feeling of it, the first number in the population column was looked up, again with regular expressions. The first number is mostly what can be classified later as "user", "L1" [2] or in some cases "ethnic population". An explanation of these labels can be found in the next chapter.

The immigrant languages are not included in the list of languages (except for edition 2024), but are in the summary table or text of the edition. Therefore, they have to be added to the lan-

[2] L1 for first language, also known as mother tongue or native language.

guage data frame in the second step of the preprocessing. For the immigrant languages there is only the name listed with mostly a undefinied speaker number and rarely with a referenced year. The enrichment process was done semi-automatically, since some of the corresponding strings were copied into the variable for the information extraction. Once the string was ready, copied manually into variable or extracted with pandas and regular expressions from the text_summary or from the summary `.csv` file, the data was transformed into a `DataFrame` with name and user number. The data from one year (edition) could then be merged.

To enable the merging of all editions into one data frame, the missing ISO codes had to be looked up. This concerns the immigrant languages and the editions before 2005. To get the ISO codes for the languages, the find function from the `langcodes` library was used. For better performance, the language names were slightly formatted and normalized, dealing with commas in the name string and normalizing the apostrophes. Nonetheless, manual validation and correction was crucial since it is the basis for linking the languages over time and creating the diachronic database. Helper functions like `get_name_iso_table` and `get_missing_iso_code` support the manual validation and correction. With this, the tables were ready to be merged, and a spreadsheet of the extracted numbers could be created (`get_spreadsheet_editions`).

With all ISO codes and the `pycountry` library, the official name is looked up. Manual validation and correction are still necessary. For example, Tutchone, as a macrolanguage, does not have an ISO code; therefore, both ISO codes of the Tutchone language form the id for the macrolanguage as a list. The language is kept in the dataset only for the sake of completeness. For later analysis, macrolanguages are not relevant. The function `get_missing_official_name` helps with the correction. The language table does not only include individual languages but, depending on the edition, also macrolanguages. The text format for macrolanguages in edition 2009 differs from other editions. It therefore is harmonized by deleting unnecessary elements like line breaks, larger white spaces and the string “More information.”. A small orthographical error in “Amacrolanguage” was corrected to “A macrolanguage”. The first preprocessing was closed by creating a `.csv` file with an overview of the languages in Canada with their ISO codes, names in *Ethnologue* and official name. Also, `.csv` and `.json` files for each edition were created.

6.3 Information Extraction

For the information extraction task, the following libraries were used `os` (for operating with system interfaces), `openai` (for access to the OpenAI API, OpenAI (2025)), `json` (for encoding

and decoding of files in JSON format), `numpy` (for for array computing, Harris et al. (2020)), `statistics` (for mathematical statistic functions) and `copy` (for shallow and deep copy operations). Furthermore, `pandas`, `re` and the functions from `preprocessing` were reused.

In this step, the basic function `open_json` was created to open a `.json` file and get an object in JSON format. Also, the function `csv_to_json` takes a `.csv` file and saves it after transforming it to a `.json` file. The function `create_json` takes an object in JSON format and saves it as a `.json` file.

The relevant information that has to be extracted are the user numbers of the languages. This information can be found in semi-structured short texts in the population column. The information that can be found is the category of user with spatial mapping, as well as the user number and the source of this information, including the year. Not always all information is present for every language. One short text also includes all the information combined for one language and, therefore, often more than one entry. In general, the texts can have information on the user in general or more specific on L1 speaker, L2 speaker, semi-speaker, and monolingual speaker. This information can be specified for Canada. Furthermore, information on L1 speakers of all languages of a macrolanguage can be present; also, for individual languages, this information is sometimes mentioned for the macrolanguage the individual language is part of. Also, information about the ethnic population of the language may be present. For these two categories, the spatial component is not relevant. It is possible that more information is given in the short texts. Examples of short texts of the "population" feature of the years 2000 and 2024 can be found in chapter 10.4.

For the information extraction task, the approach of prompt engineering was pursued. The paid large language models from OpenAI were used via the API from the `openai` library. A first function generates a client from OpenAI that gives access to the Models, this client is then also set as a global variable. The prerequisite is that an API key is in the environment variables. Another global variable contains the prompt format. The prompt is quite detailed to answer all the requirements (see chapter 10.5). It is structured as follows: first, a short summary of the overall task is given, then a longer explanation and guideline on how to perform the task are given, and the exact structure of the output that is wanted was defined. Next, more hands-on specifications for the implementation were given. To improve the output, examples are integrated into the prompt (Aarfi & Ahmed, 2024; Phoenix & Taylor, 2024; Schulhoff et al., 2024). Lastly, the text from which the information has to be extracted is in the prompt.

In addition to few-shot prompting, role prompting technique was applied. With the role "You are an expert in information extraction." to set more weight and focus on the task.

The wanted structure is in JSON format (list of dictionaries), where each instance represents

a language with keys describing the type of information extracted and the value of the information itself. The keys follow the rules defined in the prompt, and the values are extracted from the short texts. The table 6.2 describes the structure of the keys and the values. Information should be extracted for the following categories: "user" (number of users not further specified, or listed as all user), "L1" (first language or mother tongue speakers), "L2" (second language), "semi" (semi-speakers), "monoling" (monolingual speakers). These categories are further specified by the location of the information ("canada" for information explicit for Canada, "all" for not specified or not Canada exclusive, e.g. Canada and US) and the year of the information visible through the mentioned source or the year of the edition with the addition of "ed". Information on "all_languages_L1" (important for macrolanguages) and "ethnic_population" only the year or year of edition is relevant. All further information that does not fit into these categories has to be saved in "extra_info_ed_{year}", which is always edition specific. Like this, ideally, no information is lost during the information extraction process but is saved for later and enables other research purposes.

<table>
<tr><td>user (all, not specified),
L1 (L1 speaker, mother tongue),
L2 (L2 speaker),
semi (semi-speakers),
monoling (monolingual speaker)</td><td>canada (specific for Canada),
all (not specified, not only Canada)</td><td rowspan="2">{year} (year of information)
ed_{year} (year of edition if no source/year indicated)</td></tr>
<tr><td colspan="2">all_languages_L1 (L1 speaker of all languages of a macrolanguage),
ethnic_population (ethnic population of the language)</td></tr>
<tr><td colspan="2">extra_info (all other information for the language in the population feature)</td><td>ed_{year}</td></tr>
</table>

Table 6.2: This table shows the structure of the keys of the information stored in the feature "population"

The examples were created manually together with the test set. For this, random instances (languages) were chosen, and the task was performed manually. The prompt format has variables as place holders for the year of the edition, the examples with output, and the short text of the feature "population".

The information extraction pipeline is modularized. The first function formats the prompt by adding the edition-specific information to the prompt (year and examples with outputs); the underlying prompt format is by default the one from the global variable. The next function creates the messages that have to be given to the LLM, containing the roles and the prompt.

The `completion_output` function builds the chat completions with the messages. The client and the model are optional inputs. By default, the client is the one defined in the global variable, and the model is "gpt-4o". The function also parses the chat response, and the output is in JSON format (a list of dictionaries). The parameters are set as follows: "temperature" is 0, because the output is not allowed to be creative but should follow the defined structure; "frequency_penalty" and "presence_ penalty" are 0, and "stop" is none. Even though a JSON structure is wanted, the parameter "response_format" was not set because also the keys and not only the values had to be extracted from the short texts. Since the information has to be gathered from each short population text of an edition, a function was created that combined the steps of the information extraction and loops over each population text adding the extracted information to the `.json` file of each edition. With each text, the task was given at once. This method led to the best results in a first supervised try-out in which evolution the prompt was improved, and a batched prompt including all population texts at once was rejected. The quality of the data extraction is crucial for the later analysis and, therefore, the better results outweighed higher resources - time and costs - in the performance.

This works for all editions except 1996, because no population text could be extracted in the preprocessing for the information extraction with regex, but it also has to be done with a LLM. The information was in a string containing the information that is also found in the other editions already structured in the HTML code. It concerns the following features named like in the other editions: "population", "details_info", "classification", "dialects", "comments" and "status". Not all languages have information for all these features. The prompt for splitting is set as a global variable, as well as the examples for this task, since they do not change, but for better structure were separated from the prompt itself. The function `split_1996` is splitting the text in the features; functions from the before mentioned information extraction pipeline could be reused and the features were added to the `.json` file. With this split information, also the information regarding the population can be extracted, like for the other editions. The function for 1996 is slightly different because of the splitting and its nested saving structure.

The last step of the information extraction is the evaluation of the performance. For this, the test sets that were manually created together with the examples are needed. Because the test set was created in a different environment, as simple text. It did not follow the JSON structure, to get a coherent structure and correct small inconsistency problems, e.g. L1 was written in lowercase, a test set preparation step had to be added. This was solved by the function `prepare_testset` which combines the function for the structural problem using prompt engineering and the function for the key correction. The corrected test dataset is saved as a `.json` file for each edition and used for evaluation. The evaluation showed that the splitting task for

the 1996 edition could be done perfectly by the LLM and had for all tested instances an accuracy matching score of 100%.

Next, the outcome of the information task was evaluated by simply getting the scores for the percentage of exact matches for one instance (language), and then the average and median for a whole edition and, furthermore, the standard error proportion for the accuracy was calculated. Moreover, the number of languages for which the information was extracted, the number of annotated languages (languages in the test set), the percentage of tested languages per edition, and the number of languages that actually got a score during the evaluation process (some languages were manually annotated and in the test set although they did not have information in "population" and therefore did not go through the information extraction process due to missing data) were added to the evaluation summary.

The average of all average accuracies was 84.5%, whereas the oldest editions have the weakest outcome, see table 10.1, which is not surprising since the text was less standardized back then. A closer look at the mistakes, they seem acceptable. Often errors were made at the question of "extra information", which can be something like the researcher added a period at the end of the phrase and the LLM not. That does not really change things for the purpose, but makes the gold standard different from the LLM response. Also, if in a text, only two different types of information have to be extracted, then if one is wrong, it is already just at 50% for this language. So, the impact of errors depends on the language, and one error can lead to an accuracy of 50% or, for example, 80% in the case of more information available for the language. Therefore, it is not enough to look at the accuracy of a language, or the average accuracy for an edition, but at what mistakes the LLM made. For sure, also in this task the well-known problems of LLMs occurred, for example hallucination and making up categories for the data not allowed, even though this was tried to avoid through the prompt design. The total cost for the use of the LLMs for this information extraction task was 3,62$. This includes all runs from the first try outs to the final information extraction.

For the evaluation of all editions, first, one big `.json` file was created that contains all the extracted and scraped information. Then a function loops through the years and calculates the accuracy scores for each edition separately. Finally, the overviews were combined to obtain the final evaluation summary. For the big `.json` file, first, all files for the editions had to be prepared. All data junk was removed, like keys without values. The keys that show inconsistencies or were no longer suitable were dealt with. For this task, helper functions for renaming and removing keys were built. The `.json` files for each edition were re-enriched with the information from the `.csv` files with the scraped *Ethnologue* information. The editions 1996 and 2000 were identified via the old *Ethnologue* abbreviation and from 2005 on with the ISO code. The edition

specific keys had to be extended by a suffix that refers to the edition. The ISO code is excepted from this as well as keys already indicating information for a specific edition (ending with a digit). With all these preparations, the `.json` files of the editions were merged by the ISO code. To deal with the hallucinating problem from the information extraction, the keys were checked if they follow the structure and if not corrected manually. For this task, a function identifies the errors and 75 keys with their values were manually corrected for the entire data. With the few manual corrections, a second evaluation showed a small increase of the average accuracy of one percentage point, but taking the standard error into account, the lower margin is always over the 60% threshold and therefore acceptable knowing the nature of the error described above. The table 6.3 shows the final evaluation.

Edition	Number Language Entries	Size Test set (Annotated percentage)	Median Accuracy	Average Accuracy	Standard Error
1996	79	17 (21,52 %)	100	80.21	0.19
2000	90	34 (37.78 %)	100	78.95	0.14
2005	89	29 (32.58 %)	100	82.14	0.14
2009	91	31 (34.07 %)	100	88.33	0.12
2013	92	20 (21.74 %)	100	85.83	0.16
2015	94	37 (39.36 %)	100	88.10	0.11
2016	99	22 (22.22 %)	100	88.61	0.14
2017	99	25 (25.25 %)	100	98.00	0.06
2018	101	21 (20.79 %)	100	86.35	0.15
2019	101	23 (22.77 %)	100	84.78	0.15
2024	235	26 (11.06 %)	100	90.73	0.11

Table 6.3: This table shows the results of the information extraction after the corrections

For the remainder of the study, the gold standard of the test set replaces the extracted data to increase the quality.

6.4 Preprocessing 2/ Dataset construction

For the second preprocessing part, after the information extraction, and the preparation of the dataset, the library `pandas` is necessary, as well as the functions from the first preprocessing and the information extraction. The starting point for this part is the big `.json` file, which includes all information from all editions for each language as an instance.

In a first step, this `.json` file was further enriched by the official names that were earlier derived from the ISO code and are in the "`language_canada.csv`" file. Then features were created for the names, the alternate names, and autonyms which combine in a set all values for the feature from the different editions. Furthermore, a binary feature was added that indicates the macrolanguages. The split information from the 1996 edition is now unwrapped and integrated into the normal structure. A differentiation between the information already scraped in a structured form or formatted by regular expression and the information that was split with the LLM is no longer necessary, since the evaluation process is finished.

The edition-based preprocessing is done and so the preparation for the analysis can be started. First, an overview of the languages was created. It is a `.csv` file, that contains the binary value of if a language is present in an edition. Another function creates a table with an overview of the status of each language per edition. For this, a JSON object with a subset of the status and the overview of the languages is necessary, and with the help of another function, the status is clustered into 'living', 'endangered' and 'extinct', 'unestablished' and 'unknown'. Furthermore, a table was created that contains the EGIDS level for the language in Canada per edition since 2013.

For the purpose of measuring linguistic diversity, the speaker number is of interest. So, a subset with the merged *Ethnologue* data, that only contains the ISO code and the extracted data because this is the information on the population, was created. This information has now been unwrapped to get the data per year and not per edition. The goal is to create a dataset with speaker numbers for the languages per year. Since the data is not necessarily recent data from the year of the edition and can be multiple times in different editions, the unwrapping was first only done for unambiguous data. The data with contradictions were found with a function and had to be manually checked and corrected. For this, the data were also checked with the *Ethnologue* data directly from the scraped version to compare whether the information extraction caused the error. Then if only one number was indicated with a phrase with more information, such as "X or fewer" or "X or more", just the number was kept. For the case where inside a value a choice is given "X or Y", the smaller number (first number) was kept. If a range is given, the median was taken. If the contradiction cannot be solved easily through the consultation of the *Ethnologue* and the described rules, it was looked whether the information could be looked up in the cited literature and taken from the source directly, this was the case for more recent information (e.g. through the indication FFPC 2014 referring to First Peoples' Cutural Council (2014)). For older data which cannot be looked up, the number of the newest edition was the choice. In general, some information was already not clear in *Ethnologue* and other figures were contradictory just between the editions.

Further information on user numbers was not only in the extracted information, but for the immigrant languages between 2000 and 2019 the user numbers were written in the summary and not in the text. This data was already cleanly extracted, but had to be added now to the subset. The function `reorder_columns` helps to better understand the data by sorting them. For the statistical analysis, the data types of the values have to be correct and were therefore checked. Numeric fields were set as integers and all values with wrong data types were corrected manually. Often information had to be shifted into the "extra_information" feature, if a string value was given like "few monolinguals" since few is not quantifiable. The corrected `.json` file now shows all information from *Ethnologue* on the speaker. To better see the gaps in the data, the dataset was also transformed into a `.csv` file. For the later analysis only the L1 speaker and the user in general are important, so a subset with the corresponding keys was created.

6.5 Census Enrichment

Because of all the gaps, it was decided to enrich the *Ethnologue* data with the census data from the years 1996, 2001, 2006, 2011, 2016 and 2021 (Statistics Canada, 1998a, 1998b, 1998c, 2003, 2007, 2012b, 2017a, 2023a) covering the time span of the *Ethnologue* editions. The data were manually searched and downloaded from the online presence of Statistics Canada. For the `.csv` files for 2021 and 2016 the data consists of information about the "Mother tongue" and "Knowledge of language", which are mapped to the categories L1 and user, respectively. 2011, 2006 and 2001 only include detailed information on mother tongue hence L1. For 1996 the information for L1 and user can be found in different files. One for L1 and two for user (separating official and unofficial languages).

For this step only known libraries were used: `pandas`, `re`, `numpy` and `json` together with the functions from `information_extraction` and `preprocessing`. Just before the actual enrichment task the downloaded data were brought in a rather unified form that the working with it is facilitated. Some rows and columns were deleted to get the real column titles; the deletion of unnecessary columns helps to provide a visual overview of the data but is not necessary for the technical process. Once the first manual preparation is done, some more preparations were necessary to just keep important information, these were done inside the function `create_census_languages_df`. Here, the ISO codes were also added for each language as identifier. The automatic mapping of ISO code to language name is not without problems, and hence the ISO codes needed a manual correction.

For the years 2011 to 2021 with the help of the column "Topic" the specific information can

be filtered for all on "Mother tongue" and for 2016 and 2021 also for "Knowledge of language", the characteristics column indicates the language and the column "Total" contains the counts, so the number of speakers. 2001 and 2006 already have the filtered information for L1. In the process the tables were prepared by deleting unnecessary rows and just including non-ambivalent data. Therefore, the rows containing the word languages were discarded since they indicated language families/ macrolanguages rather than individual languages. Languages specified with "n.i.e." (not included elsewhere) and "n.o.s." (not otherwise specified) were also removed. For older census (2011 and earlier) data "n.o.s." was included, because of the imbalance and lack of Indigenous languages as individual languages in the dataset. Although they were filtered out for the analysis because of their scope as macrolanguage, it is purely for complementing the dataset. For the preparation, the function `create_census_languages_df` processes the census data in line with the respective year and returns a prepared pandas data frame. The census data from 1996 were available in multiple files, as described above. The preparation therefore was done for each file separately, but in the same function (`create_census_1996_languages_df`) and analogously to the other years: it returns three data frames. The biggest difference is made for the official languages English and French for the user category, since they are separate from the non-official languages and more detailed listed. Therefore, the count for the user had to be done first by adding the count for knowledge in "English and French" to the counts of "English only" and "French only". This method is in line with the numbers from the newer censuses of 2021 and 2016.

In the preparation step, the ISO codes were looked up automatically. This had to be checked manually, since the ISO codes as the identifiers are crucial. Especially for 2021, this was important because the census uses autonyms for the language identification, but they are not necessarily automatically connectable with the ISO codes using the `langcodes` library (function defined in preprocessing, 6.2). Two common problems were that "Flemish" was identified as "ndl" which is "dutch" and not "vls". For Persian "fas" was automatically detected, which is the macrolanguage, but the data refers to the individual language "Iranian Persian" therefore "pes", following the entries in of *Ethnologue*. The census 2011 has an entry for "Creoles", which is neither a macrolanguage nor an individual language. But it exists an ISO 639-5 code. This code set includes Alpha-3 codes for language families and groups. For Creoles the code "crp" was set, which indicates "creoles and pidgins". Still, it is only included for completeness and is not relevant for the analysis.

For an overview of which languages are in the census and which in *Ethnologue* the function `find_shared_data` returns the shared and the individual ISO code as well as more detailed information on the *Ethnologue* information that also has information only in the census.

Continuing with the actual enrichment. Depending on which data are in the prepared census data frame, there are three functions. One that can enrich L1 and user, and one each for just L1 or user. In general, if there is already data in *Ethnologue* for which census data exist, the *Ethnologue* data is updated by the census data. If there is no data yet, then it is simply added (with the name of the language). Following the thought that the census is more precise, since *Ethnologue* obtains data from the census, and the direct source is the census. It can be seen that the number for user in *Ethnologue* (when the source is the census) corresponds to the number given in the census, but it is often not the exact number, but rounded. In contrast, the number for L1, even if the source refers to the census, does not agree with the census data for mother tongue. Also, not for any other language category in the census. *Ethnologue* always indicates a bigger number than the census. In this project, the focus is on individual languages; therefore, most of the more precise data from the latest census are not included in the enrichment process because it is about macrolanguages and dialects. Omitting this prevents the data from being doubled in the dataset. Especially regarding L1, where it is the 'ideal' situation to just have one L1 per person.

A Helper function was built for the saving of the versions after each enrichment for a census year. Two other helper functions enable the whole data frame to be seen in the output and to reset the option to the shortened view.

Once all data was merged, a second control process was necessary. For this, subsets were created for languages that were also in *Ethnologue* before the enrichment and the languages from the census where no existing ISO code was found in the *Ethnologue* data. It was checked if a language was in the *Ethnologue* data with another ISO code, so the ISO codes were adapted to *Ethnologue*, for example, for yiddish the ISO code was changed from the macrolanguage ISO code yid to the individual ISO code "ydd" for Eastern Yiddish, since this is the code used for yiddish by *Ethnologue* in Canada. Instances with no ISO code were deleted, this concerns dialects. In this process also the ISO code for Odia was corrected. *Ethnologue* refers to Odia by "ori", which is the macrolanguage Oriya, the individual language has the code "ory". This confusion may have arisen after the ISO code changes in 2012 and 2015 and kept with intention for identifier consistency. However, for this project, it was changed according to the latest version of the codes. The dataset was updated with the corrections and saved in JSON format as well as `.csv` file.

In a last step, a `.csv` file was created that summarizes the code characteristics for the dataset. It therefore includes the ISO codes present in the dataset, the scope and the language type. The scope of the languages can be "I" (individual), "M" (macrolanguage) or "C" (collective). Collective is just relevant for "crp" (creoles and pidgins). The language type is either "L" (living) or "E"

(extinct). The missing data for "[tce,ttm]", "aig?" and "crp" was added manually.

6.6 Imputation

For the imputation task the libraries `pandas` and `numpy` are necessary, furthermore the function from `dataset_creation`, `preprocessing` and `information_extraction` are useful. As already described, the important speaker numbers are "L1" and "user". Therefore, four subsets were created, for "L1_canada", "L1_all", "user_canada" and "user_all".

For the creation of the two datasets "L1" and "user" with data for the languages per year in "canada", the keys had to be cleaned, do that just the year is kept. For data where no source was referenced by *Ethnologue* the indicated year was the year of the edition. First, it was checked if for the year of the edition other precise data is available. If not, the information from the edition was taken for that year.

The goal is to have two datasets for "L1" and "user" with the data about Canada with languages per year. Since the data in “all” are not necessarily not Canada, but rather not specified and potentially Canada. First, the "canada" dataset was enriched by the "all" dataset for "L1" and "user", respectively. If there are 0 in "L1_all", it can be assumed that also there are zero L1 speaker in Canada. So, the zeros were taken over for the specific years. For the merging of the "canada" dataset and the general dataset, some rules were created. Most importantly, if there is already information for a specific year in the Canadian dataset, nothing will change there. If there is data given in the general dataset for a year 2021 or later, it is checked, if data for 2021 exists in the Canadian dataset, if so, nothing changes. If not, the data will be taken over for a manual check to enrich the Canadian dataset. Also, if there is data for a year, that is not in the Canadian dataset, it will be put in the set for manual check. With this check dataset it was decided individually which data will be included in the "canada" dataset.

For "L1" the decision was made based on a few rules: if the language is supposed to be a Canadian language, the data is taken over. To not integrate data was decided because of two reasons, first, it is not a Canadian language and in the details of the language entry in *Ethnologue* it is written that the language has only a few speakers in Canada, the number is not representing the speaker in Canada and therefore not taken over. In another case the number does not fit into the data series for the language already present in the "canada" dataset.

Equivalent the dataset for "L1", the dataset for "user" was created, the beforehand rule found rule that Canadian languages are taken over was integrated in the automated process. The manual decision for the rest of the data was made based on the data series. But the data was not only compared to the user data but also to the "L1" data, since during the information ex-

traction task the data which is not clearly marked was collected as "user" information. But if there is no data series to compare with, again user was selected. If the data does not fit neither in "L1" nor in "user", the data was discarded and not integrated in one of the datasets. Hence, the classification is context sensible, but researcher based. Like this there are actually three datasets, one that contains the information for user, and two for "L1", one of them containing the secured information on "L1" which is explicitly written like this in *Ethnologue* and the census. The other "L1" dataset also integrates the information that is probable "L1". A list of which data was integrated or discarded in the checking process can be found in the appendix 10.7. Interesting observations could be made: often the numbers from 2009 fall out of the line, sometimes the user number decreases while the number for L1 increases.

All now remaining gaps are filled in the actual imputation task for which the datasets were prepared. For the imputation two methods were chosen: Linear imputation and moving average. For linear imputation the data points are connected through a linear function (Blu, Thevenaz, & Unser, 2004; Genolini, Ecochard, & Jacqmin-Gadda, 2013; Kosten, 1967; Meijering, 2002; Wittmann, 2019). The moving average (Hyndman, 2010) calculates the missing values by using the average of the n data points around a data point. This means the value depends more on the surrounding and it is period sensible but smooths outlier. To be able to calculate the average the missing data was filled beforehand with last observation carried forward and next observation carried backward. However big gaps of missing data can lead to difficulties, as it reacts quite late on changes. Especially in the case of this study where a window size of 3 was chosen which means that only the values in direct neighborhood are relevant. A window of three represents in this study 10% of the time frame. Also with a bigger window size most of the data would be still be no real observations, since there are bigger gaps in any case. Linear interpolation is based only on the values actually in the dataset and is a curve fitting method. With this method approximate values are to be found and again the smaller the gap the more likely it's more approximate. In both cases the methods are quite simple and fast. This gives a good basis for the analysis. Further imputation methods were not used or compared.

First, all missing year columns were added to the data frames to have all empty cells created, that have to be filled. Then using the `pandas` library, the missing data was interpolated with linear regression. The moving average imputation was also done using the `pandas` library. In any case, it was only imputation done strictly between the data, and no extrapolation. Therefore, there is still no data for every year for every language. The first and the last data is always data from *Ethnologue* or census and therefore real observed data.

The imputation finalizes the preprocessing steps and the dataset creations. In the end there are six datasets on which the analysis can be performed.

7 Analysis

The analysis was also performed in python. The functions were built following the methods described in 5. For the implementations the following libraries were used: `pandas`, `json`, `re`, `numpy`, as well as `scipy` (for fundamental algorithms and mathematical tools, Virtanen et al. (2020)), `matplotlib`, `plotly`, `seaborn` and `squarify` (for plotting and data visualizations, Hunter (2007); Laserson (2024); Plotly Technologies Inc. (2015); Waskom (2021)) and `Jinja2` (for data export in LaTeX format, Pallets (2025)).

In the following, the analysis is done based on the extended L1 ("L1_plus") dataset that used the linear imputation method. If notable differences are seen in the other datasets, this will be discussed occasionally in the sections or referenced to the appendix. First, the L1 datasets are preferred over the user dataset, since most of the methods are designed for a unilingual context. In addition, the dataset does not show a consistency in the speaker number, which may be caused by the wide definition of the category and the amount of missing data. After the imputation "user" has more data points and covers more languages, but the extended L1 dataset had more data points before the imputation and is therefore based on more real values than the other datasets; see table 10.2. The linear imputed datasets show, over all, a smoother curve (see figures 10.7, 10.8, 10.9). In general, the choice of the imputation method does not change the interpretation of the results.

7.1 Languages in Canada - Descriptive Analysis

For Canada, there are 231 living languages reported by *Ethnologue* in the edition of 2025 (Eberhard et al., 2025a). Including the two official languages as well as Indigenous languages and immigrant languages. In the prepared dataset, there are 291 different languages listed, with no differentiation on the scope or the status. More precise information on the count can be found in the next section. Information on the distribution and language size can be found in section Linguistic Evenness. For all years, the smallest speaker number (6) in the L1 plus dataset is Heiltsuk (hei, Indigenous language) in 2014. However, in the user dataset there are more languages recorded with less than 20 speaker, down to one, like Munsee (umu, Indigenous language) in

2022 or Tlingit (tli, Immigrant language) in 2018. This might be because there is no further information on the user because of a lack of quality. Or it is a sign for language shift that there are few user, but they are not seen as L1 speaker. Figure 7.1 shows the languages by their speaker number and the mean of the different groups. It is visible that Indigenous languages have smaller speaker communities than immigrant languages. The biggest languages are always the official languages English and French.

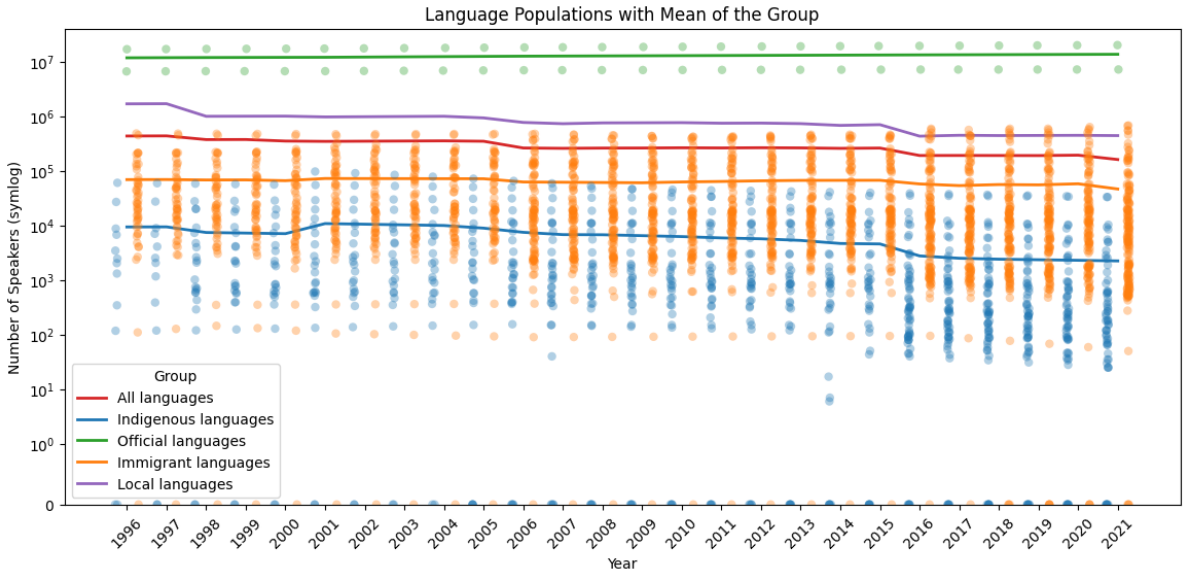

Figure 7.1: This strip plot shows the language populations. The color indicates the group of the language. Green are the official languages, blue the Indigenous languages and orange the immigrant languages. The line plot adds information on the mean size of the languages in the language groups.

7.2 Linguistic Richness

Unsurprisingly, the linguistic richness of the official languages is always two: English and French. Immigrant and Indigenous languages both experience growth throughout the period, as can be seen in figure 7.2. The dataset contains 59 individual living languages for 1996. In addition to the two official languages, 10 Indigenous languages and 47 immigrant languages are listed. In 2021, there is data for 205 languages, of which 54 are Indigenous languages and 149 are immigrant languages. For a complete overview, see table 10.3. The group of the immigrant languages has the highest linguistic richness and also shows the biggest increase between 1996 and 2021.

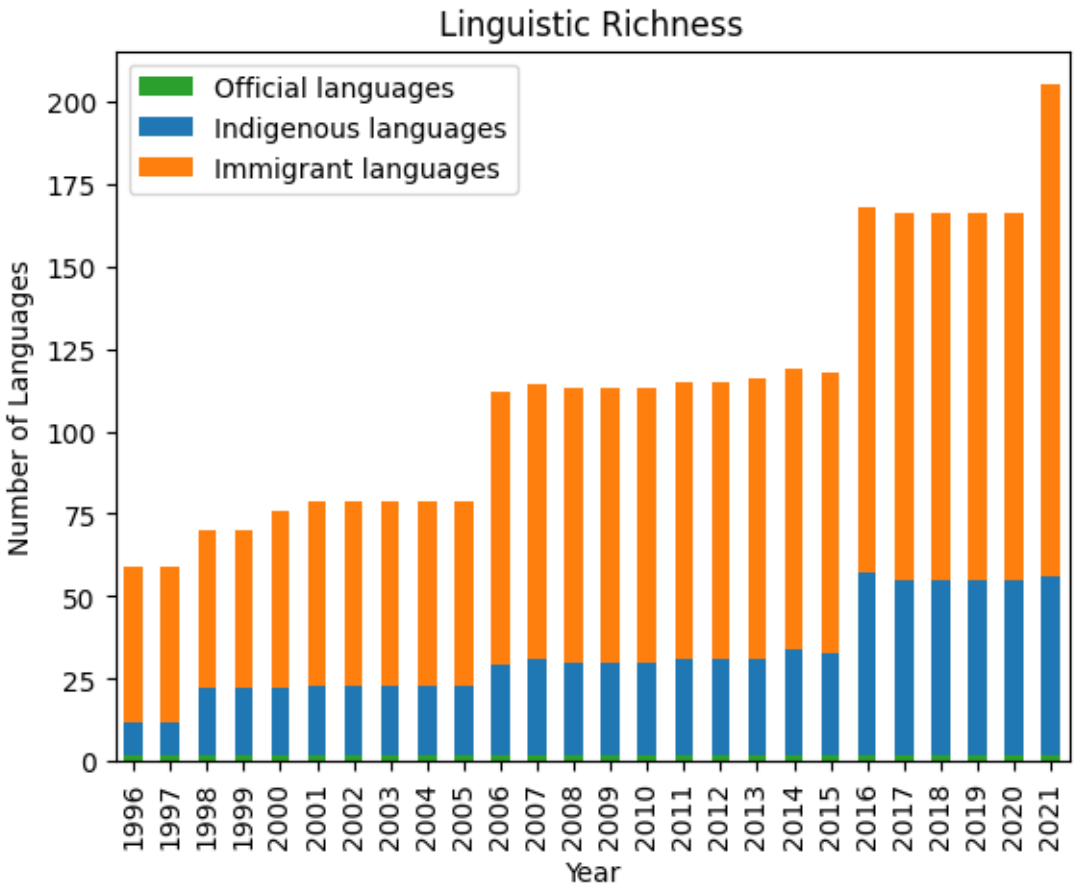

Figure 7.2: This stacked bar plot shows the linguistic richness of Canada, it is measured by the count of all living individual languages represented by group.

The increase can be explained less by the addition of 'new' languages and more by the improvement in data quality. Only individual languages are counted, whereas earlier data only included information on macro languages. Significant jumps can also be seen, reflecting the publication years of the editions and the years of the census, as only internal imputation was used. On the other hand, design decisions and internal *Ethnologue* processes can also be identified. With the complete switch to annual publication and the creation of the paywall, it can be assumed that the financing was intended to improve quality. At the same time, developments in the census, e.g. the policy of listing more languages in 2021, can also be seen. This also has an impact on *Ethnologue*, as *Ethnologue* obtains data from the census.

The linguistic richness shows a diversity for Canada that increases over time which means that more languages are present in 2021 than in 1996. The increasing diversity is primarily linked to the growth in the immigrant language group. Linguistic diversity measured by richness shows the effectiveness of the multiculturalism policy of Canada.

7.3 Linguistic Evenness

Over time, not only the linguistic richness increases, but also the abundance of the language groups, meaning the number of speakers within the groups; see line plot 7.3. This is in accordance with population growth, fig. 10.10. All language groups are getting more speakers, only the indigenous languages do not show a significant increase. The plot 10.8 of the comparisons of the imputation methods for Indigenous languages show the speaker numbers in more detail. It reveals the number of speakers increased until 2001 and decreased then again. In 2021 there are slightly more speakers of Indigenous languages than in 1996. Both other groups show an increase in the number of speakers. This increase is roughly the same size, which is why the increase in speakers of all languages is based approximately equally on the increase in these two groups. For the local languages, the number of speakers of the official languages is most important, and Indigenous languages are practically insignificant.

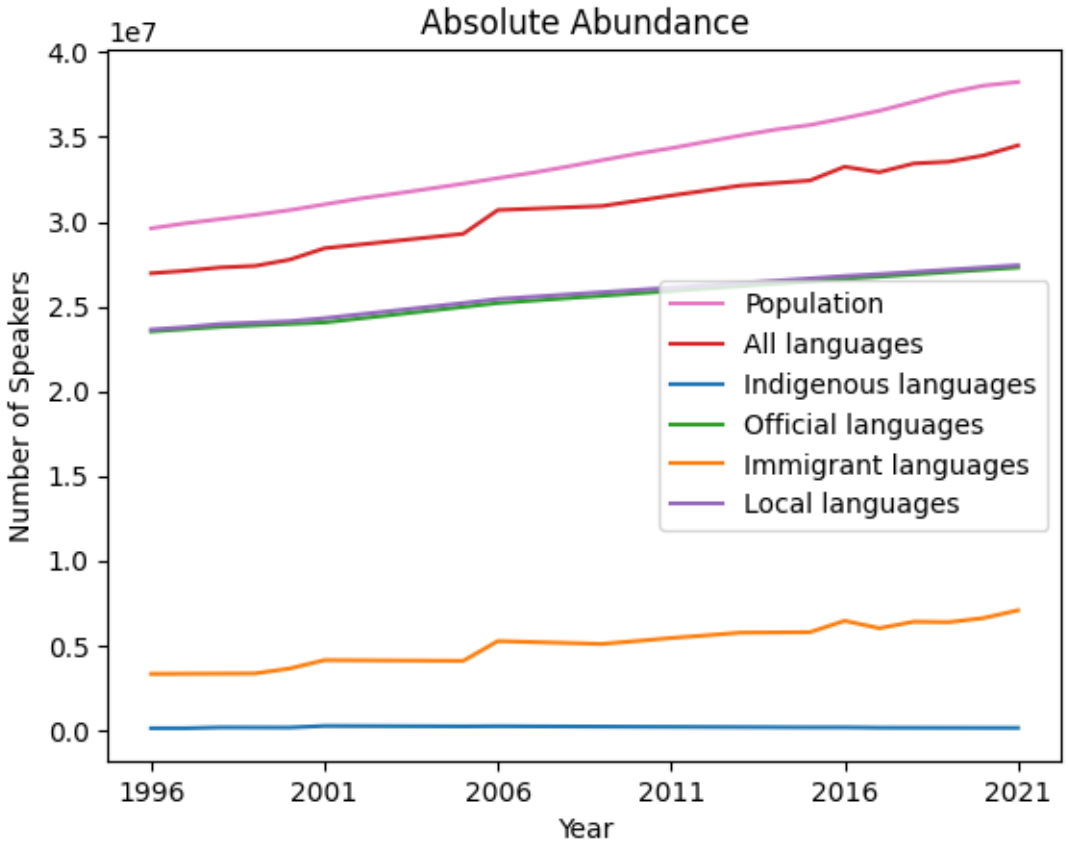

Figure 7.3: This plot shows the absolute abundance of the languages, measured by the number of speaker. In addition it shows the population estimates of Canada (pink). The red line (all languages) indicates the sum of all speaker represented in the dataset.

The plot 7.4 shows that the data available cover around 90 % of the population; see also table 10.4. The relative proportion of people who have an official language as their first language falls from around 80 % to approximately 72 %. In 1996, around 11 % had an immigrant language as their first language, while in 2021 more than 18 % spoke an immigrant language. The number of speakers of Indigenous languages is almost invisible. The growth from 0.38 % in 1996 up to 2001 can be attributed to the data situation. Indigenous languages made up 0.8 % in 2001, and this already very small proportion will fall to 0.35 % in 2021. So, even though the linguistic richness of Indigenous languages increases and also the number of speakers is slightly higher in 2021 than in 1996, their relative abundance decreases.

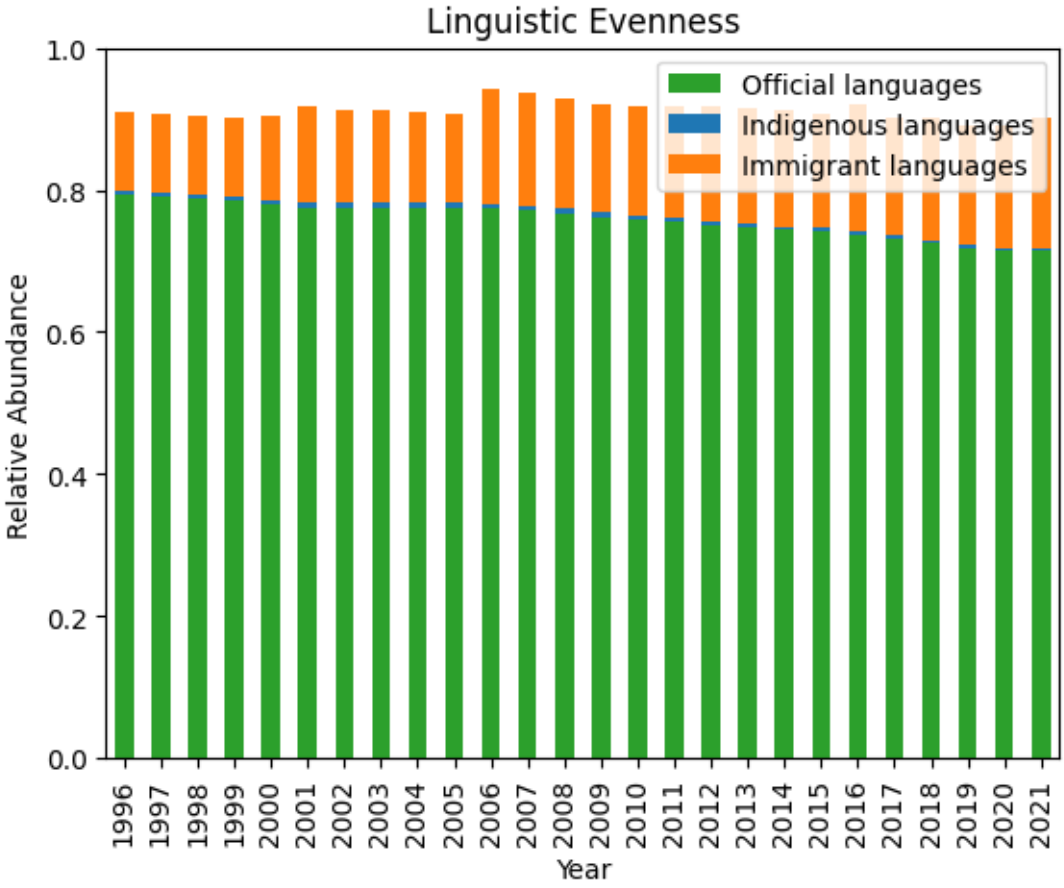

Figure 7.4: This plot shows the linguistic evenness, normalized with the population estimates. The blank space between the bar and the upper border of the plot represents the missing data in the dataset.

The tree maps 7.5, 7.6 and 7.7 show the distribution of speakers across languages. As expected, the official languages take up the largest part. English (eng) is represented by the largest area and French (fra) is the second most represented language, but clearly behind. This is the

case throughout the entire period under review. In 1996, the largest immigrant languages were Italian (ita) and German (deu). The largest Indigenous language, with 60,000 L1 speakers, is Plautdietsch (pdt). It should be noted here that the Indigenous languages group was formed on the basis of *Ethnologue*. Indigenous languages in this study are languages that *Ethnologue* considers Canadian, or that *Ethnologue* assigns to Canada as their primary country. Thus, Indigenous languages that are native to both Canada and the US, and are found, for example, in the border region, may fall into the group of immigrant languages if the primary country is not Canada. The largest Indigenous language in 2001 is Woods Cree (cwd) and the third largest immigrant language is Yue Chinese (yue). Since only the macro languages Chinese and Cree are present in the 1996 dataset, it is clear that the level of detail increased between 1996 and 2001, and increased even further by 2021, which was already indicated by the linguistic richness. Although in 2001 there was one Indigenous language among the 20 largest languages in Canada, in 2021 only immigrant languages are listed alongside the official languages. These are no longer mainly European languages, but Mandarin (cmn) and Punjabi (pan) are the largest immigrant languages, followed by Yue, Spanish (spa), Tagalog (tgl), and Italian. In comparison to the distribution of speaker across languages based on L1, the treemap 10.12 based on user (not only L1 speaker) shows that the evenness is quite similar, with English and French being dominant. But more Indigenous languages become visible, which might indicate language shift in action, that the Indigenous language is not the first language, but still in the language profile of the people.

The linguistic evenness shows the high dominance of the official languages in the Canadian landscape. But at the same time, the linguistic evenness shows, that Canada is getting linguistically more diverse due to the immigrant languages. Compared to linguistic richness, the Indigenous language endangerment is visible through linguistic evenness by not having an impact on the measurement.

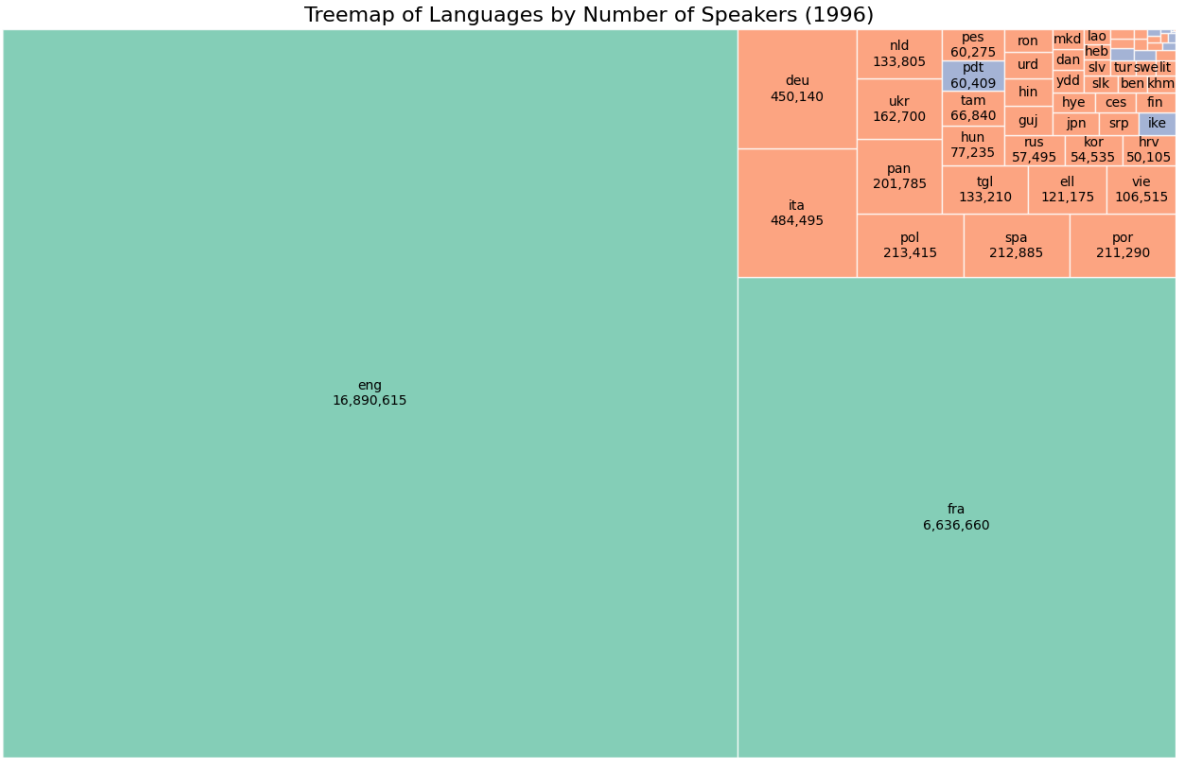

Figure 7.5: This tree map shows the size of languages based on the number of L1 speakers in 1996. Around 9 % of the Canadian population is not represented in this plot. Green are the official languages, orange are the immigrant languages and blue are the Indigenous languages. The ISO codes are indicated for all languages with more than 9000 speakers. The top 20 languages are accompanied by information on their number of speakers.

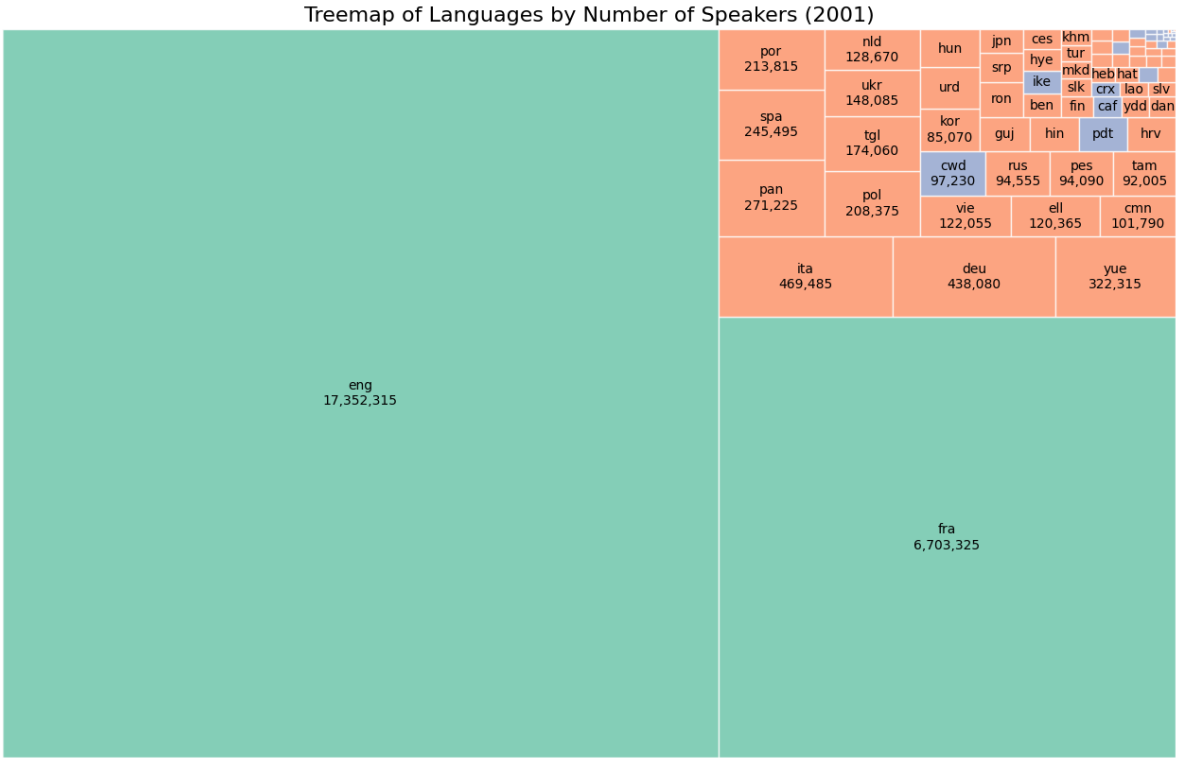

Figure 7.6: This tree map shows the size of languages based on the number of L1 speakers in 2001. Around 8 % of the Canadian population is not represented in this plot. Green are the official languages, orange are the immigrant languages and blue are the Indigenous languages. The ISO codes are indicated for all languages with more than 10000 speakers. The top 20 languages are accompanied by information on their number of speakers.

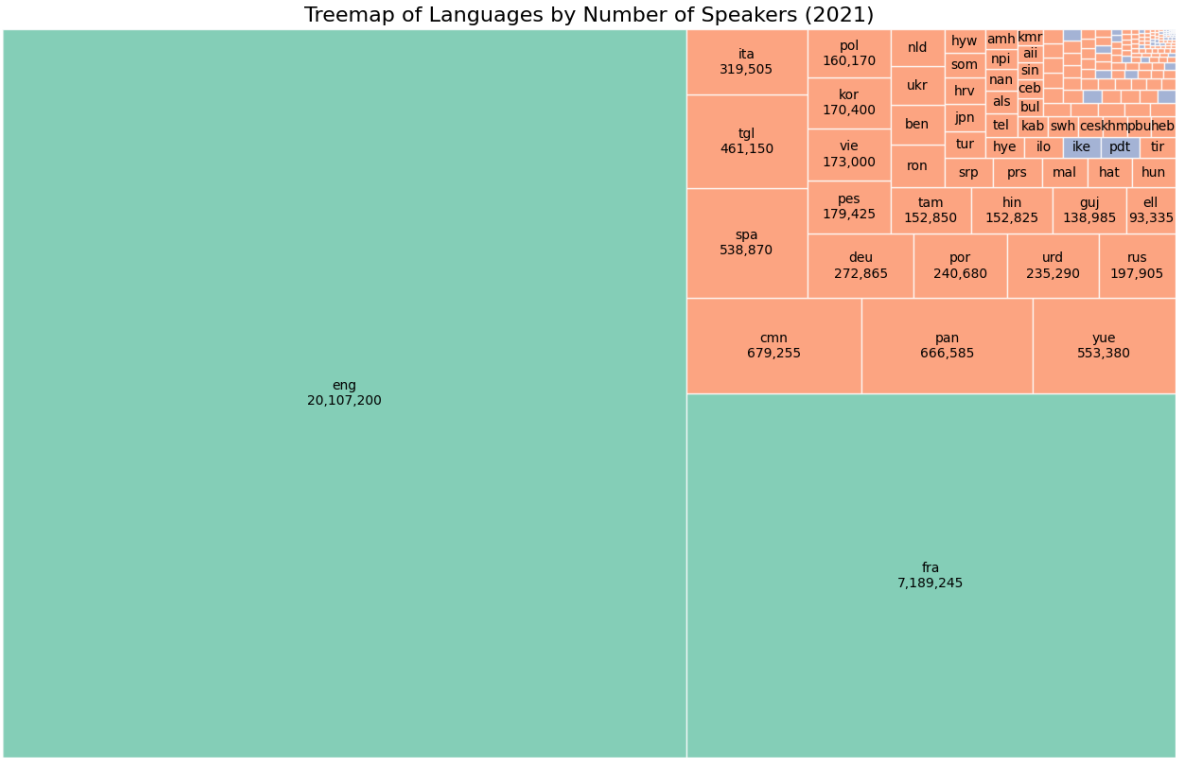

Figure 7.7: This tree map shows the size of languages based on the number of L1 speakers in 2021. Around 10 % of the Canadian population is not represented in this plot. Green are the official languages, orange are the immigrant languages and blue are the Indigenous languages. The ISO codes are indicated for all languages with more than 16000 speakers. The top 20 languages are accompanied by information on their number of speakers.

7.4 Index of Linguistic Diversity by Harmon and Loh

Figure 7.8 shows the Index of Linguistic Diversity for Canada between 2001 and 2021 which measures the changes of the relative distribution. The baseline was set to 2001 for data quality reasons. For comparison the figure 10.13 (ILD with baseline 1996) can be consulted.

For the diversity based on the population (with around 10 % missing data) or the diversity for the data present in the dataset, the development looks quite similar. Until 2006 the diversity increases, to be than quite stable until 2016, since then it is decreasing. In 2021, there is still a higher diversity compared to 1996 or 2001, respectively. This might be linked to the increase of the linguistic richness seen in the dataset. With 2006 to 2015, there is a richness which just slightly increases. In 2016 it increases by nearly 50 %, since there is no increase in the ILD visible,

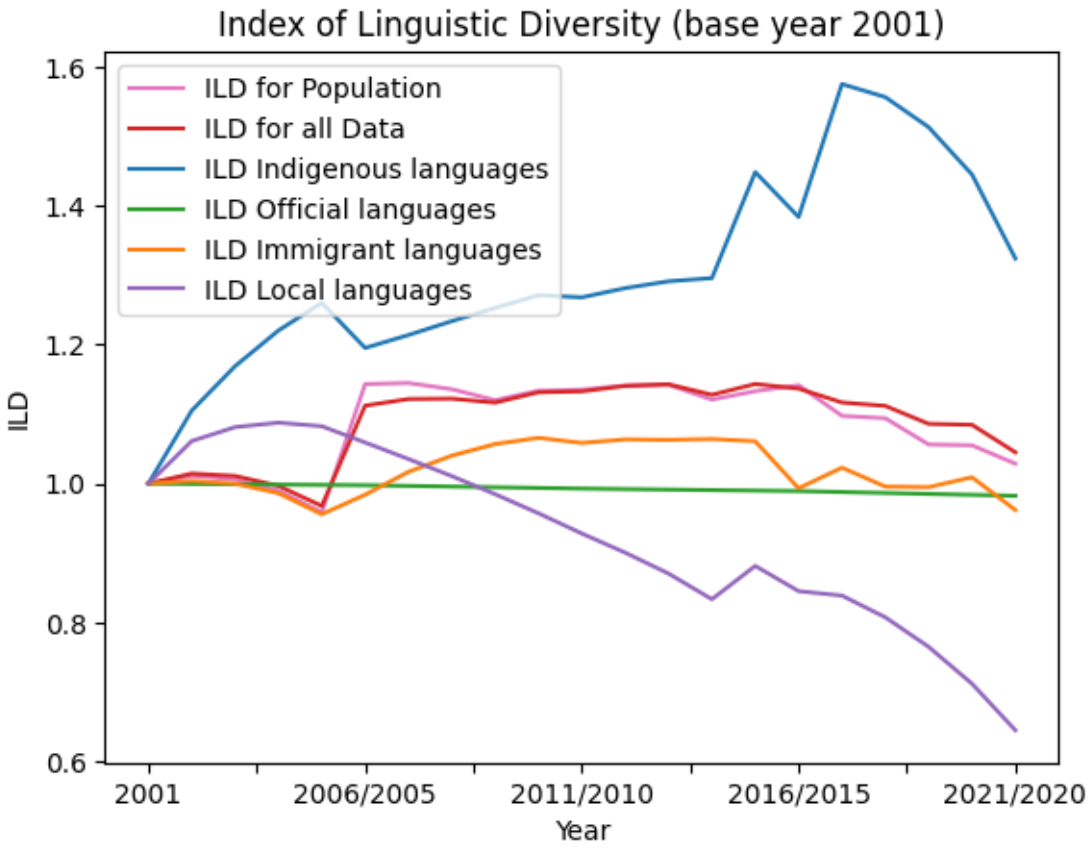

Figure 7.8: This plot shows the Index of Linguistic Diversity with the baseline set on 2001. The ILD that is computed on all data that is normalized with the population estimates is indicated by the pink line. The red line represents the ILD of all data that is normalized by the sum of this data. The other groups are also normalized by the sum of their group and it shows therefore the internal ILD of the groups.

the dataset must have become less evenly distributed. Regarding the official languages no big changes can be observed compared to 2001. Just a slight decrease of diversity; this is linked to English, which became a bit more dominant in Canada.

In general immigrant languages show fluctuations but no big change.They became a bit more diverse first, but since 2015 the diversity decreased. The small increase can be linked to the increase of the linguistic richness. But with time this is not the most driving power, but the distribution has more impact, so even if there are more languages, the fact that they are less evenly distributed is more important. With this change, the linguistic diversity decreases because of the dominance of few languages. This is even more visible at the ILD based on 1996. The immigrant languages are less diverse in 2021 than in 2001 and 1996.

Indigenous languages seem to get more diverse, which is because of the increase of the rich-

ness, from 2017 on the diversity decreases, which means that the discrepancy between the languages increases. The increase of richness, does not prevail the decrease of evenness.

A different picture emerges when looking at the local languages (official and Indigenous languages) combined. A significant decrease from 2001 to 2021 to under 0,65 is visible. This suggests that even if the richness increases, the dominance of the official languages, especially English outweighs the Indigenous languages. It can be seen that dominant Indigenous languages become smaller, which leads on the one hand to a higher diversity inside the Indigenous languages because they become more evenly distributed, on the other hand, it increases the gap to the official languages.

It is therefore important to not only look on the diversity in general, but on the groups as well and the combination of them. Compared to the other imputation model, it can be seen that developments are delayed when using moving average imputed data; see 10.14.

For Canada in general, linguistic diversity has not changed much according to the ILD. However, the ILD clearly shows the dominance of the official languages on the cost of Indigenous languages, while the Indigenous languages alone indicate a more balanced relative distribution of speakers with positive changes since 2001.

7.5 Linguistic Diversity Index by Greenberg

The Linguistic Diversity Index defines linguistic diversity as the likelihood that two people who meet by chance will speak not the same language. Figure 7.9 indicates that overall the linguistic diversity in Canada increased between 1996 and 2021. Measuring diversity in relation to the population, where 10 % are not further defined, it is logical that diversity is higher than when based solely on the data.

The two official languages are not very diverse, and their diversity is declining furthermore slightly, which is related to the fact that English is becoming slightly more dominant. Indigenous languages became more diverse until 2011, then stabilized at a certain level until around 2016, since then the diversity declined slightly again. The Indigenous languages show the biggest gain of diversity with LDI values of 0.64 in 1996 to 0.86 in 2021.

The local languages taken together show slightly greater diversity than the official languages. However, the official languages have so much influence that diversity decreases slightly, as is also the case with the official languages. This shows the high degree of imbalance between the official and Indigenous languages. With this in mind, the overall increase in diversity must be attributed to the third language group, immigrant languages. These indeed show the highest diversity, with the LDI rising from 0.93 to 0.95 between 1996 and 2006 and has since stabilized

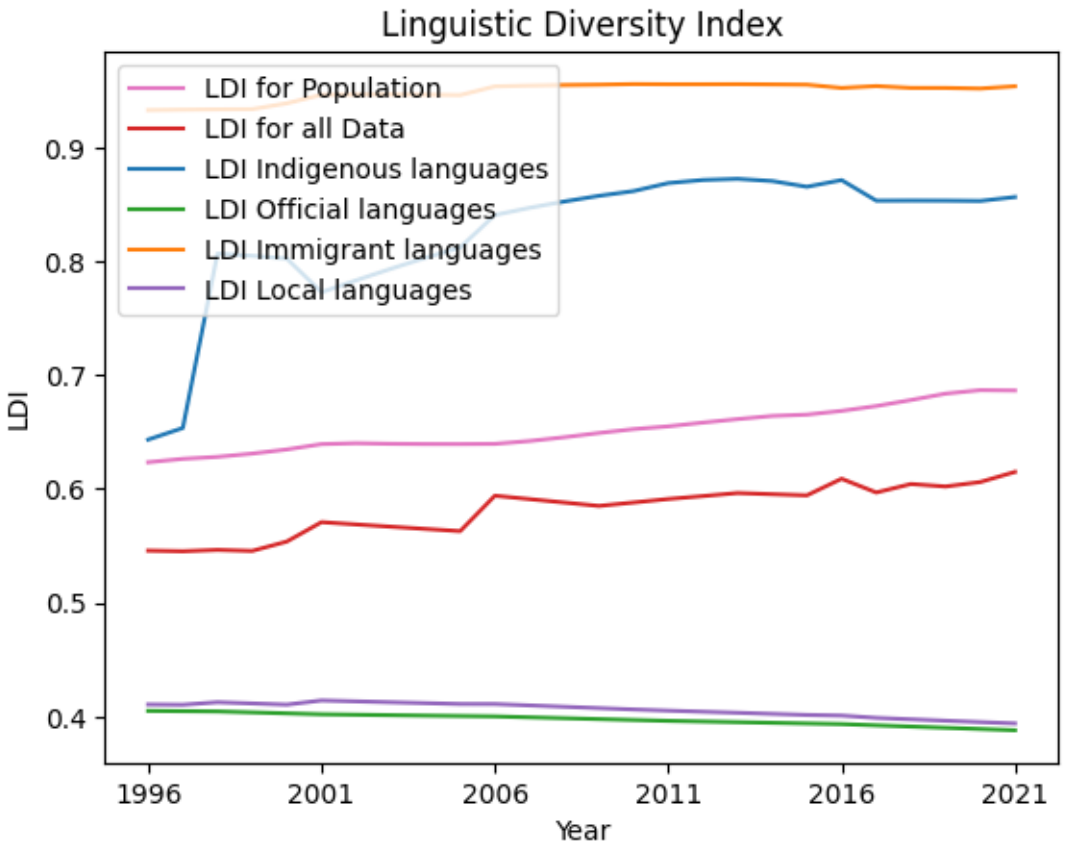

Figure 7.9: This plot shows the Linguistic Diversity Index for Canada between 1996 and 2021. The LDI that is computed on all data that is normalized with the population estimates is indicated by the pink line. The red line represents the LDI of all data that is normalized by the sum of this data. The other groups are also normalized by the sum of their group and it shows therefore the internal LDI of the groups.

at this level. The increase at the beginning of the period under review can be attributed, just as for the Indigenous languages, to the quality of the dataset and the significant increase in linguistic richness. Although other languages were added later, they do not appear to have had a significant impact on diversity due to their small size. This is consistent with the insensitivity of the diversity measure to small languages.

The increase in linguistic diversity in Canada is therefore the result of growth due to immigrant languages but continues to be influenced by the dominance of English and French. The languages become more evenly distributed, and it is less likely that two people randomly picked in Canada have the same first language.

7.6 Shannon's Entropy

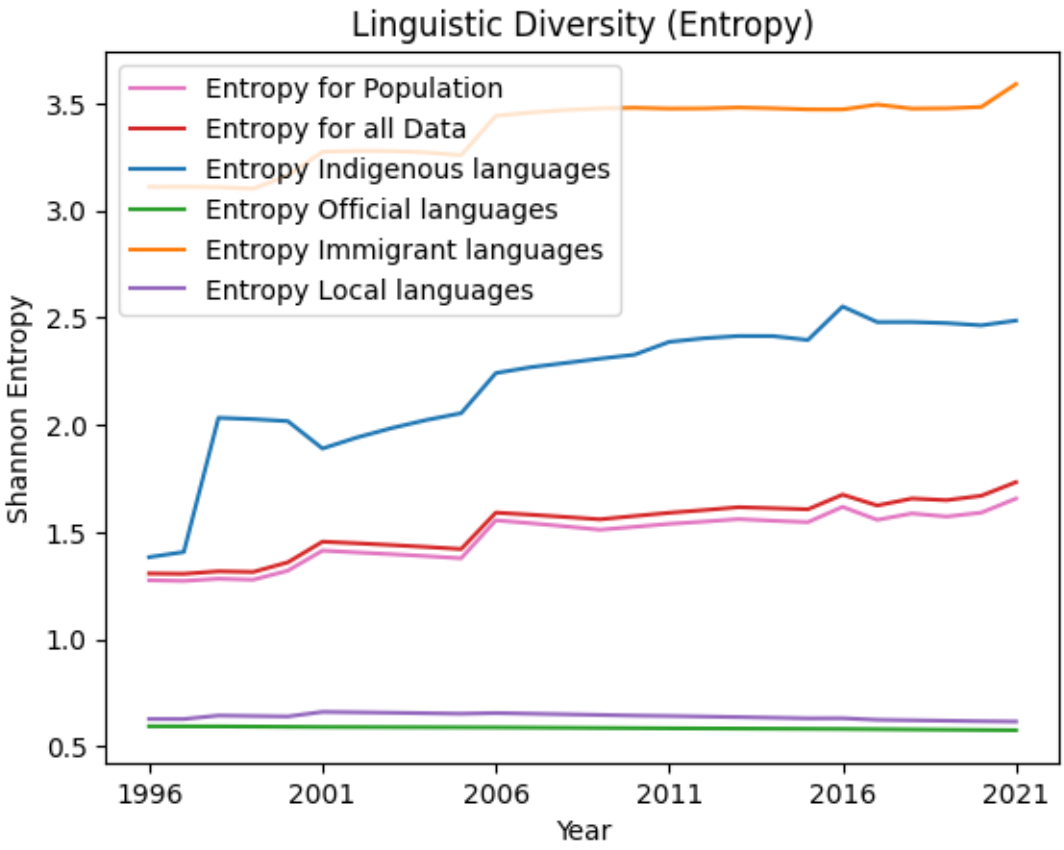

Figure 7.10: This plot shows the linguistic diversity for Canada between 1996 and 2021 computed on the basis of the entropy. The entropy that is computed on all data that is normalized with the population estimates is indicated by the pink line. The red line represents the entropy of all data that is normalized by the sum of this data. The other groups are also normalized by the sum of their group and it shows therefore the internal entropy of the groups.

Following the approach that linguistic diversity can be computed through Shannon's entropy by determining how uncertain the linguistic identity of a person is in a community. The plot 7.10 for the entropy shows similar patterns like the LDI. The entropy based on the population and based on all available data increases but is very similar and runs parallel, with the entropy for all data being slightly above that of the population. This is a small difference from the LDI. The position of the entropy normalized with the population estimates is below the curve showing the entropy normalized to all available data. This can be explained by the missing 10% and the way entropy and LDI are calculated.

The lowest entropy is found within the official languages. This is declining slightly, which means that it is becoming less surprising which language a person speaks. This is related to the fact that English is becoming more dominant, making it more likely that a person will speak English. However, only a very slight decline in entropy is apparent.

The entropy of Indigenous languages shows an increase until 2016, then declines slightly and remains at that level. The peak in 2016 is also the time where the linguistic richness is the largest for the Indigenous languages. This shows the impact of small languages on the entropy based linguistic diversity. The importance of the linguistic evenness is visible through the seasonal peak in 1998, which is also reflected through the LDI.

For the local languages, entropy behaves similarly to that of the official languages; it is low in comparison, albeit slightly higher than for the official languages alone. A slight decline can be observed until 2021. It rose minimally until 2001, but has been falling since then. The dominance of the official languages is far too big that Indigenous languages do not play a role for the linguistic diversity calculated by Shannon's entropy or the LDI.

The immigrant languages show the highest entropy of the groups examined. Here, too, it can be seen that they became more diverse between 1996 and 2021. This happened in stages, with periods of little change: 1996 to 2000, 2001 to 2006, and 2007 to 2020 and bigger shifts between them.

The overall increase of entropy values suggests a more balanced distribution of speaker numbers across languages over time. Just like for the LDI, the immigrant languages are decisive for the overall diversity in Canada, making it less certain which first language a person has in Canada. Even though within the Indigenous language group the diversity increases, they have no significant impact on the overall diversity, which here again emphasizes the endangered status of the Indigenous languages by their small speaker communities.

7.7 Hill Numbers Framework

The Hill Numbers Framework allows different measures of diversity to be compared with each other, see figure 7.11. The numbers equivalent of order $q = 0$ corresponds to the linguistic richness. At $q = 2$, small languages are neglected and it is a transformation of Greenberg's Linguistic Diversity Index, also known as inverse Simpson index. With the Hill number of order $q = 1$, small languages are neither neglected nor emphasized. This balance is also found in Shannon's entropy.

The shape of the diversity curves is similar to their numerical counterparts and emphasizes higher diversities. Therefore, the increase in diversity is higher for immigrant languages, which

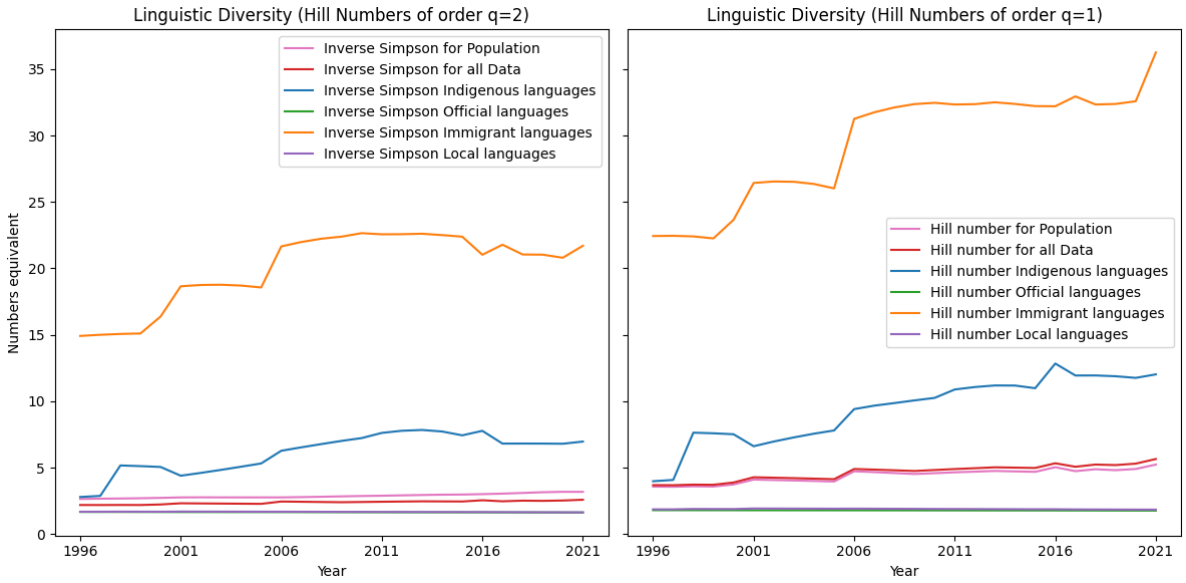

Figure 7.11: This plot shows the linguistic diversity in Canada between 1996 and 2021 as a comparison between the Hill numbers of order one and two. The hill numbers that are computed on all data that is normalized with the population estimates is indicated by the pink line. The red line represents the entropy of all data that is normalized by the sum of this data. The other groups are also normalized by the sum of their group and it shows therefore the internal entropy of the groups. The Hill numbers of order zero correspond to the linguistic richness.

have the highest diversities in entropy and LDI. The LDI did not show any major changes in diversity, but the numerical counterparts ($q = 2$) show that diversity initially increased and then stabilized. The number equivalents for immigrant languages are between 14.9 and 22.6, for Indigenous languages between 2.8 and 7.9, official languages stay around 1.7 with a minimal decrease to 1.6. The local languages can be represented as well by 1.7 (just slightly above the official languages) and all languages range between 2.2 and 2.6, or normalized by the population between 2.7 and 3.2.

For the Hill numbers of order one, the same trends as for the entropy are visible. With immigrant languages having the highest diversity (22.4 to 36.2) and Indigenous languages showing an increase from 1996 to 2021 (with values between 4.0 to 12.8). The official languages can constantly be represented by the numbers equivalent of 1.8. The local languages, slightly above the official languages again have values of 1.9. The total diversity of Canada increased from 3.7 to 5.7.

To compare these two orders with the linguistic richness, the numbers of order zero, only the diversity of the official languages is likewise small (two languages). But the linguistic richness is

for all groups higher than the hill numbers of order one or two. Following the linguistic richness, the diversity increases for all groups, except for the official languages.

Notable is that the numbers equivalents of order one are higher than of order two, suggesting that the Canadian linguistic landscape is indeed more imbalanced by having small languages. Especially in the group of immigrant languages it can be seen that the order needs higher numbers to represent the group. This is accordance with the linguistic richness. The numbers for Indigenous languages are also higher even though less distinct between the orders one and two. This suggests that they are more evenly distributed overall. To represent the total linguistic diversity, there are smaller numbers equivalents necessary. Again order one higher than inverse Simpson. This means that most of the languages are invisible in the linguistic landscape of Canada. Which is proven by the different orders of the hill number framework. It is therefore not easy to answer the the question of if the hypothesis is true and the linguistic diversity in Canada decreases. The diversity in fact increases if considered in total. But this is because of the increase of languages, which is linked to the increase of quality of the data, especially for immigrant languages. However, the linguistic diversity is not declining because of the Indigenous languages, because they were already quasi invisible in the linguistic landscape and do not really influence the diversity measurements (neither entropy nor LDI, as well as the the hill numbers based on those measurements).

7.8 Status

For the analysis of the status of the languages, missing data was imputed by the last observation carried forward (LOCF) method. This is because the status, or the EGIDS level respectivly, is a categorical value. It is assumed that an information is valid until it is changed. There are only data on this topic listed in *Ethnologue*, so any enriched data from the census has no correspondent information on the status or the EGIDS level. The information is always based on the edition and has no further information on currentness, therefore, the data reflects the publication years of the editions. No extrapolations were performed. The number of languages included in this part of the study may differ from the first part (from the linguistic richness), and also other languages may be included or excluded from this study depending of their, first, existence in the *Ethnologue* edition and second, on their availability of speaker numbers.

The analysis of the status of the languages and their changes precedes the analysis on the threat level of the languages in Canada. The edition of 1996 does not give an explicit information on the living languages. For this study, it is therefore assumed that languages listed and not marked particularly as endangered or extinct can be seen as living languages. From 2000

on however, also living languages are explicitly assigned their status, and languages can therefore have an unknown status if information is not given. This is mainly the case for immigrant languages; see also chapter 5.6. At the same time *Ethnologue* focuses on living languages, that is why extinct languages may not be included in this study, but more extinct languages exist. This is not much of a problem since it is more about the status changes than the exact number of extinct languages.

Until 2013 the majority of the classified languages are living, this changes with the introduction of the EGIDS in *Ethnologue*. Most of the languages are from there on endangered. Even though the linguistic richness increased, the number of living languages decreased and changed to endangered; see fig. 10.15. In the edition of 1996, 60 Indigenous languages are listed. With the following edition, the number of listed Indigenous languages increased to 75. And just in 2013 and 2024 the number increased by one, meaning being quite consistent. However, the overall threat on Indigenous languages is very visible in figure 10.16, and especially in the last 12 years. The steep rise in the number of immigrant languages is not covered with information on the status (fig. 10.17). There are some endangered languages, but just like for the Indigenous languages, in 2013 the number increased significantly. Since then, more living languages are classified, but also extinct languages appear on the list. Most of the extinct immigrant languages are languages that are native to the United States according to the primary country mapping of *Ethnologue*. So they are still Indigenous languages to North America.

The EGIDS shows the threat level for the languages since 2013. Also here (fig. 10.18), it is possible to see that the languages get more threatened over time. The language shift is visible especially for Indigenous languages; see fig. 10.19, the number of developing languages (level 5) decreases and the number of threatened languages (level 6b) increases. The same, shifting languages (level 7) are decreasing, but the level 8a, moribund, encompasses more languages over time. Some fluctuations can be seen for nearly extinct languages (level 8b), but at the same time more languages become dormant and extinct (levels 9 and 10). Most immigrant languages do not have an EGIDS Score for Canada (fig. 10.20. So, the number of developing languages increases, but so does the number of nearly extinct and dormant languages.

To have a better understanding the changes of the threat levels over time are of interest, see fig. 7.12. Not including the changes into the dataset, a total of 85 changes can be detected (see table 10.5). Of the 85 changes, 78 were to a higher threat score, and the most changes can be seen between the categories 'living' and 'endangered'. The two official languages are always classified as living. Most of the immigrant languages have the weight one, since their status is 'unknown'. For Indigenous languages the threat level increases. However, this also shows the inconsistencies of the dataset. As Wendat (wdt, Indigenous language) is listed as 'unknown'

from 2013 to 2023 even though before that and in 2024 the language is listed as 'extinct' and the status of the Maritime Sign Language (nsr, Indigenous language) changes in 2000 from 'endangered' to 'living' and in 2005 back to 'endangered'. These are the only changes to a status with a higher weight in the group of Indigenous languages. The five changes that decrease the threat levels in the group of the immigrant languages are from 'unknown' to 'living'.

The EGIDS scores for the languages since 2013 are more precise and therefore interesting for the understanding of process within the status described above (see fig. 7.13). In 2013, the EGIDS level 7 (shifting) counts the most Indigenous languages, in 2024 however, more languages can be found with the score 8a (moribund). There are a few more improvements in the score, but still more languages become more threatened, see table 10.6. Most of the changes are between neighboring levels. However, Eastern Canadian Inuktitut (ike, Indigenous language) and Inuinnaqtun (ikt, Indigenous language) changed their status from provincial (level 2) to threatened (level 6b). Dakota (dak, immigrant language), is the only language with two changes from 6b to 8a in 2016 and to 8b in 2019. In general, most changes can be seen between 2018 and 2019, with the biggest flow from level 7 to level 8a. The only inconsistency shows Kaska (kkz, Indigenous language) that is assigned with level 7 and just in 2018 with level 8a. Otherwise, all languages just have one change maximum ignoring changes into/from the dataset and to/ from 'unknown'.

In general, the trend of Indigenous becoming more endangered is visible through the status. On immigrant languages, it is difficult to find an overall valid trend, since most do not have a status assigned. Also the degree of endangerment increases, which is visible through the EGIDS. Threatened languages (mostly Indigenous languages) become even more endangered, also in the last 10 years.

Figure 7.12: This figure shows an alluvial diagram that visualizes the status changes over time. The dimensions are the years and the values of the dimensions are the status weights. Blue represents the official languages, green the Indigenous languages and red the immigrant languages. The value with no assigned weight represents the languages in the dataset, but with no data for the specific year. An interactive version of this visualization can be found in the gitlab repository.

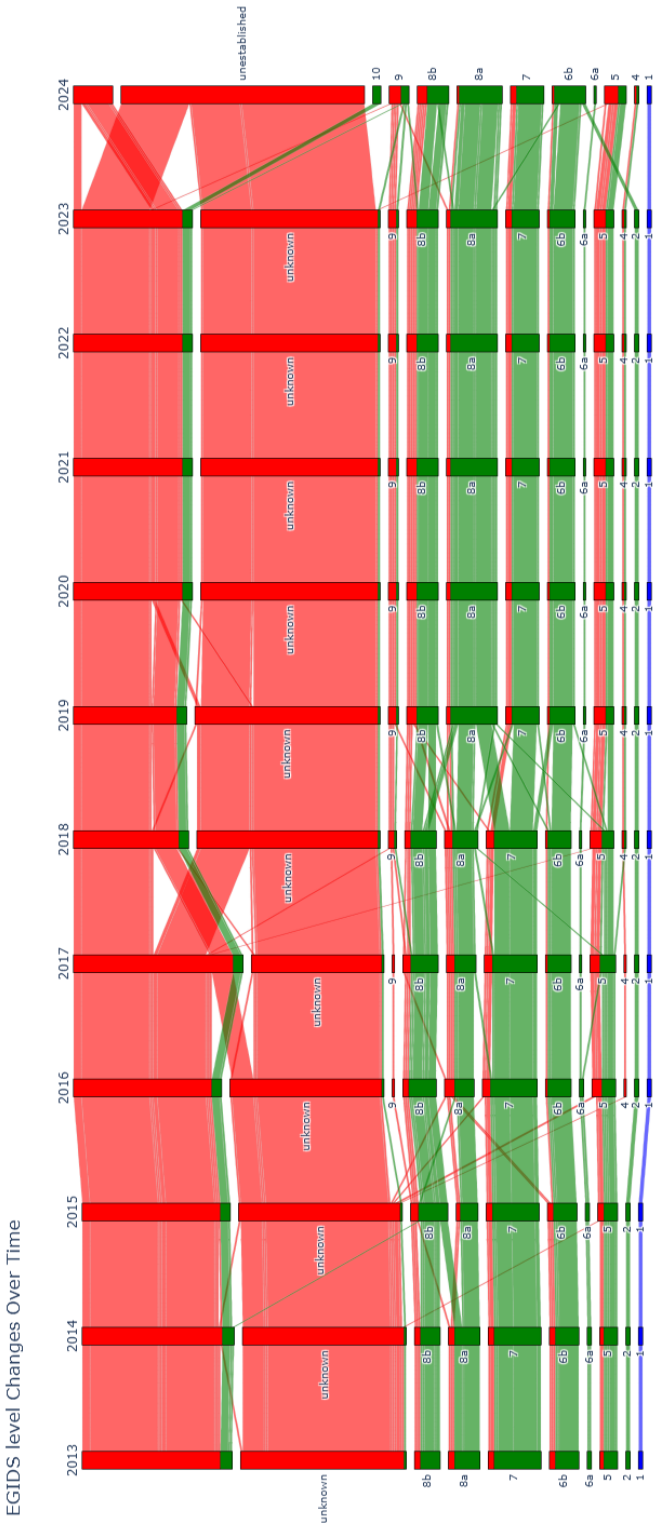

Figure 7.13: This figure shows an alluvial diagram that visualizes the EGIDS score changes over time. The dimensions are the years and the values of the dimensions are the EGIDS levels, "unknown" and "unestablished" form their own level. Blue represents the official languages, green the Indigenous languages and red the immigrant languages. The value with no assigned weight represents the languages in the dataset, but with no data for the specific year.

7.9 Red List Index

To get more insides on the overall status of the Canadian languages, the Red List Index (RLI) is computed. The weights are set like described in chapter 5.6. The status of official languages are constant 'living' and 'national', respectively, for EGIDS. They have the highest status in the dataset. Therefore the RLI is also high and the languages are considered the most stable in Canada. The RLI on the status is 1, so the maximum. The RLI based on the EGIDS however is for the official languages at 0,91. This is due to the fact, that above 'national', there exists the level 'international', which has the weight 0. It is notable that since the EGIDS is given based on the country level, the maximum score cannot be reached, but is 'national'. The change of the status

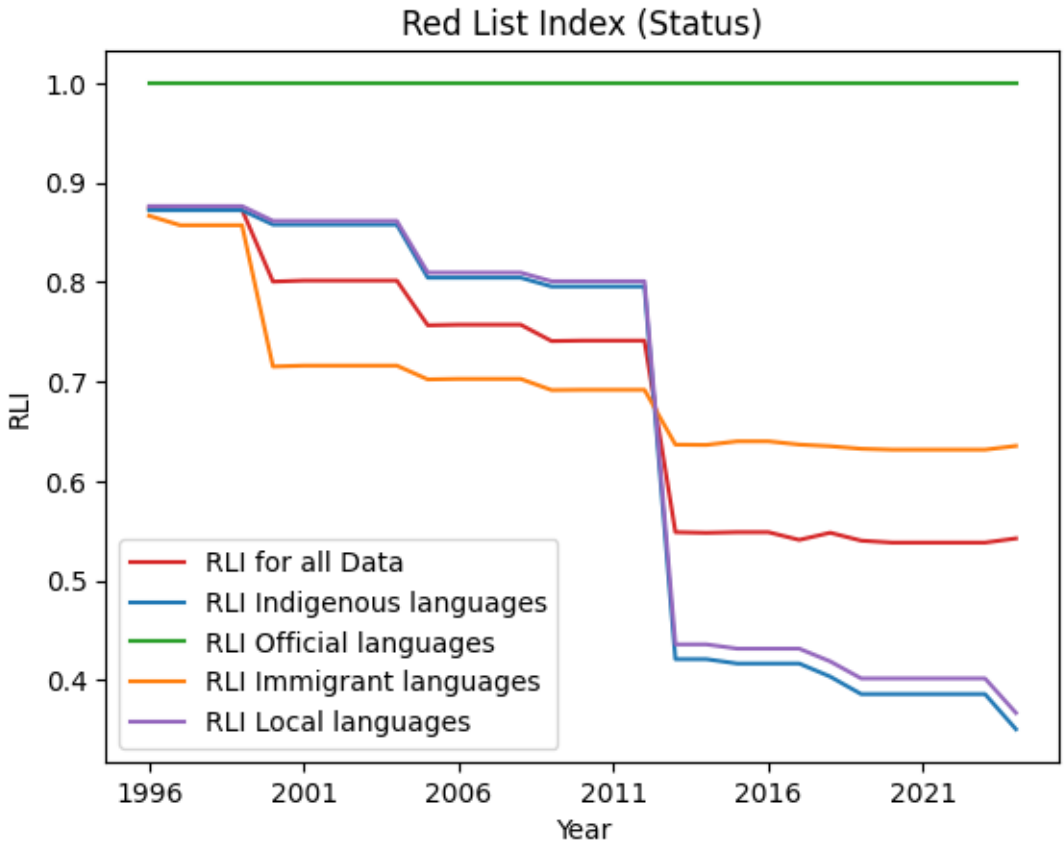

Figure 7.14: This plot shows the Red List Index for Canada between 1996 and 2021 based on the status with the cluster defined for this analysis. The weights for computing this RLI can be find in table 5.2

with the introduction of the EGIDS in 2013 that is seen in the previous chapter is also visible in figure 7.14. It shows a big drop between 2012 and 2013 in the stability of the languages. The RLI of Indigenous languages started in 1996 at 0.87 and decreased until 2012 to 0.8. After the

drop the RLI decreased further from 0.42 to 0.35. This means the Indigenous languages are not stable in Canada and highly under threat. The RLI of the immigrant languages also drops with the introduction of the EGIDS, but not that significantly. More important is the drop with the edition of 2000 from 0.86 to 0.72. This is related to the inclusion of 41 immigrant languages as 'unknown', and therefore with the weight one. In the following the linguistic richness increases, with mostly unclassified immigrant languages. From 2000 to 2013 the RLI decreases to 0.64 and rests around this niveau until 2024. The RLI of the local languages is just slightly higher than the one of the Indigenous languages, showing the official languages do not have an impact on it, since the metric is more based on the linguistic richness than on the linguistic evenness. Which is the only way to make the endangerment of the languages visible. The RLI of all languages is between the RLI of Indigenous and immigrant languages.

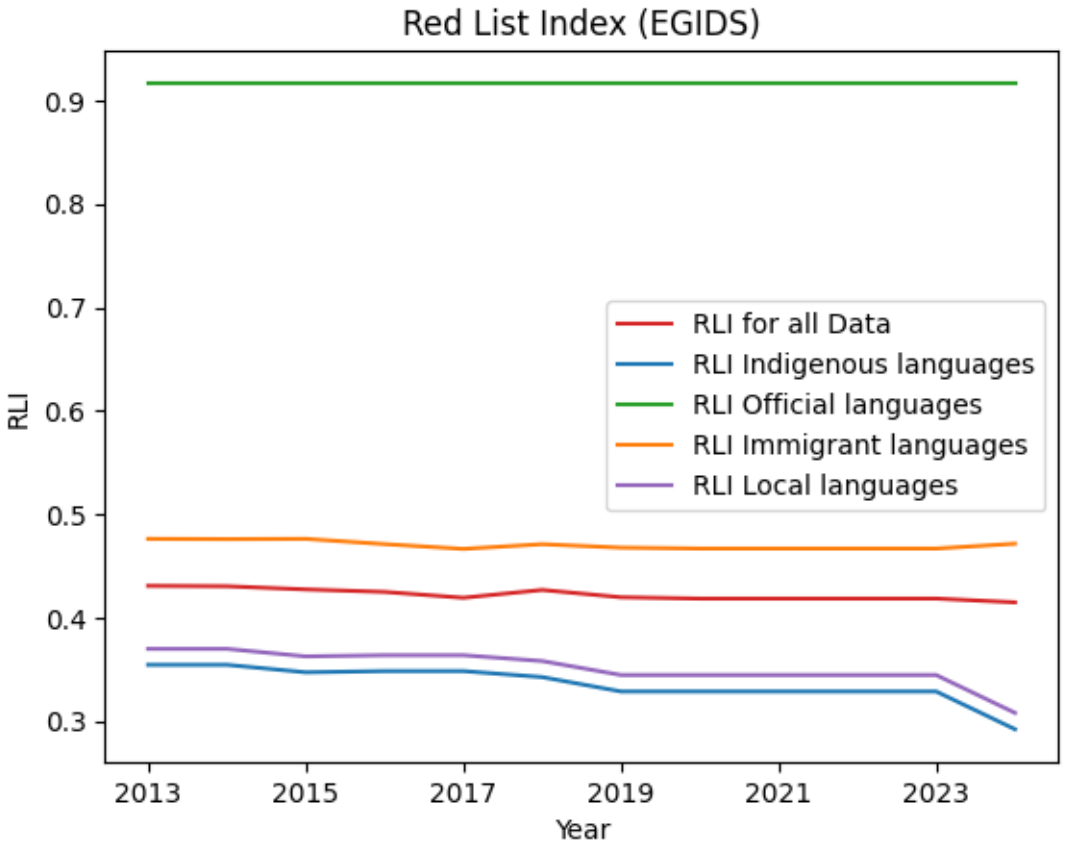

Figure 7.15: This plot shows the Red List Index for Canada between 1996 and 2021 based on the EGIDS. The weights for computing this RLI can be find in table 5.3.

The figure 7.15 shows the RLI based on the EGIDS with 'unknown' and 'unestablished' set to the weight 6. The RLI is quite constant over time for the immigrant languages, which is in accordance with the RLI seen based on the status. The Indigenous languages decrease from 0.35

to 0.29. In comparison with the RLI computed only on the EGIDS ignoring the 'unknown' and 'unestablished' languages (figures 10.21 and 10.22), the differences between the RLI of the Indigenous languages and of the immigrant languages are smaller. Also the immigrant languages have a smaller RLI 2013 than the Indigenous languages. Otherwise, the Indigenous languages have the smallest RLI from 2015 onwards. The change of the weights has no effect on the Indigenous languages. The RLI of the immigrant languages increases from 2013 to 2016 from 0.33 to 0.36 and decreases after that to 0.32 in 2024.

In general, it can be seen that the Indigenous languages are less stable and more threatened. Their stability decreases also the steepest in the last 30 years. The status of immigrant languages are not stable, either, but more or less stable at their niveau. The official languages are in the best status, which is not surprising. This confirms the previous findings, and the second part of the hypothesis, that Indigenous languages are the languages in Canada, that have the worst standing.

8 Results

	All languages	Official languages	Indigenous languages	Immigrant languages	Local languages
Language size	↘	→	↘	↘	↘
Richness	↗	→	↗	↗	↗
Abundance	↗	↗	↗	↗	↗
Evenness	→	↘	↘	↗	↘
ILD	↗	→	↗	→	↘
LDI	↗	↘	↗	↗	↘
Entropy	↗	↘	↗	↗	↘
Hill numbers	↗	↘	↗	↗	→
Status	↗	→	↗	↗	↗
EGIDS	↗	→	↗	↗	↗
RLI status	↘	→	↘	↘	↘
RLI EGIDS	↘	→	↘	↘	↘

Table 8.1: This table summarizes the findings of the analysis. The arrows indicate the direction of the global trend in the timeframe analyzed by the measure. The gradient of the arrow does not reflect the gradient of the curve calculated. Some changes are minimal. This overview is a simplification, for more details consult the chapter 7 or read the results.
Explanation: higher richness/ evenness/ ILD/ LDI/ entropy/ Hill number = higher diversity; higher status and EGIDS = higher endangerment level; higher RLI = higher stability

The official languages are the least diverse, but the most stable. They are the fewest languages, but cover the largest part of the population. Over the last 30 years, the number of speakers has increased, but their relative abundance has decreased. English has also become slightly more dominant compared to French. Linguistic diversity, regardless of the measure (apart from linguistic evenness), has otherwise remained fairly constant, and there have been no changes in status. The changes seen in the overview table 8.1 are minimal, and it can be said, also in comparison to the other language groups, that the diversity does not change much.

Indigenous languages are in the worst position in Canada. Although they have become more

diverse within their group (linguistic richness, ILD, entropy, and hill numbers) and are quite diverse (LDI), their relative frequency is declining and they are becoming noticeably more endangered. They have little influence on the total Canadian linguistic diversity, as they have already been marginalized too far from the linguistic landscape. The combination of the analysis of Indigenous languages as own group and together with the official languages as the local languages group give valuable insights on their endangerment and their invisibility.

The immigrant languages group is the most diverse in Canada, and its diversity has increased over the last 30 years (linguistic richness, relative abundance, LDI, entropy, Hill numbers), with only the ILD declining. In addition, these languages are becoming more endangered, although they are not as severely affected as Indigenous languages. Interpreting their status is relatively difficult, as information on the status of many immigrant languages is not available. Nevertheless, immigrant languages have remained stable since 2013, with developments in individual languages being offset by counter-developments in other languages.

In terms of linguistic evenness, local languages are dominated by official languages, but Indigenous languages are relevant for linguistic richness. The dominance of official languages can be seen in most measures of diversity, and local languages show low diversity, which remains practically unchanged with a slight downward trend (LDI, entropy, Hill numbers). However, diversity has decreased significantly over the last 30 years when the ILD is taken into account. This highlights the discrepancy between official and Indigenous languages. In terms of stability, however, there has been a clear negative trend (RLI), which is dominated by linguistic richness and thus indigenous languages.

Similar to local languages, different language groups have a greater influence on the diversity of Canada's overall linguistic landscape, depending on the measurement used. In terms of linguistic evenness, the official languages dominate again, but the immigrant languages within the dataset have the highest richness. Linguistic richness and abundance are increasing, as is linguistic evenness, because the official and immigrant languages are becoming more balanced. Overall, the language situation is moderately diverse and is becoming more diverse over the last 30 years (LDI, entropy, Hill numbers), and the ILD has also increased, although it has been declining since 2016. The generally positive picture of the development of linguistic diversity should be viewed with caution, because overall, the languages in Canada and their status are deteriorating, which means that most languages are losing stability.

9 Conclusion

The study began with a qualitative description of Canada's history and language policy. Particular emphasis is placed on the discrimination against the Indigenous population throughout history, which has its origins in colonization. Today, Canada pursues a policy of multiculturalism and multilingualism. This is important both for the country's official bilingualism and for its migration policy.

Furthermore, previous research was presented, highlighting the necessity of this study due to a research gap. This study used already known methods, which were also explained. Since good data are crucial for the analysis and no diachronic dataset was available in a usable form, it was newly created for this study. The underlying data for the analysis came from eleven editions of *Ethnologue* and six censuses. These two data sources were presented, and the creation of the diachronic dataset was described. The workflow included data collection through web scraping and information extraction with few-shot prompting, data cleaning and preparation by merging the data sources, and imputation of missing data. This master's thesis also serves as a guide on how to create a diachronic dataset using *Ethnologue*. It should be noted, however, that the workflow is still tailored specifically to Canada, as it involves many individual decisions and manual corrections. The first step in particular, web scraping, can be reused for other countries and can be adopted in whole or in part, depending on resources and requirements, and combined with customized data preprocessing. The code and data processing scripts are publicly available in the GitHub repository (`https://github.com/julianebe/DiaLingDivCan`), ensuring transparency and reproducibility. Due to licensing restrictions, the *Ethnologue* data themselves cannot be shared, but all intermediate files and aggregation scripts are accessible.

With this preparation, the statistical analysis could be conducted. After a descriptive analysis of the linguistic situation in Canada, linguistic richness, linguistic evenness, the Index of Linguistic Diversity, the Linguistic Diversity Index, Shannon entropy, and, in comparison, linguistic diversity using the Hill Numbers Framework were examined based on the dataset of speaker numbers. In addition, linguistic diversity was examined, as revealed by the status and status change of the languages. The Red List Index was also calculated for this purpose.

The answer to the research question "How has linguistic diversity in Canada changed in the

last 30 years?" can be summarized as follows: linguistic diversity increased in Canada in the last 30 years. However, this can be traced back to the fact, that the language groups of official languages and immigrant languages become more even and the number of immigrant languages and their speaker increases. At the language group level, there are no major changes in the linguistic diversity of the official languages. Immigrant languages are becoming more diverse, which can be linked to the multiculturalism policy. The opening up of the country can also be seen in linguistic diversity. The group of Indigenous languages is becoming more diverse within its own group. However, the extremely difficult situation, which is rooted in history, is evident in the fact that the Indigenous languages do not really influence the overall linguistic diversity of the country. This study also clearly demonstrated through the local languages group the dominance of official languages over Indigenous languages. When looking at linguistic diversity based on status, it becomes clear how bad the situation is for Indigenous languages. It also shows that over the last 30 years, Indigenous languages have become increasingly threatened and are losing stability.

The hypothesis, that linguistic diversity declines as more languages become endangered has to be rejected. In total, this study paints a contrasting picture: linguistic diversity increases even though the languages become more endangered. In overall, the Canadian linguistic landscape is becoming more diverse (LDI values increased from 0.55 to 0.62 and entropy from 1.3 to 1.7) while the languages become less stable overall (RLI status drops from 0.87 to 0.54). This can also be seen for the group of Indigenous languages, which is why the hypothesis claiming that Indigenous languages particularly support the first hypothesis must also be rejected. But it must be noted, that this does not contradict the underlying meaning of the hypothesis, that is, the difficult situation of the Indigenous languages in total. The increase of the linguistic diversity cannot be linked to the better standing of the Indigenous languages in Canada throughout the last 30 years. Especially when the group of local languages is explored, it can be seen, that the Indigenous languages do not play a big role, and were already invisible in the linguistic landscape of Canada so that they are not decisive for the increase of linguistic diversity.

A comparison of the methods shows that the diversity measurements, linguistic richness, linguistic diversity index, Shannon entropy, and thus also the Hill numbers, behave in the same way for the individual language groups, with immigrant languages having the greatest linguistic diversity and official languages being the least diverse. Only in the case of richness (also Hill number of order q=0) do the combined diversities of all languages or local languages deviate from the line, as they are greater than in the other measures. The Index of Linguistic Diversity is, by definition, the only diachronic measure, whereby it is always relative to the set starting point, which distorts the results. Still the decline of diversity in the local languages is notable. The Red

List Index for Canadian languages is based on the subset of languages based on the *Ethnologue* editions that includes the status of the languages. Since it does not directly calculate diversity but is rather an indicator of the stability or threat to the languages, it paints a different picture, but one that supports the statements on linguistic diversity, with the official languages as the most stable languages and the Indigenous languages having become more threatened over the last 30 years.

The limitations of this study will be discussed below, followed by suggestions for further research.

9.1 Limitations

The study can only be as good as the data on which it is based. Even though efforts were made to achieve a high quality dataset, it is not perfect. Firstly, the dataset depends on the accuracy of the data in *Ethnologue* and the census. This is beyond the scope of this study. However, as described in chapter 3, it is the best option available. Related to this is the fact that the timeliness of the data for the individual languages varies, which is why the second data source, the census, is advantageous and many comparable data points could be integrated into the dataset at the same time. Furthermore, the LLM in the information extraction step also influences the accuracy of the data, and errors cannot be ruled out. The dataset is also influenced by the researcher's decisions, and mistakes cannot be precluded, especially in the final merging process to create the final dataset. However, they have been minimized as much as possible through several rounds of corrections and clear rules in the decisions. Not all *Ethnologue* editions since 1996 have been used. The missing editions have less influence on the speaker data, as this data is not up to date anyway and does not change every year. The analysis on the speaker numbers and on the status does not rely on the same number of languages. However, since the grouping follows the same rules, the results can still considered complementary and usable at this stage. Missing data were imputed using a single imputation method. It cannot be ruled out that more complex imputation methods would have yielded more accurate results. However, this could not have been proven, and since diachronic analysis is more about trends than exact values, a more complex method was not considered necessary. Despite the imputation, the dataset still has gaps. This is partly because it includes languages for which there are no L1 speaker numbers or status information available, and partly because some languages do not appear in the dataset at all or do not appear throughout the entire period, which cannot be remedied by imputation. Compared to the total population, the speaker numbers dataset always lacks approximately 10% of the population. This raises the question of whether linguis-

tic diversity within Indigenous languages would still increase if more data were available and linguistic richness included all languages for the whole time frame.

The workflow created for generating the dataset is structured in such a way that it can be adopted for similar projects. However, the workflow is not fully automated due to the many manual corrections and decisions required. Nevertheless, the modules and functions are reusable.

The analysis also has its limitations. It is merely descriptive in nature and could be expanded to identify the reasons behind the changes in linguistic diversity and what the drivers of these changes are, in order to be able to respond even more effectively in the fight against language loss. Furthermore, linguistic diversity is only examined at the level of language groups (official, Indigenous, immigrant languages) and their combinations (all languages and local languages). Linguistic diversity was examined exclusively on the basis of speaker numbers and status, and the relationship between languages, such as interlinguistic distance, was not taken into account. Linguistic diversity based on the number of speakers is calculated using L1 data. However, L1 does not say anything about actual language skills, and passive comprehension could be higher than active language skills.

9.2 Outlook

Further research could, of course, address these limitations. The completeness of the data could be improved by integrating the missing editions from 2020. Big changes in the speaker numbers are not expected, but it would be interesting for the analysis of the status, as the status, and the EGIDS, depend on the edition. The dataset could also be expanded to include non-digital data from before 1996, if desired. Additional imputation methods could be tested in the workflow and the results could be compared.

The analysis reveals many possibilities for further research into Canada's linguistic diversity. The analysis is already quite comprehensive, but in terms of evenness, for example, diversity at the language level and changes between years could be examined in even greater detail. The languages selected in this study provide a good overview for entering the discussion about official languages. It would be advantageous to examine English and French separately and to form another language group from the non-official languages that includes Indigenous and immigrant languages. Instead of examining linguistic diversity by language group, as is done here, it would be interesting to find out how linguistic diversity and, above all, language endangerment has developed for language families. To move away from linguistic diversity measures only on the level of speaker numbers and status, it would be interesting to see if similar

trends for the three language groups (official languages, Indigenous languages, and immigrant languages) are visible, if the similarity of the languages is also taken into account by adding weights to the languages in the statistical analysis or even to focus on the functional diversity of languages solely.

The geographical scope also reveals further research opportunities. On the one hand, it would be possible to compare Canada with other countries; the workflow for creating the datasets for the other countries can be adopted from this study. For example, a study on linguistic diversity in the USA would be interesting knowing the recent change in the language policy that declared English as official language. With such a study, compared and combined with the knowledge of linguistic diversity in Canada, one would gain insides on linguistic diversity in North America and if the findings are Canada specific or not. An alternative would be to look into the linguistic diversity in Australia to find out if the trends in Canada are similar and therefore global and not restricted to North America. Australia is just like Canada a big country of the Commonwealth with similar characteristics in history, politics and economic power, which leads to the expectations that the language dynamics are comparable and linguistic diversity trends similar. However, it is also possible to move beyond the national level and integrate a diachronic regional comparison of linguistic diversity based on provinces or cities, or the degree of urbanity, as has been done in some previous research. However, the dataset created for this study cannot serve as a basis for such a study, as the geographical dimension has not been scaled. Such an analysis would be possible, however, on the basis of census data.

The analysis could also be deepened, as already mentioned in the previous chapter, in which linguistic diversity was discussed in connection with influences and drivers. With regard to Indigenous languages, it would be interesting to take a closer look at the influence of colonization or migration for immigrant languages. In order to gain a better insight into the linguistic reality of people in Canada, a qualitative study, that also takes the nuances of active and passive language skills into consideration, could be beneficial.

The chosen time span of the last 30 years also reflects the digital age, and the connection between linguistic diversity and digitalization would be another interesting field of research. Here, too, a mixed-method study with quantitative and qualitative analysis would be an interesting approach.

10 Appendix

10.1 Census

References: Statistics Canada (1997, 1999, 2004a, 2004b, 2010a, 2010b, 2012a, 2012c, 2017b, 2018a, 2023b, 2023c)

Mother tongue

1996, 2001, 2006

Question: What is the language that this person **first learned** at home **in childhood** and **still understands**?

Definition: Refers to the first language learned at home in childhood and still understood by the individual at the time of the census.

2011

Question: What is the language that this person **first learned** at home **in childhood** and **still understands**?

Definition: Refers to the first language learned at home in childhood and still understood by the individual on May 10, 2011.

2016

Question: What is the language that this person **first learned** at home **in childhood** and **still understands**?

Definition: 'Mother tongue' refers to the first language learned at home in childhood and still understood by the person at the time the data was collected. If the person no longer understands the first language learned, the mother tongue is the second language learned. For a person who learned two languages at the same time in early childhood, the mother tongue is the language this person spoke most often at home before starting school. The person has two mother tongues only if the two languages were used equally often and are still understood by the person. For a child who has not yet learned to speak, the mother tongue is the language spoken most often to this child at home. The child has two mother tongues only if both languages are spoken equally often so that the child learns both languages at the same time.

2021

Question: What is the language that this person **first learned** at home **in childhood** and **still understands**?

Definition: **Mother tongue** refers to the first language learned at home in childhood and still understood by the person at the time the data was collected. If the person no longer understands the first language learned, the mother tongue is the second language learned. For a person who learned more than one language at the same time in early childhood, the mother tongue is the language this person spoke most often at home before starting school. The person has more than one mother tongue only if they learned these languages at the same time, and still understands them. For a child who has not yet learned to speak, the mother tongue is the language spoken most often to this child at home. A child who has not yet learned to speak has more than one mother tongue only if these languages are spoken to them equally often so that the child learns these languages at the same time.

Knowledge of official languages

1996, 2001

Question: Can this person speak English or French well enough to conduct a conversation?

Definition: Refers to the ability to conduct a conversation in English only, in French only, in both English and French or in neither of the official languages of Canada.

2006, 2011

Question: Can this person speak English or French well enough to conduct a conversation?

Definition: Refers to the ability to conduct a conversation in English only, in French only, in both English and French, or in neither English nor French.

2016

Question: Can this person speak English or French well enough to conduct a conversation?

Definition: 'Knowledge of official languages' refers to whether the person can conduct a conversation in English only, French only, in both or in neither language. For a child who has not yet learned to speak, this includes languages that the child is learning to speak at home.

2021

Question: Can this person speak English or French well enough to conduct a conversation?

Definition: **Knowledge of official languages** refers to whether the person can conduct a conversation in English only, French only, in both or in neither language. For a child who has not yet learned to speak, this includes languages that the child is learning to speak at home.

Knowledge of non-official languages

1996, 2001, 2006

Question: What language(s), **other than English or French**, can this person speak well enough to conduct a conversation?

Definition: Refers to languages, other than English or French, in which the respondent can conduct a conversation.

2011

Question: -

Definition: -

2016

Question: What language(s), **other than English or French**, can this person speak well enough to conduct a conversation?

Definition: 'Knowledge of non-official languages' refers to whether the person can conduct a conversation in a language other than English or French. For a child who has not yet learned to speak, this includes languages that the child is learning to speak at home. The number of languages that can be reported may vary between surveys, depending on the objectives of the survey.

2021

Question: What language(s), **other than English or French**, can this person speak well enough to conduct a conversation?

Definition: **Knowledge of non-official languages** refers to whether the person can conduct a conversation in a language other than English or French. For a child who has not yet learned to speak, this includes languages that the child is learning to speak at home. The number of languages that can be reported may vary between surveys, depending on the objectives of the survey.

10.2 Code

The code for the whole workflow and the analysis can be found in the following github repository: `https://github.com/julianebe/DiaLingDivCan`. The datasets cannot be shared because of the proprietary character of *Ethnologue*. In the repository also the visualizations from this Master's thesis are stored.

The step information extraction of the workflow uses LLMs from OpenAI as tool. Furthermore, ChatGPT was used for this Master's thesis as coding assistant.

10.3 Description HTML Structure

This section describes the HTML structure.

Ethnologue 1996

Description of the HTML structure of *Ethnologue* 1996 (see 10.1)
The summary of the country is as a text in the first `<p>` tag.
The information on the languages can be found in the tags `<p class="HIN">`.
The language name is inside the first `<a>` tag.
Alternate names are in the `<i class="NAL">` tag, if any.
The abbreviation is inside squared brackets.
All other information is written as a string following this information. Any links (for the language family) are ignored.

```
<h1 align="CENTER">Canada</h1>
▼<p>
  "27,567,000 (1993 Johnstone). Indian (800,000) and Eskimo (32,000) ethnic total (1993): 146,285
  mother tongue speakers (1981 census). Literacy rate 96% to 99%. Also includes Arabic 93,000, Chinese
  780,000, from India and Pakistan 280,000, from the Philippines 95,000, speakers of many European
  languages. Information mainly from Chafe 1962, 1965, SIL 1991. Christian, secular, Muslim, Jewish.
  Blind population 27,184. Deaf institutions: many. Data accuracy estimate: Data accuracy estimate: A1,
  A2. The number of languages listed for Canada is 79. Of those, 76 are living languages and 3 are
  extinct."
</p>
▼<p class="HIN">
  ▼<b class="PNAM">
    <a name="ABE">ABNAKI-PENOBSCOT</a>
  </b>
  " ["
  <a href="/web/19990424112038/http://www.sil.org/ethnologue/lookup?ABE">ABE</a>
  "] 20 speakers (1991 M. Krauss) out of 1,800 population including USA (1982 SIL). Total population
  probably evenly divided between the two dialects. Quebec on St. Lawrence River between Montreal and
  Quebec City. "
  <a href="/web/19990424112038/http://www.sil.org/ethnologue/families/Algic.html">Algic</a>
  ", Algonquian, Eastern. Dialects: WESTERN ABNAKI (ABENAKI, ST. FRANCIS, ABENAQUI), PENOBSCOT (EASTERN
  ABNAKI). Dictionary. Grammar. Bible portions 1844. Nearly extinct."
</p>
▼<p class="HIN">
  ▼<b class="PNAM">
    <a name="ALG">ALGONQUIN</a>
  </b>
  <i class="NAL">(ALGONKIN)</i>
  " ["
  <a href="/web/19990424112038/http://www.sil.org/ethnologue/lookup?ALG">ALG</a>
  "] 3,000 speakers out of 5,000 population (1987 SIL). Southwestern Quebec, northwest of Ottawa and in
  adjacent area of Ontario. "
  <a href="/web/19990424112038/http://www.sil.org/ethnologue/families/Algic.html">Algic</a>
  ", Algonquian, Central, Ojibwa. Several dialects. In the west children prefer the national language,
  although some may speak Algonquin; most adults speak Algonquin; young adults may prefer the national
  language. Elsewhere Algonquin is the principal means of communication and spoken by the majority of
  all ages. 75% to 100% literate. Bible portions 1980-1993. Work in progress."
</p>
```

Figure 10.1: HTML Structure *Ethnologue* 1996

Ethnologue 2000-2009

The HTML structure (see 10.2) is not exactly the same for all years (also the aesthetics are different), but the structure containing the information of interest is similar.
The summary of the country is as a text in the tag `<blockquote>`.
The list of languages can be found in the `<table cellspacing="12">` tags. Each language is inside the `<tr valign="TOP">` tag.
The name is inside the `<td width="25%">` tag.
The information on the language is in `<td width="75%">` inside a paragraph tag.
The abbreviation (2000) or the ISO code (since 2005) is inside squared brackets. Followed by details on the language (see table 4.1). This paragraph however is more structured by highlighting the features (like “Alternate names”, “Classification” or “Dialects”) with the `<i>` tag, which allows to extract the information more easily.
If the language is endangered, it is visible through a link `<a href="nearly_extinct.asp">`.

```
<h1 align="CENTER">Languages of Canada</h1>
▸ <p align="CENTER"> … </p>
▾ <blockquote>
    "National or official languages: English, French. 30,563,040 (1998 UN). Indian (800,000) and Eskimo
    (32,000) ethnic total (1993): 146,285 mother tongue speakers (1981 census). 4,120,770 non-English or
    French mother tongue, or 15.3% (1991 census). Literacy rate 96% to 99%. Also includes Afrikaans
    2,353, Tosk Albanian 29, Judeo-Moroccan Arabic, Najdi Spoken Arabic 28,550, Armenian 2,265, Assyrian
    Neo-Aramaic 5,000, Belarusan 2,280, Bulgarian 1,630, Chaldean Neo-Aramaic, Yue Chinese 250,000,
    Corsican, Czech 27,780, Danish 27,395, Dutch 159,000, Estonian 14,520, Western Farsi 15,000, Finnish
    36,725, Western Frisian, Irish Gaelic, Scots Gaelic 5,000, Standard German 470,505, Greek 104,455,
    Haitian Creole French 12,317, Hebrew, Hungarian 86,835, Icelandic, Italian 514,410, Iu Mien 100,
    Japanese 43,000, Kashubian, Korean 73,000, Lao, Latvian 15,000, Lithuanian, Macedonian 3,995,
    Maltese, Bokmaal Norwegian 27,405, Nung, Eastern Panjabi 214,530, Western Panjabi, Plains Indian Sigr
    Language, Polish 222,355, Pontic, Portuguese 222,870, Vlax Romani, Romanian 11,300, Russian 31,745,
    Serbo-Croatian 93,860, Sinhala 3,004, Slovak 17,370, Slovenian 6,415, Southwestern Caribbean Creole
    English, Spanish 228,580, Swedish 21,650, Tagalog 158,210, Turkish 8,863, Turoyo, Ukrainian 174,830,
    Vietnamese 60,000, Welsh 3,160, Eastern Yiddish 49,890, Arabic 166,150 (1998), Chinese 736,015
    (1998), from India and Pakistan 280,000, speakers of many European languages. Information mainly from
    W. Chafe 1962, 1965; SIL 1951-1999. Christian, secular, Muslim, Jewish. Blind population 27,184. Deaf
    population 1,704,551. Deaf institutions: many. Data accuracy estimate: A1, A2. The number of
    languages listed for Canada is 90. Of those, 85 are living languages and 5 are extinct. Diversity
    index 0.55."
    <br>
  </blockquote>
  <h3>Spoken living languages</h3>
▾ <table cellspacing="12">
  ▾ <tbody>
    ▾ <tr valign="TOP">
        <td width="25%">ABNAKI, WESTERN</td>
      ▾ <td width="75%">
        ▾ <p>
            "[ABE] 20 speakers (1991 M. Krauss) out of 1,800 ethnic population including Eastern Abnaki ir
            USA (1982 SIL).  Quebec on St. Lawrence River between Montreal and Quebec City. "
            <i>Alternate names:</i>
            " ABENAKI, ABENAQUI, ST. FRANCIS.  "
            <i>Classification:</i>
            " Algic, Algonquian, Eastern.  "
            <a href="nearly_extinct.asp">Nearly extinct.</a>
            <a href="show_language.asp?code=ABE">More information.</a>
          </p>
        </td>
      </tr>
    ▾ <tr valign="TOP">
        <td width="25%">ALGONQUIN</td>
      ▾ <td width="75%">
        ▾ <p>
            "[ALG] 2,275 mother tongue speakers (1998 Statistics Canada) out of 5,000 population (1987
            SIL).  Southwestern Quebec, northwest of Ottawa and in adjacent areas of Ontario
            (Golden Lake and Maniwaki). "
            <i>Alternate names:</i>
            " ALGONKIN.  "
            <i>Classification:</i>
            " Algic, Algonquian, Central, Ojibwa.  "
            <a href="show_language.asp?code=ALG">More information.</a>
          </p>
        </td>
      </tr>
    ▸ <tr valign="TOP"> … </tr>
```

Figure 10.2: HTML Structure *Ethnologue* 2000

Ethnologue 2013-2019

The information for these editions is on two websites.
The country website contains the summary, the languages are on a subsite from this.
The information on the summary (see 10.3) is inside the first `<div class="view-content">` tag.
The Information is further structured by label and content. The labels can be accessed through the `class "views-label"`, which is sometimes inside a `<em>` and sometimes inside a `<span>` tag. The content is in `<div class="field-content">`. The section is structured like a table but without the specific table tags.
The language information (see 10.4) on the other webpage is also inside the first occurrence of the tag `<div class="view-content">`. Each language is in a row defined by `<div class="views-row">`. Due to the nested structure of the divisions, it is possible to access the same data through different tags.
The language name is in `<div class="title">`.
The information on the language is inside `<div class="content">`
The ISO code is inside the first `<a>` tag in squared brackets.
The other information follows the ISO code, first with the details, then the features highlighted through the `<em>` tag.

```
<div class="view-content">
  <div class="views-row views-row-1 views-row-odd views-row-first views-row-last">
    <div class="views-field views-field-field-population-figure">
      <em class="views-label views-label-field-population-figure">Population</em>
      <div class="field-content">31,613,000</div>
    </div>
    <div class="views-field views-field-field-national-languages">
      <em class="views-label views-label-field-national-languages">National Languages
      </em>
      <div class="field-content">English, French</div>
    </div>
    <div class="views-field views-field-field-literacy-rate">
      <em class="views-label views-label-field-literacy-rate">Literacy Rate</em>
      <div class="field-content">99% (2011 USDS)</div>
    </div>
    <div class="views-field views-field-field-immigrant-languages">
      <em class="views-label views-label-field-immigrant-languages">Immigrant
      Languages</em>
      <div class="field-content">
        "Afrikaans (9,810), Akan (14,600), Amharic (19,160), Arabic (374,000),
        Armenian (31,700), Assyrian Neo-Aramaic (5,000), Belarusan (580), Bengali
        (64,500), Bosnian (12,200), Bulgarian (19,630), Central Khmer (21,100),
        Chaldean Neo-Aramaic, Corsican, Croatian (52,300), Czech (24,500), Danish
        (14,900), Dutch (121,000), Eastern Frisian (2,200), Eastern Panjabi (460,000),
        Eastern Yiddish (16,300), Finnish (18,500), Greek (118,000), Gujarati
        (101,000), Haitian (102,000), Hakka Chinese (5,460), Hebrew (20,200), Hindi
        (106,000), Hmong Njua, Hungarian (71,100), Icelandic (1,670), Ilocano
        (19,500), Iranian Persian (177,000), Italian (438,000), Japanese (43,000),
        Judeo-Moroccan Arabic, Kannada (3,140), Kashubian, Konkani (3,540), Korean
        (143,000), Lao (14,200), Latgalian, Lithuanian (7,600), Macedonian (18,300),
        Malay (11,900), Malayalam (17,700), Maltese (7,000), Mandarin Chinese
        (255,000), Marathi (6,660), Min Nan Chinese (18,820), Najdi Spoken Arabic
        (20,000), Northern Kurdish (10,300), Northern Pashto (13,100), Norwegian
        (6,280), Nung, Plains Indian Sign Language, Polish (201,000), Pontic,
        Portuguese (226,000), Romanian (93,100), Russian (170,000), Scottish Gaelic
        (2,320), Serbian (58,500), Sindhi (12,900), Sinhala (15,700), Slovak (18,500),
        Slovene (11,300), Somali (33,700), Southeastern Dinka (15,900), Southwestern
        Caribbean Creole English, Spanish (439,000), Standard Estonian (6,610),
        Standard German (430,000), Standard Latvian (6,460), Swahili (12,100), Swedish
        (8,100), Sylheti, Tagalog (384,000), Tamil (143,000), Telugu (10,700), Thai
        (8,420), Tigrigna (10,800), Tongan, Turkish (31,200), Turoyo, Ukrainian
        (120,000), Urdu (194,000), Vietnamese (153,000), Vlax Romani, Welsh (1,440),
        Western Panjabi, Yue Chinese (389,000)"
      </div>
    </div>
    <div class="views-field views-field-field-deaf-population">…</div>
    <div class="views-field views-field-field-deaf-institutions">…</div>
    <div class="views-field views-field-field-country-references">…</div>
    <div class="views-field views-field-field-language-counts">…</div>
  </div>
</div>
</div>
```

Figure 10.3: HTML Structure *Ethnologue* 2013 Summary

```
▼<div class="view-content">
  ▼<div class="views-row views-row-1 views-row-odd views-row-first">
     <div class="title"> Abenaki, Western </div>
    ▼<div class="content">
       <a href="/web/20131204112434/http://www.ethnologue.com/language/abe">[abe]</a>
       " Quebec, Odanak Reserve on the Saint Francois River. Also in United States. 10
       in Canada (Golla 2007). "
       <em>Status:</em>
       " 8b (Nearly extinct). "
       <em>Alternate Names:</em>
       " Abenaki, Abenaqui, St. Francis, Western Abnaki "
       <em>Classification:</em>
       " Algic, Algonquian, Eastern Algonquian, Abenaki "
       <em>Comments:</em>
       " Used for source material for linguistic studies and language revival programs.
       "
      ▼<p>
         <a href="/web/20131204112434/http://www.ethnologue.com/language/abe">More
         Information</a>
       </p>
     </div>
   </div>
  ▼<div class="views-row views-row-2 views-row-even">
     <div class="title"> Algonquin </div>
    ▼<div class="content">
       <a href="/web/20131204112434/http://www.ethnologue.com/language/alq">[alq]</a>
       " Southwest Quebec, northwest of Ottawa and adjacent areas of Maniwaki and
       Golden Lake, Ontario. 1,760 (2011 census), decreasing. Less than 10%
       monolinguals. Ethnic population: 5,000 (1987 SIL). "
       <em>Status:</em>
       " 6b (Threatened). "
       <em>Alternate Names:</em>
       " Algonkin, Anishinaabemowin "
       <em>Dialects:</em>
       " Northern Algonquin, Southern Algonquin (Nipissing). Northern Algonquin and
       Southern Algonquin varieties very different. "
       <em>Classification:</em>
       " Algic, Algonquian, Ojibwa-Potawatomi "
       <em>Comments:</em>
       " Christian. "
      ▼<p>
         <a href="/web/20131204112434/http://www.ethnologue.com/language/alq">More
         Information</a>
       </p>
     </div>
   </div>
  ▶<div class="views-row views-row-3 views-row-odd">…</div>
```

Figure 10.4: HTML Structure *Ethnologue* 2013 Languages

Ethnologue 2024

The summary (see 10.5) of the country can be found through `<section id="summary">` which embraces a description list `<dl>`.
The labels are in `<dt>` and the information in `<dd>` tags.
The language information (see 10.6) is in `<section id="languages">`.
The name and the ISO code are in `<div class="languages__label">`, the name as a string and the code in the `<a>` tag.
The rest of the information on the language follows the division in an unordered list `<ul>`.
The `<ul>` contains the features highlighted by `<i>` with the information as strings.
Macrolanguages are a special case, where the information on the including languages is not marked by the html.
This is also valid for the Edition from 2025.

```
  <h1 class="title__home">Canada</h1>
</header>
▼<section class="section__country section" id="summary" aria-labelledby="summaryHeading">
  <h2 id="summaryHeading" class="section__title">Summary</h2>
 ▼<dl class="description-list well">grid
  ▼<dt class="entry__label">
     <a href="/methodology/#Population">Population</a>
   </dt>
   <dd class="entry__content">41,013,000</dd>
  ▼<dt class="entry__label">
     <a href="/methodology/#principal">Principal Languages</a>
   </dt>
   <dd class="entry__content">English, French</dd>
  ▶<dt class="entry__label">…</dt>
   <dd class="entry__content">99% (Roser and Ortiz-Ospina 2018)</dd>
  ▶<dt class="entry__label">…</dt>
  ▶<dd class="entry__content">…</dd>
  ▶<dt class="entry__label">…</dt>
  ▶<dd class="entry__content">…</dd>
  ▶<dt class="entry__label">…</dt>
   <dd class="entry__content">1,704,500</dd>
  ▶<dt class="entry__label">…</dt>
  ▶<dd class="entry__content">…</dd>
  ▶<dt class="entry__label">…</dt>
  ▶<dd class="entry__content">…</dd>
 </dl>
</section>
```

Figure 10.5: HTML Structure *Ethnologue* 2024 Summary

```
  <h1 class="title__home">Canada</h1>
 </header>
▸<section class="section__country section" id="summary" aria-labelledby="summaryHeading">…
 </section>
▾<section id="languages" class="section languages" aria-labelledby="languagesHeading">
  <h2 class="section__label" id="languagesHeading">Languages</h2>
 ▾<div class="description-list flow__list">  grid
  ▾<div id="abe" class="languages__label entry__label">
     "Abenaki, Western "
     <a class="chip" href="/language/abe">abe</a>
   </div>
  ▾<ul class="languages__content entry__content">
     <i>Location:</i>
     " Quebec province: Odanak Reserve on Saint Francois river."
     <br>
     <i>Users:</i>
     " 10 in Canada (Golla 2007). Total users in all countries: 14."
     <br>
     <i>Status:</i>
     " 8b (Nearly extinct). Language of recognized indigenous peoples: Odanak, Première Nation des
     Abénakis de Wôlinak."
     <br>
     <i>Alternate Names:</i>
     " Abenaki, Abenaqui, Alnombak, Saint Francis, Western Abnaki."
     <br>
     <i>Autonym:</i>
     " Alnôbak."
     <br>
     <i>Classification:</i>
     " Algic, Algonquian, Eastern Algonquian, Abenaki."
     <br>
   </ul>
  ▾<div id="aar" class="languages__label entry__label">
     "Afar "
     <a class="chip" href="/language/aar">aar</a>
   </div>
  ▾<ul class="languages__content entry__content">
     <i>Users:</i>
     " 2,500 in Canada (2024 IMB), based on ethnicity."
     <br>
     <i>Status:</i>
     " Unestablished."
     <br>
     <i>Classification:</i>
     " Afro-Asiatic, Cushitic, East, Saho-Afar."
     <br>
   </ul>
  ▸<div id="afr" class="languages__label entry__label">…</div>
  ▸<ul class="languages__content entry__content">…</ul>
```

Figure 10.6: HTML Structure *Ethnologue* 2024 Languages

10.4 Example Short texts Population

Following this examples from the *Ethnologue* editions 2000 and 2024 are shown. Including the name, the abbreviation or ISO code and the short text about the population.

***Ethnologue* 2000**


```
{
    "name":"Cree, Plains",
    "abbreviation":"crp",
    "population":"34,000 or more in Canada. Population total both
    countries 35,000 or more speakers out of 53,000 or more population
    (1982 SIL)."
},
{
    "name":"Munsee",
    "abbreviation":"umu",
    "population":"7 or 8 speakers out of 400 population (1991 M. Dale
    Kinkade)."
},
 {
    "name":"Plautdietsch",
    "abbreviation":"grn",
    "population":"80,000 or more first language speakers and 20,000
    second language speakers in Canada (1978 Kloss and McConnell).
    Total German mother tongue speakers in Canada including standard
    German, 561,000 (J.A. Hawkins in B. Comrie 1986). Population total
    all countries 400,000, of whom 150,000 use it habitually. 110,735
    or more in Latin America are fairly monolingual."
},
{
    "name":"Tlingit",
    "abbreviation":"tli",
    "population":"145 mother tongue speakers in Canada (1998
    Statistics Canada) out of 1,000 population in Canada (1995 M.
    Krauss)."
```

},

***Ethnologue* 2024**


```
{
    "name":"Assiniboine",
    "iso-code":"asb",
    "population":"370 in Canada (2021 census), all users. L1 users:
    190 (2021 census). No monolinguals (2021 census). Ethnic
    population: 3,500 in Canada and the United States (Golla 2007).
    Total users in all countries: 370 (as L1: 190; as L2: 180)."
},
{
    "name":"Babine",
    "iso-code":"bcr",
    "users":"520 (FPCC 2018). 295 semi-speakers (FPCC 2014). 100 fluent
    speakers and 100 passive speakers of Wetsuwet’en. 200 speakers of
    all degrees of fluency of Babine Proper (Golla 2007). No
    monolinguals (2021 census). Ethnic population: 5,400 (FPCC 2018)."
},
{
    "name":"Ditidaht",
    "iso-code":"dtd",
    "population":"7 (FPCC 2018). Ethnic population: 1,350 (FPCC 2018)."
},
{
    "name":"Pentlatch",
    "iso-code":"ptw",
    "population":"No known L1 speakers. Joe Nimnim, the last speaker,
    died in 1940."
},
```


10.5 Information Extraction Prompt

This is the prompt for the information extraction task:

```
prompt_old = """
Your task is to extract the information and structure it in the
following way.

You should extract any information on: user, L1 user, L2 user,
semi-speaker, monolinguals
The information should be specified if it's information on only
canada or not (all)
regarding the specification on canada, it should be mentioned
explicitely in the text or the source points to it
Furthermore the year should be added, the year is within the
source in brackets
If no year is provided for the information add the year of the
edition (ed_year) to the structure
You should also extract information all languages L1 and ethnic
population
For this information there is no specification canada or not
needed, but information on the year
Any further information should also be saved with the year of
the edition
No information can be lost, but not all possible features are in
one text.
Provide a json format.

###
like this the following information structure is wanted:
user_all_year/ user_canada_year/ user_all_ed_year/ user_canada_ed_year
L1_
L2_
semi_
monoling_
all_languages_L1_year/ all_languages_L1_ed_year
ethnic_population_
extra_info_ed_year
```

```
###
year should be changed into the real year
the edition year is {edition}
no known speaker is equivalent to 0 speaker
_canada_ if "in Canada" mentioned for the information or the source
is census or source is FPCC, source is Ministere de la Sante et des
Services Sociaux or Govt. report or MSSS, or the source contains
the words Canada or Quebec
a range is to be written in a list, with the min, max
the first number, if not stated precisely otherwise, is the number
for user all or canada.

###
Example 1 with output: {example1}  {output1}
Example 2 with output: {example2}  {output2}
Example 3 with output: {example3}  {output3}

###
extract the information for the following texts,
if a text is an empty string, the output for the text is an empty
json object
{user_text}

"""
```

10.6 Results Information Extraction

Edition	Number Language Entries	Size Test set (Annotated percentage)	Median Accuracy	Average Accuracy
1996	79	17 (21,52%)	100	78,65
2000	90	34 (37,78%)	100	73,58
2005	89	29 (32,58%)	100	79,76
2009	91	31 (34,07%)	100	88,33
2013	92	20 (21,74%)	100	79,17
2015	94	37 (39,36%)	100	84,07
2016	99	22 (22,22%)	100	88,61
2017	99	25 (25,25%)	100	98,00
2018	101	21 (20,79%)	100	86,34
2019	101	23 (22,77%)	100	82,61
2024	235	26 (11,06%)	100	90,73

Table 10.1: This Table shows the results of the information extraction before any corrections

10.7 Dataset creation

Manuel transfer from "L1_all"

Takeover to "L1_canada"

asb chp ono tsi

Discarded

pot moh str

Manuel transfer from "user_all"

Takeover to "L1_canada"

ara	slv	cmn	kat
hai	spa	mar	khk
hin	swe	pbu	har
oji	tgl	snd	hil
afr	tur	som	ibo
hye	ukr	dks	kab
aii	vie	swa	kin
bel	cym	tam	lin
bul	yue	tel	cdo
ces	cld	tha	nep
dan	heb	tir	uzn
ydd	isl	urd	ory
ell	lao	aka	pam
hat	lit	amh	pag
hun	mlt	bos	run
ita	kmr	ilo	sna
jpn	srp	nan	azb
kor	ben	nor	uig
mkd	hak	lvs	war
pol	hrv	fry	wol
por	guj	bam	wuu
ron	kan	mya	yor
rus	kok	ceb	ewe
sin	msa	bin	
slk	mal	lug	

Takeover to "user_canada"

chn	tau	nld	hat
zho	tus	vls	hun
hai	[tce,ttm]	pan	ita
hin	bel	ydd	ium
esi	nob	est	lav
fil	ces	fin	mkd
den	dan	ell	ars

pol	swe	isl	iku
por	tgl	gle	kok
gla	ukr	lao	dks
hbs	vie	lit	swa
slk	cym	mlt	ekk
slv	pes	cre	kea
spa	khm	hrv	
deu	heb	frs	

Discarded

dak	moh	see	cre
pdc	oji	als	msa

10.8 Imputation

Dataset	Number of data points before imputation	Number of data points after imputation	Number of languages
L1_plus	1266	3681	240
L1_enriched	816	3593	240
user	1095	4988	276

Table 10.2: This table shows the number of data points in the different Datasets before and after Imputation.

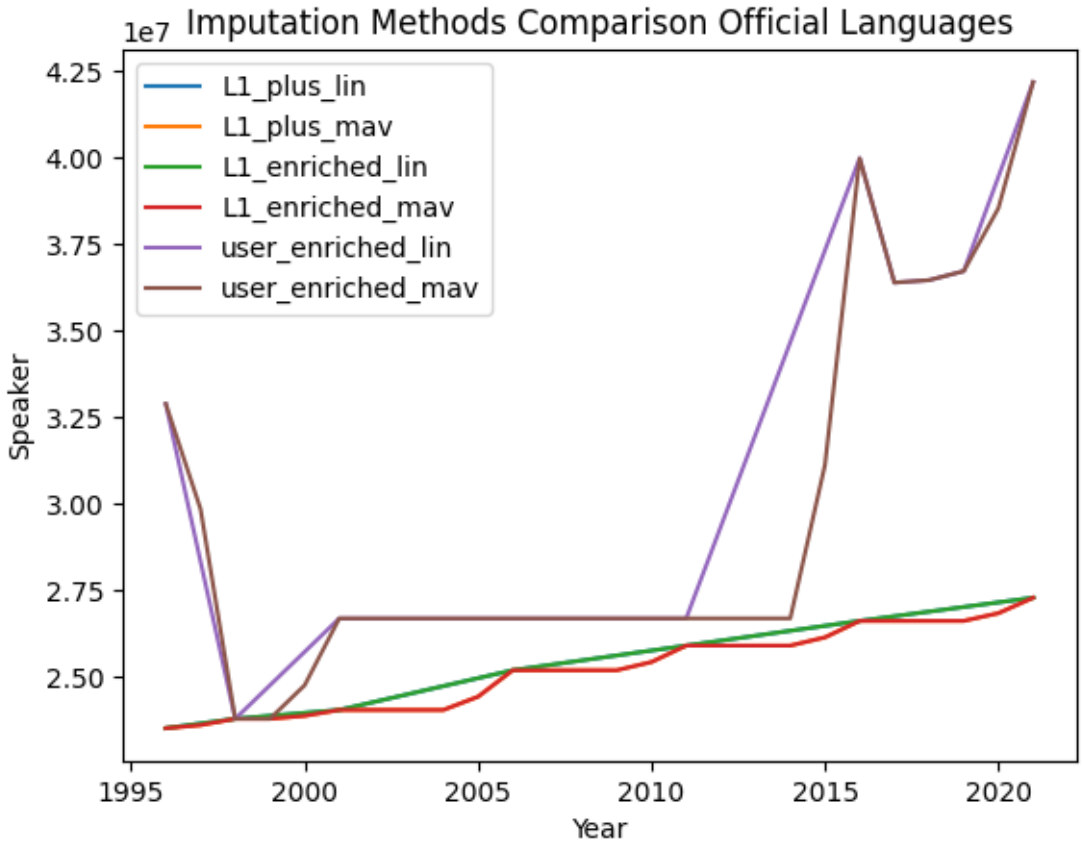

Figure 10.7: Imputation Methods Comparison Official Languages

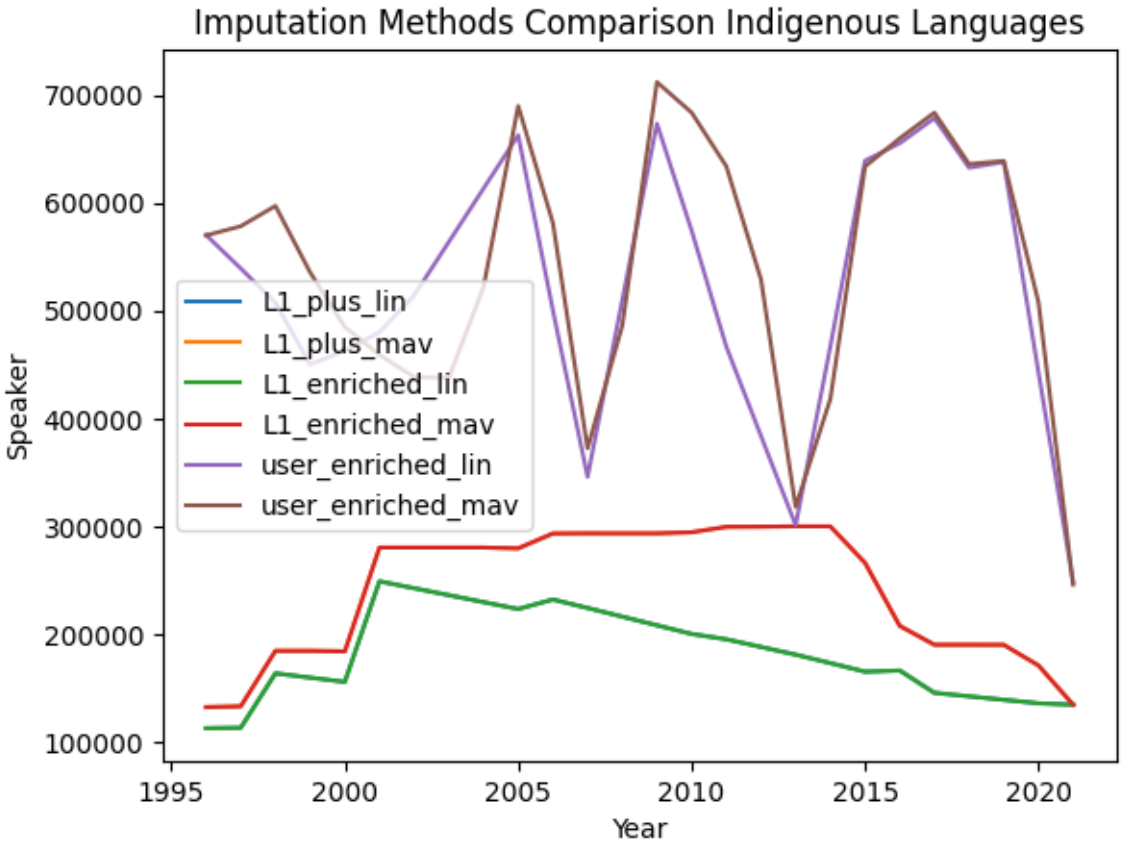

Figure 10.8: Imputation Methods Comparison Indigenous Languages. The dataset of L1_plus and L1_enriched show the same number of languages and therefore the same data, which is imputed by the methods.

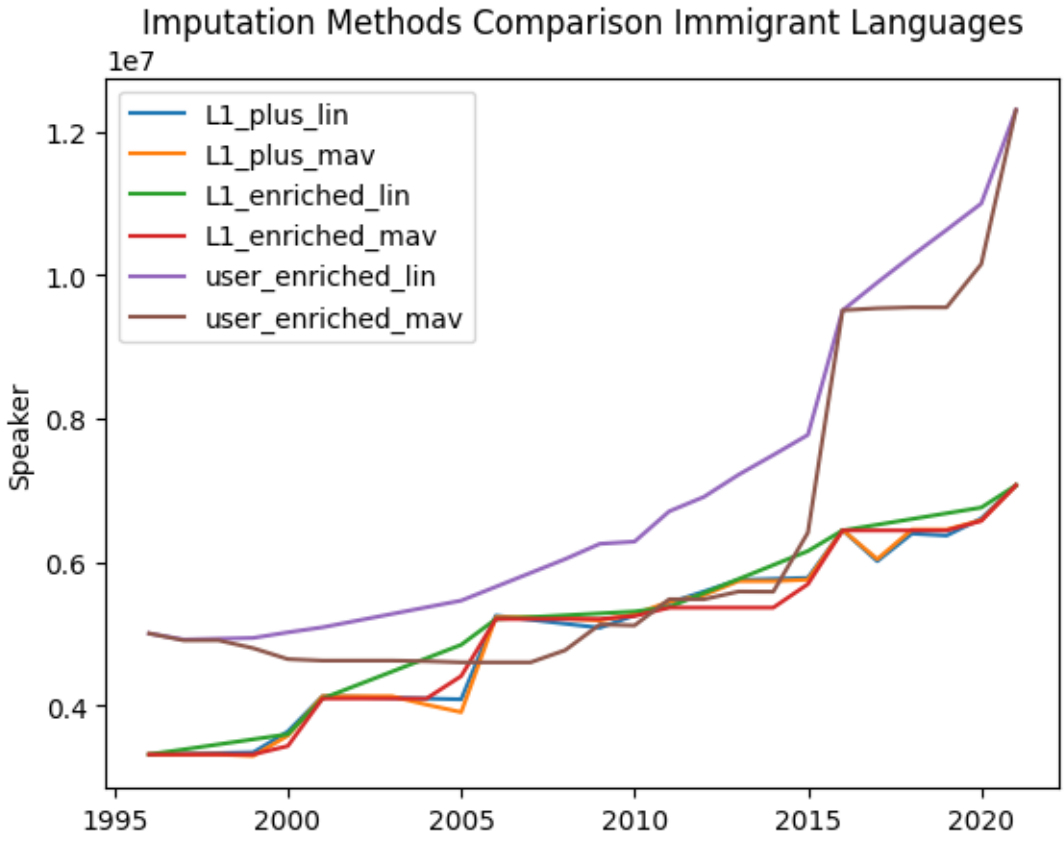

Figure 10.9: Imputation Methods Comparison Immigrant Languages

10.9 Linguistic Richness

	All languages	Indigenous languages	Official languages	Immigrant languages	Local languages
1996	59	10	2	47	12
1997	59	10	2	47	12
1998	70	20	2	48	22
1999	70	20	2	48	22
2000	76	20	2	54	22
2001	79	21	2	56	23
2002	79	21	2	56	23
2003	79	21	2	56	23
2004	79	21	2	56	23
2005	79	21	2	56	23
2006	112	27	2	83	29
2007	114	29	2	83	31
2008	113	28	2	83	30
2009	113	28	2	83	30
2010	113	28	2	83	30
2011	115	29	2	84	31
2012	115	29	2	84	31
2013	116	29	2	85	31
2014	119	32	2	85	34
2015	118	31	2	85	33
2016	168	55	2	111	57
2017	166	53	2	111	55
2018	166	53	2	111	55
2019	166	53	2	111	55
2020	166	53	2	111	55
2021	205	54	2	149	56

Table 10.3: This table shows the linguistic richness based on the L1 plus dataset (linear and moving average imputation).

10.10 Linguistic Evenness

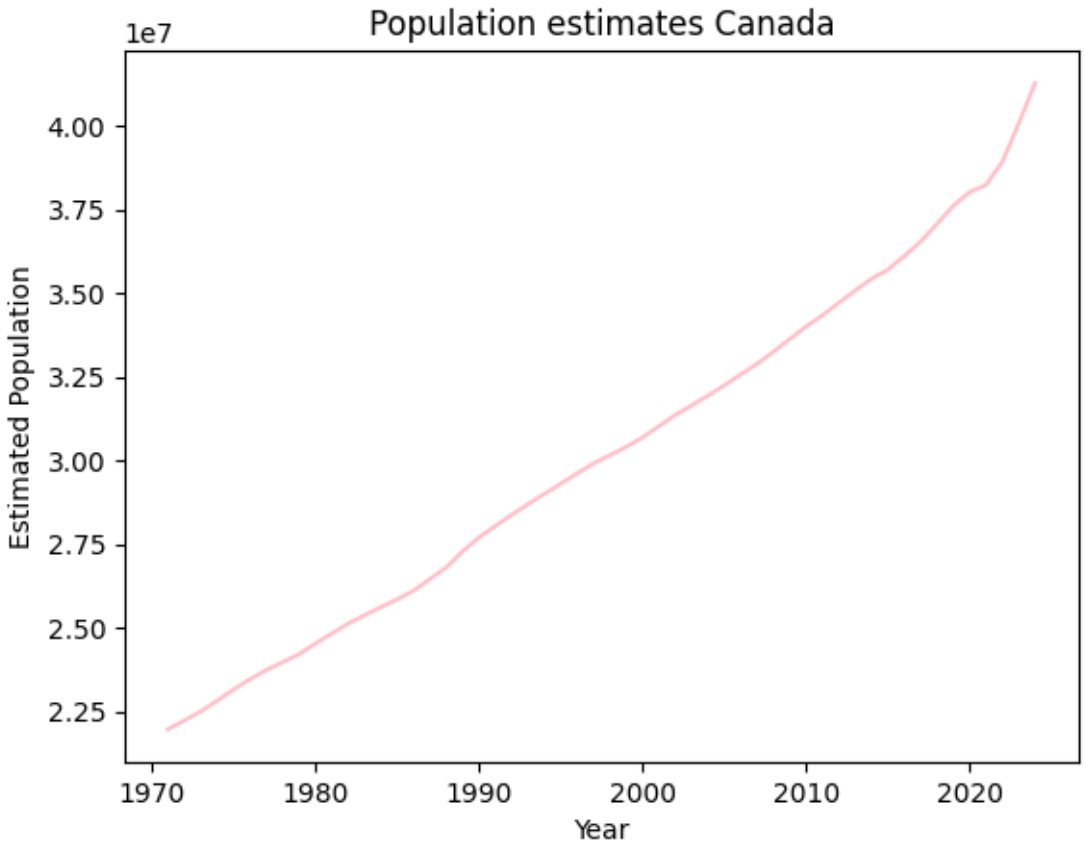

Figure 10.10: This plot shows the population estimates for Canada from 1971 to 2025. The data were published by Statistics Canada (2025b).

	All languages	Indigenous languages	Official languages	Immigrant languages	Local languages
1996	0.910462	0.003803	0.794566	0.112093	0.798369
1997	0.906340	0.003780	0.791269	0.111291	0.795049
1998	0.905360	0.005422	0.789251	0.110687	0.794673
1999	0.901004	0.005248	0.785665	0.110091	0.790913
2000	0.904773	0.005071	0.781159	0.118543	0.786230
2001	0.916621	0.008023	0.775467	0.133131	0.783490
2002	0.913507	0.007732	0.774415	0.131360	0.782147
2003	0.912043	0.007459	0.774732	0.129852	0.782191
2004	0.910233	0.007189	0.774726	0.128318	0.781915
2005	0.908243	0.006922	0.774538	0.126783	0.781459
2006	0.942209	0.007123	0.773770	0.161316	0.780893
2007	0.935491	0.006816	0.770621	0.158053	0.777437
2008	0.927756	0.006500	0.766593	0.154663	0.773093
2009	0.919521	0.006188	0.762098	0.151235	0.768286
2010	0.918389	0.005883	0.757859	0.154647	0.763743
2011	0.918594	0.005681	0.754646	0.158266	0.760327
2012	0.917084	0.005415	0.750623	0.161045	0.756038
2013	0.915962	0.005156	0.746826	0.163980	0.751982
2014	0.911048	0.004883	0.743412	0.162753	0.748295
2015	0.908325	0.004623	0.741779	0.161923	0.746401
2016	0.920586	0.004601	0.737385	0.178600	0.741985
2017	0.900849	0.003979	0.732283	0.164587	0.736262
2018	0.901914	0.003836	0.725471	0.172607	0.729307
2019	0.891569	0.003695	0.718500	0.169375	0.722195
2020	0.891630	0.003570	0.714269	0.173791	0.717839
2021	0.902264	0.003515	0.713822	0.184927	0.717337

Table 10.4: Linguistic Evenness: Coverage of Population by Language Groups

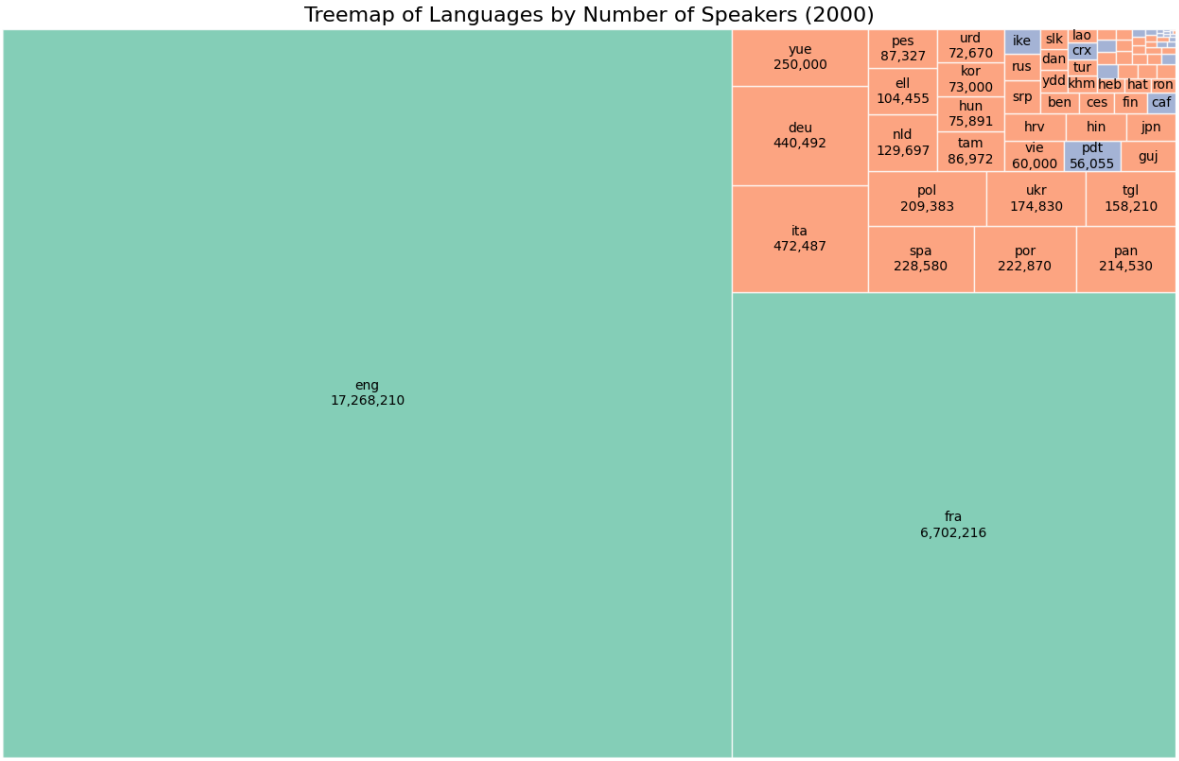

Figure 10.11: This tree map shows the size of languages based on the number of L1 speakers in 2000. Around 10 % of the Canadian population is not represented in this plot. Green are the official languages, orange are the immigrant languages and blue are the Indigenous languages. The ISO codes are indicated for all languages with more than 10000 speakers. The top 20 languages are accompanied by information on their number of speakers.

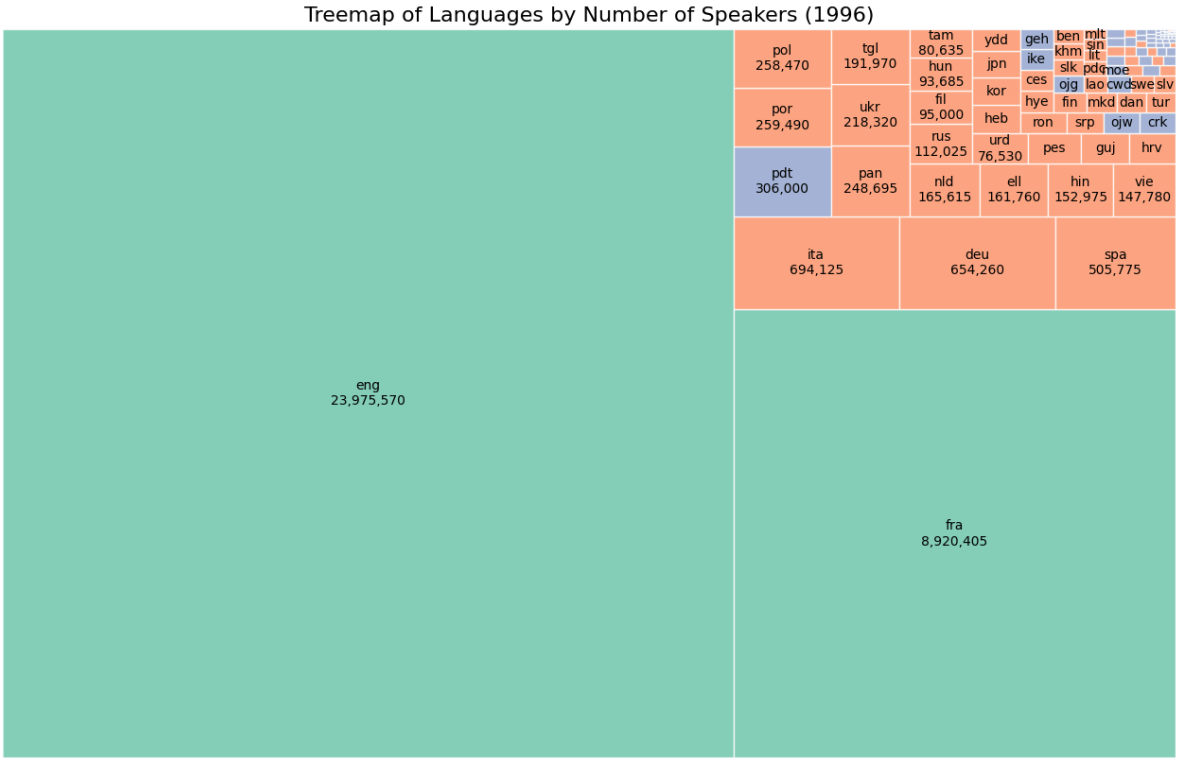

Figure 10.12: This tree map shows the size of languages based on the number of speakers (user) in 2000. Around 10 % of the Canadian population is not represented in this plot. Green are the official languages, orange are the immigrant languages and blue are the Indigenous languages. The ISO codes are indicated for all languages with more than 9000 speakers. The top 20 languages are accompanied by information on their number of speakers.

10.11 Index of Linguistic Diversity

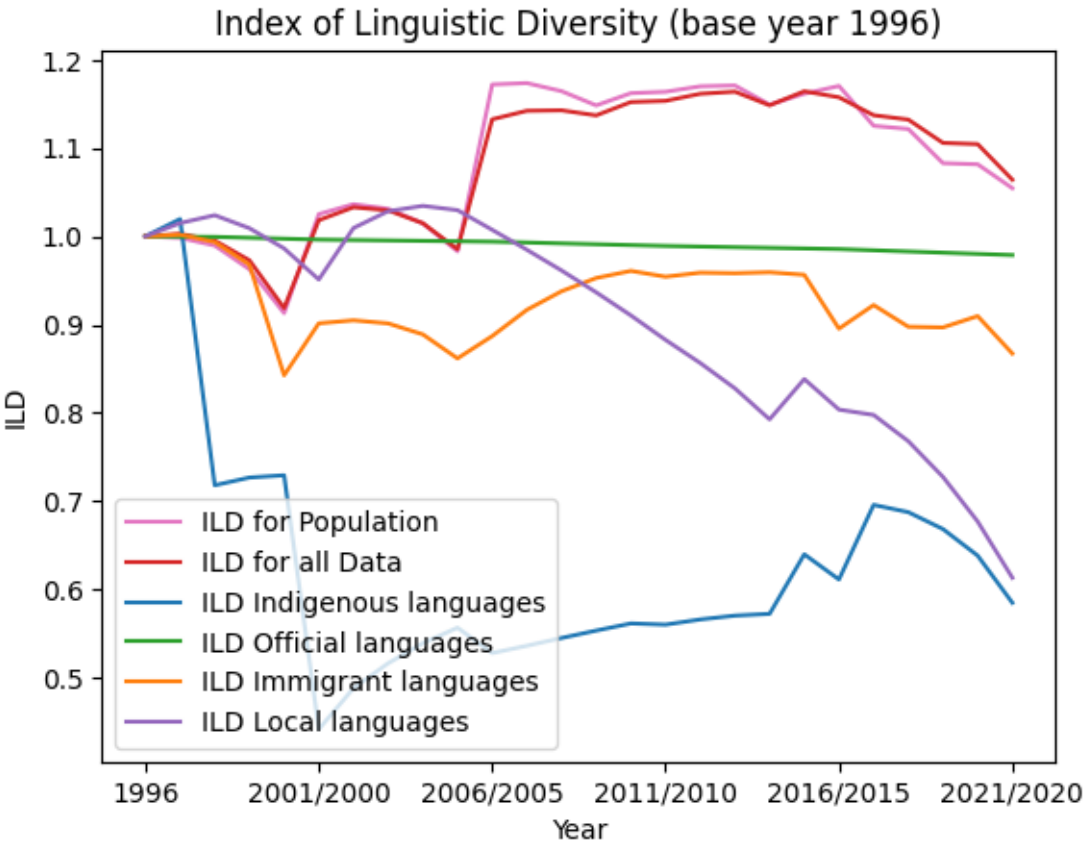

Figure 10.13: This plot shows the Index of Linguistic Diversity with the baseline set on 1996. The ILD that is computed on all data that is normalized with the population estimates is indicated by the pink line. The red line represents the ILD of all data that is normalized by the sum of this data. The other groups are also normalized by the sum of their group and it shows therefore the internal ILD of the groups.

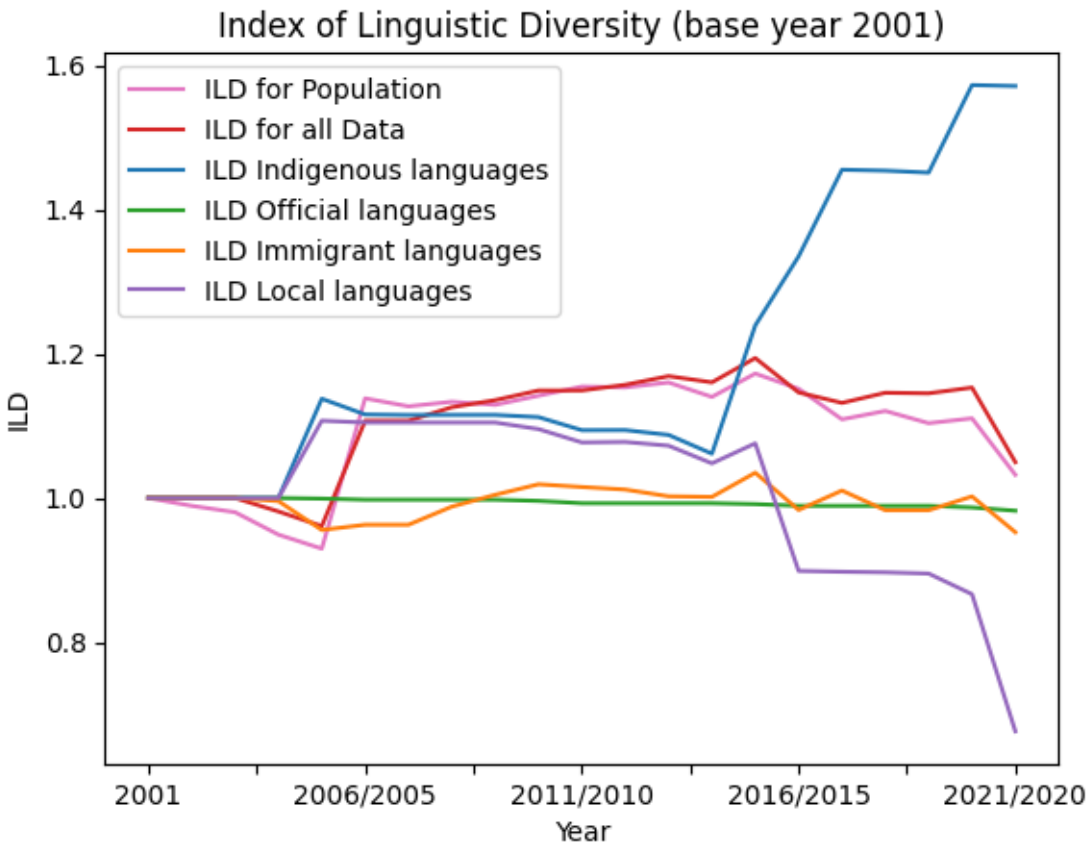

Figure 10.14: This plot shows the Index of Linguistic Diversity with the baseline set on 2001. The ILD that is computed on all data that is normalized with the population estimates is indicated by the pink line. The red line represents the ILD of all data that is normalized by the sum of this data. The other groups are also normalized by the sum of their group and it shows therefore the internal ILD of the groups. The plot shows the ILD based on moving average imputation.

10.12 Status

Status

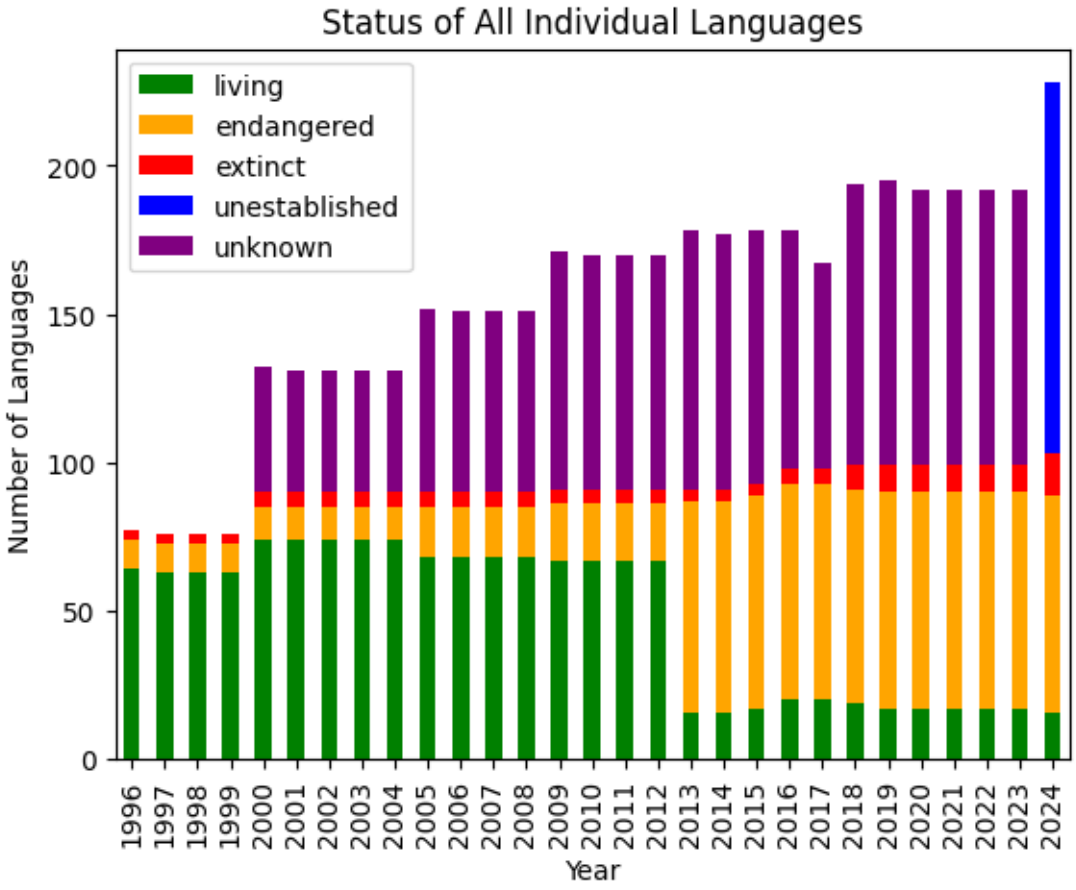

Figure 10.15: This stacked bar plot shows the status of all individual languages in Canada.

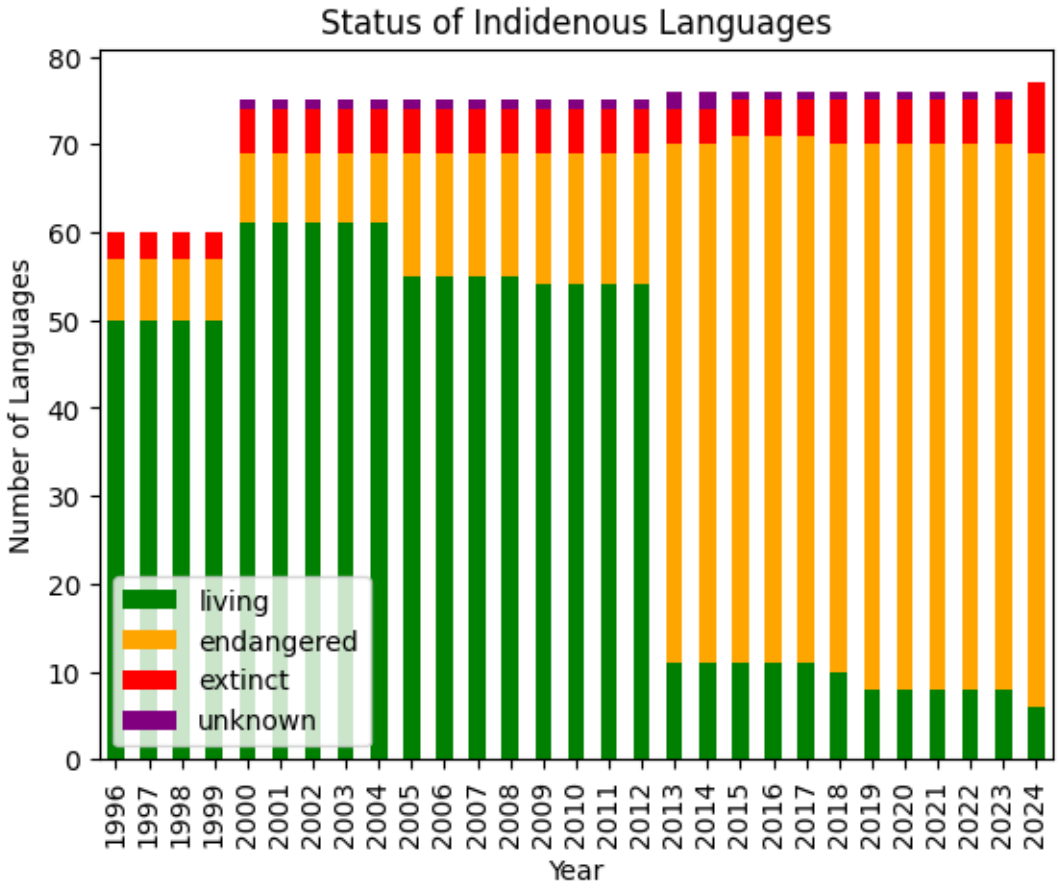

Figure 10.16: This stacked bar plot shows the status of Indigenous languages in Canada.

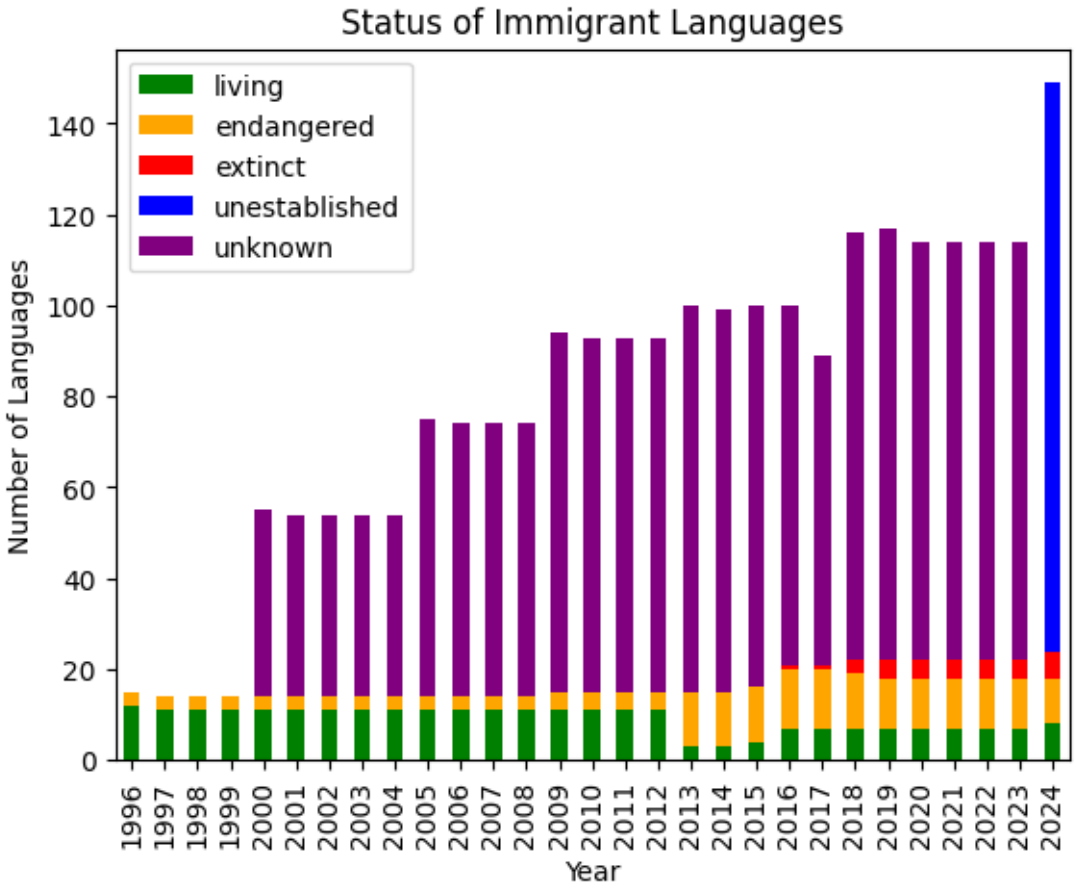

Figure 10.17: This stacked bar plot shows the status of immigrant languages in Canada.

	All languages	Indigenous languages	Immigrant languages
Number of changes	85	62	23
Number of positive status changes	78	60	18
Number of negative status changes	7	2	5
Net changes	135	112	23
Total change size	153	120	33

Table 10.5: This table shows the status changes. The number of changes, how many of these changes are positive (getting a higher status level, meaning the language gets less stable) or negative (lower level means lower threat level). The net change and the total change size (absolute sum of changes). This excludes changes, that are related to appearance and disappearance in the dataset

EGIDS

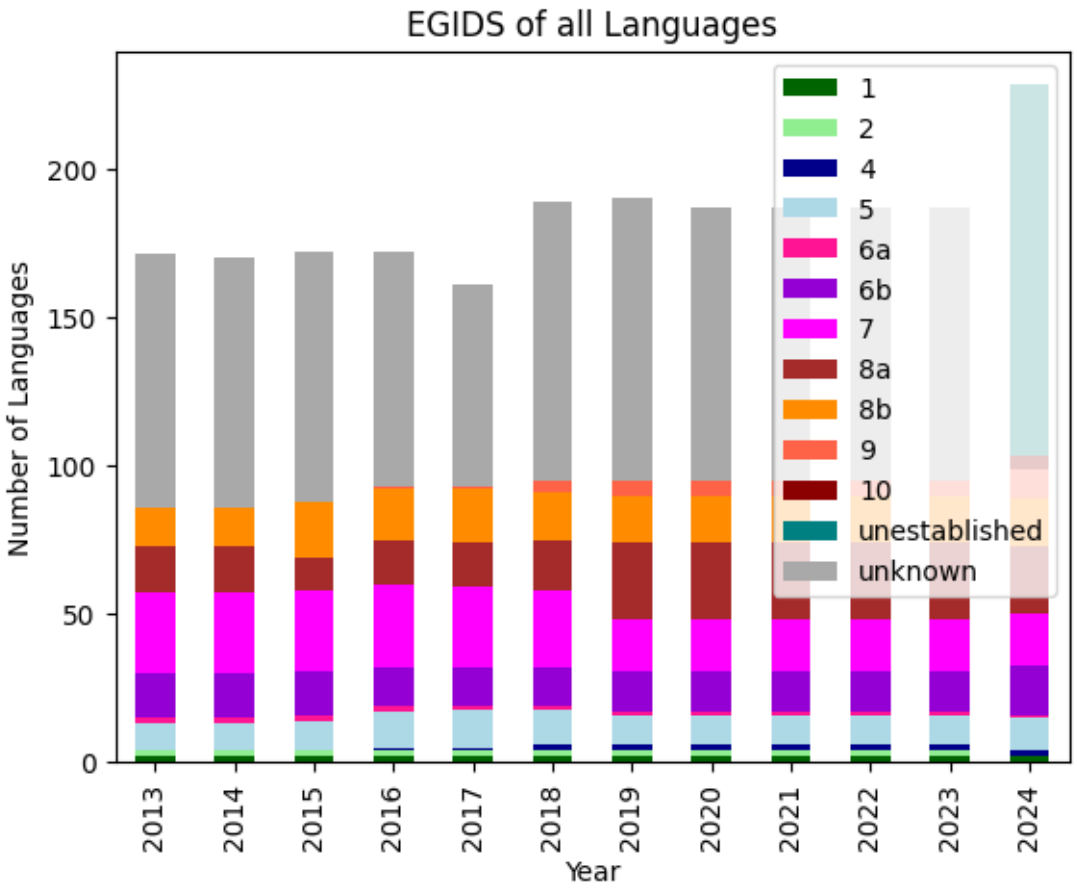

Figure 10.18: This stacked bar plot shows the EGIDS level of all individual languages in Canada.

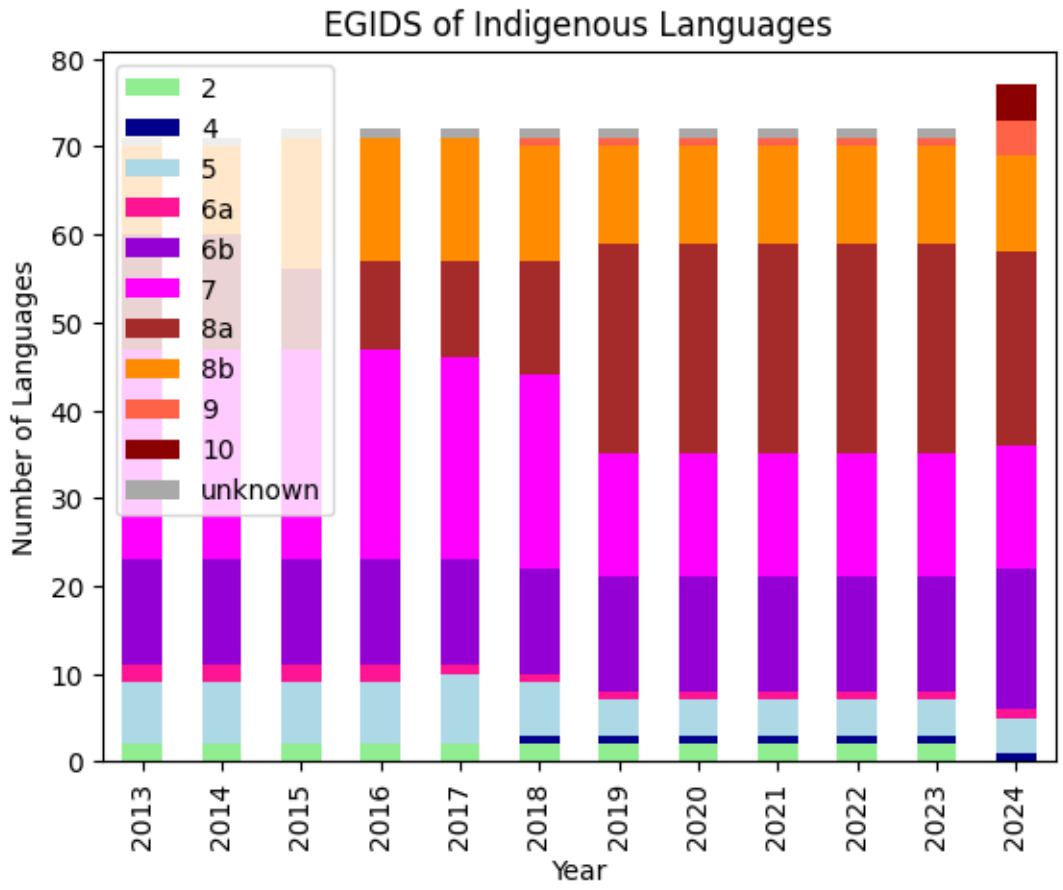

Figure 10.19: This stacked bar plot shows the EGIDS level of Indigenous languages in Canada.

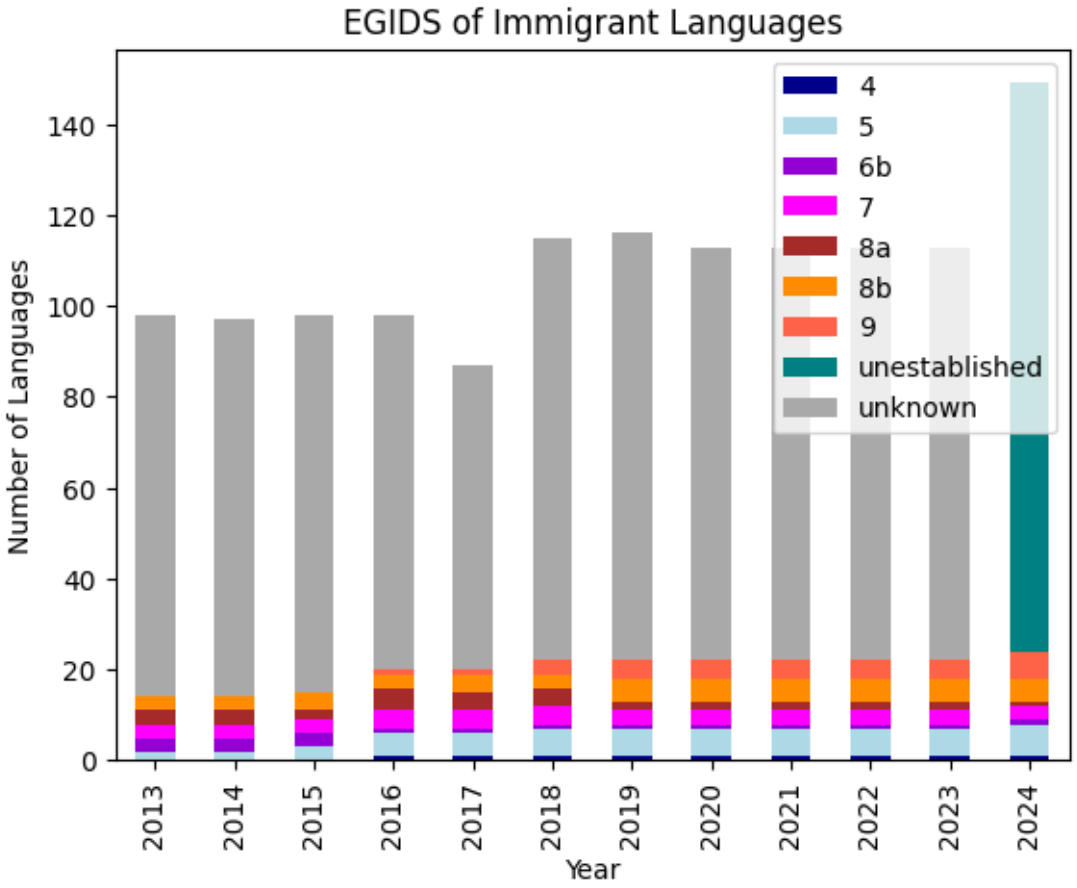

Figure 10.20: This stacked bar plot shows the EGIDS level of immigrant languages in Canada.

	All languages	Indigenous languages	Immigrant languages
Number of changes	45	35	10
Number of positive status changes	35	25	10
Number of negative status changes	10	10	0
Net changes	45	30	15
Total change size	67	52	15

Table 10.6: This table shows the egids level changes. The number of changes, how many of these changes are positive (getting a higher status level, meaning the language gets less stable) or negative (lower level means lower threat level). The net change and the total change size (absolute sum of changes). This excludes changes, that are related to appearance and disappearance in the dataset

10.13 Red List Index

EGIDS Level	Weight
’0’	0
’1’	1
’2’	2
’3’	3
’4’	4
’5’	5
’6a’	6
’6b’	7
’7’	8
’8a’	9
’8b’	10
’9’	11
’10’	12

Table 10.7: This table shows the weights needed for the RLI assigned to the EGIDS (EGIDS only version). All languages that are unestablished and or the EGIDS is unknown are not included in the computation.

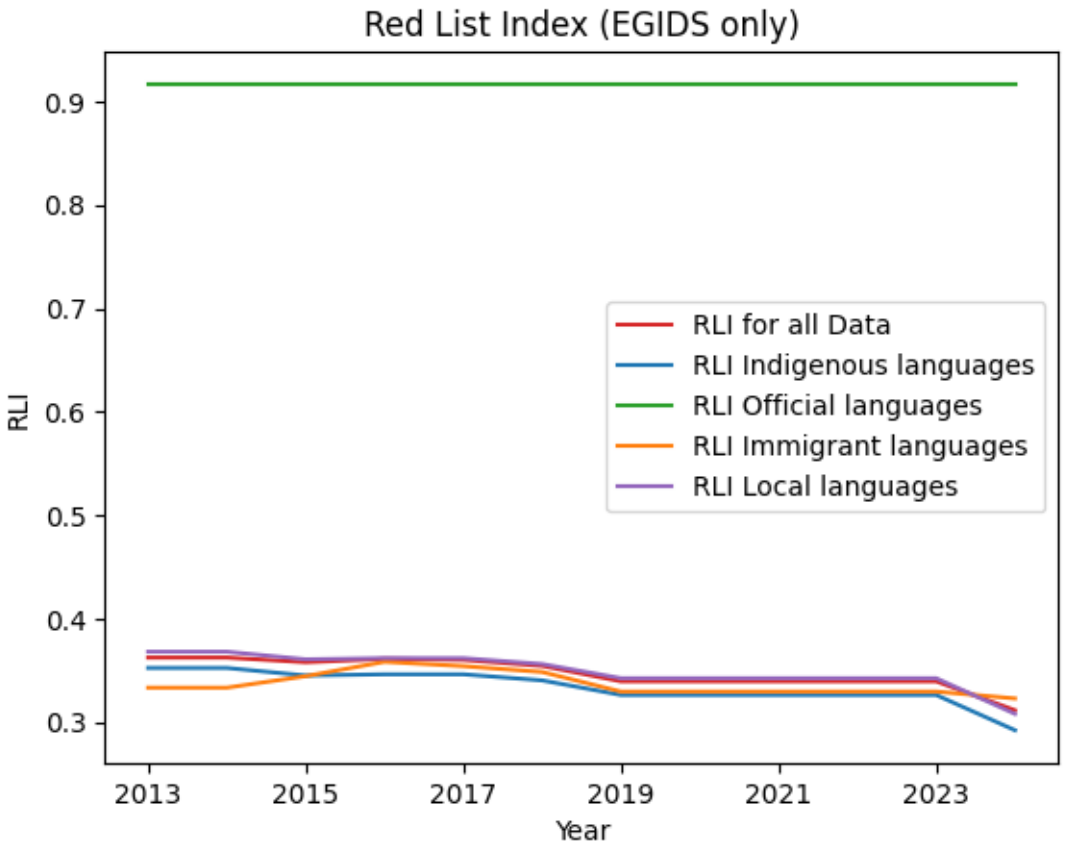

Figure 10.21: This plot shows the Red List Index for Canada between 1996 and 2021 based on the EGIDS. The weights for computing this RLI can be find in table 10.7.

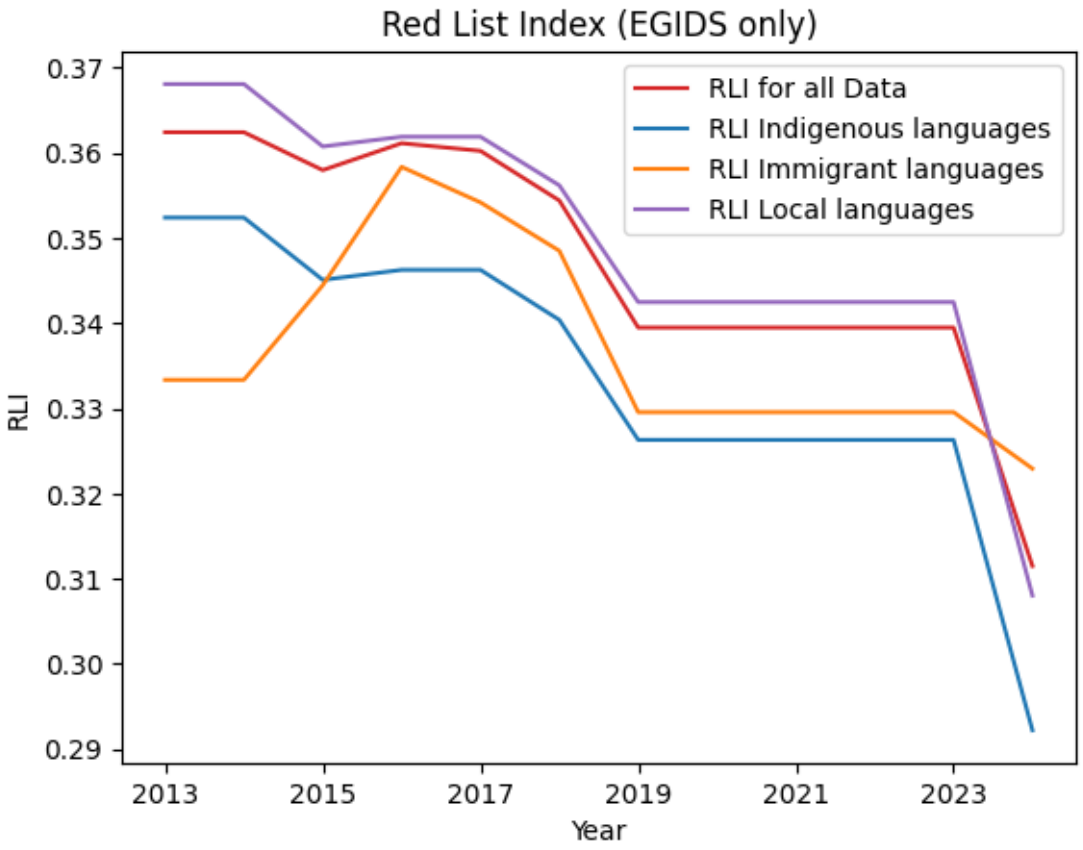

Figure 10.22: This plot shows a zoom of the Red List Index for Canada between 1996 and 2021 based on the EGIDS. The weights for computing this RLI can be find in table 10.7. The RLI of the official languages is not represented in this plot.

List of Figures

List of Tables

References

Aarfi, S. A., & Ahmed, N. (2024). Prompt engineering for generative ai: practical techniques and applications. *Scientific Engineering*, *24*, 102–115.

Arsenault Morin, A., & Geloso, V. (2020). Multilingualism and the decline of french in quebec. *Journal of multilingual and multicultural development*, *41*(5), 420–431.

Axelsen, J. B., & Manrubia, S. (2014). River density and landscape roughness are universal determinants of linguistic diversity. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1784), 20133029.

Baumann, A., Lenzner, B., Fellner, H. A., & Essl, F. (2024). The relation between european colonialism and linguistic diversity. In J. Nölle et al. (Eds.), *The evolution of language: Proceedings of the 15th international conference (evolang xv).* Retrieved from `https://evolang2024.github.io/proceedings/paper.html?nr=55` doi: 10.17617/2.3587960

Belshaw, J. D. (2016). *Canadian history: Post-confederation*. BCcampus.

Bickford, J. A., Lewis, M. P., & Simons, G. F. (2015). Rating the vitality of sign languages. *Journal of Multilingual and Multicultural Development*, *36*(5), 513–527.

Bishop, P. (2015). *Pier 21.* Historica Canada. Retrieved from `https://thecanadianencyclopedia.ca/en/article/pier-21` (Last accessed 04 September 2025)

Blu, T., Thevenaz, P., & Unser, M. (2004). Linear interpolation revitalized. *IEEE transactions on image processing*, *13*(5), 710–719.

Boissonneault, M., & Hammarström, H. (2025). Evaluating the validity of census data for tracking speaker numbers: an investigation of canada's indigenous languages. *Current Issues in Language Planning*, 1–23.

Boissonneault, M., Tallman, A., Gast, V., & Greenhill, S. J. (2025). Projected speaker numbers and dormancy risks of canada's indigenous languages. *Royal Society Open Science*, *12*(2), 241091.

Bouckaert, R., Redding, D., Sheehan, O., Kyritsis, T., Gray, R., Jones, K. E., & Atkinson, Q. (2022). Global language diversification is linked to socio-ecology and threat status. *SocArXiv*.

Brinklow, N. T., Littell, P., Lothian, D., Pine, A., & Souter, H. (2019). Indigenous language technologies & language reclamation in canada. In *Collection of research papers of the 1st international conference on language technologies for all* (pp. 402–406).

Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., … Hua, X. (2022). Global predictors of language endangerment and the future of linguistic diversity. *Nature ecology & evolution*, *6*(2), 163–173.

Butchart, S. H. M., Resit Akçakaya, H., Chanson, J., Baillie, J. E., Collen, B., Quader, S., … Hilton-Taylor, C. (2007). Improvements to the red list index. *PloS one*, *2*(1), e140.

Butchart, S. H. M., Stattersfield, A. J., Bennun, L. A., Shutes, S. M., Akçakaya, H. R., Baillie, J. E. M., … Mace, G. M. (2004, 10). Measuring global trends in the status of biodiversity: Red list indices for birds. *PLOS Biology*, *2*(12), null. Retrieved from `https://doi.org/10.1371/journal.pbio.0020383` doi: 10.1371/journal.pbio.0020383

Cambridge University Press. (2025). *diversity.* Retrieved from `https://dictionary.cambridge.org/dictionary/english/diversity` (Last accessed 09 september 2025)

Canadian Geographic. (2018a). *First nations.* Canadian Geographic. Retrieved from `https://indigenouspeoplesatlasofcanada.ca/section/first-nations/` (Last accessed 04 September 2025)

Canadian Geographic. (2018b). *Inuit.* Canadian Geographic. Retrieved from `https://indigenouspeoplesatlasofcanada.ca/section/inuit/` (Last accessed 04 September 2025)

Canadian Museum of Immigration at Pier 21. (2025). *Timeline.* (Visited on Mai 30, 2025)

Castonguay, C. (2019). Quebec's new language dynamic: French fading fast. *Language Problems and Language Planning*, *43*(2), 113–134.

Catalogue of endangered languages. (2024). University of Hawaii at Mānoa. Retrieved from `http://www.endangeredlanguages.com` (Last accessed 07 October 2025)

Collin, R. O. (2010). Ethnologue. *Ethnopolitics*, *9*(3-4), 425–432.

Currie, T. E., & Mace, R. (2009). Political complexity predicts the spread of ethnolinguistic groups. *Proceedings of the National Academy of Sciences*, *106*(18), 7339–7344.

Dirks, G. E. (2024). *Immigration policy in canada.* Historica Canada. Retrieved from `https://thecanadianencyclopedia.ca/en/article/immigration-policy` (Last accessed 03 September 2025)

Drude, S. (2018). Reflections on diversity linguistics: Language inventories and atlases.

Duchêne, A., & Humbert, P. N. (2018). Surveying languages: the art of governing speakers with

numbers. *International Journal of the Sociology of Language*, *2018*(252), 1–20.

Eaton, J., & Turin, M. (2022). Heritage languages and language as heritage: the language of heritage in canada and beyond. *International Journal of Heritage Studies*, *28*(7), 787–802.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2019). *Ethnologue: Languages of the world* (22nd ed.). SIL International.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2024). *Ethnologue: Languages of the world* (27th ed.). SIL International.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2025a). *Ethnologue: Languages of the world* (28th ed.). Dallas, Texas: SIL International. Retrieved from `www.ethnologue.com`

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2025b). *Ethnologue: Languages of the world - languagecodes.tab* (28th ed.). Dallas, Texas: SIL International. Retrieved from `www.ethnologue.com/codes/LanguageCodes.tab`

Edmonston, B., & Poulin, J. (2025). Statistics canada. In *The canadian encyclopedia.* Historica Canada. Retrieved from `https://thecanadianencyclopedia.ca/en/article/statistics-canada`

Essfors, H. N. H. (2025). *Global linguistic diversity: incorporating similarity measures using large-scale cross linguistic databases* (Master's thesis, Universität Wien). doi: 10.25365/thesis.78808

Fedor, P., & Zvaríková, M. (2019). Biodiversity indices. In B. Fath (Ed.), *Encyclopedia of ecology (second edition)* (Second Edition ed., p. 337-346). Oxford: Elsevier. Retrieved from `https://www.sciencedirect.com/science/article/pii/B9780124095489105585` doi: https://doi.org/10.1016/B978-0-12-409548-9.10558-5

Filice, M. (2016). *Numbered treaties.* Historica Canada. Retrieved from `https://thecanadianencyclopedia.ca/en/article/numbered-treaties` (Last accessed 02 September 2025)

Fincher, C. L., & Thornhill, R. (2008). A parasite-driven wedge: Infectious diseases may explain language and other biodiversity. *Oikos*, *117*(9), 1289–1297.

First Peoples' Cutural Council. (2010). *Report on the status of b.c. first nations languages 2010.* Retrieved from `https://fpcc.ca/wp-content/uploads/2020/07/2010-report-on-the-status-of-bc-first-n` (Last accessed 04 September 2025)

First Peoples' Cutural Council. (2014). *Report on the status of b.c. first nations languages 2014* (2nd ed.). Retrieved from `https://fpcc.ca/wp-content/uploads/2020/07/FPCC-LanguageReport-141016-WEB.pdf`

(Last accessed 04 September 2025)

First Peoples' Cutural Council. (2018). *Report on the status of b.c. first nations languages* (3rd ed.). Retrieved from https://fpcc.ca/wp-content/uploads/2020/07/FPCC-LanguageReport-180716-WEB.p (Last accessed 04 September 2025)

First Peoples' Cutural Council. (2022). *Report on the status of b.c. first nations languages* (4th ed.). Retrieved from https://fpcc.ca/wp-content/uploads/2023/02/FPCC-LanguageReport-23.02.14-FI (Last accessed 04 September 2025)

First peoples' cultural council. (2025). Retrieved from https://fpcc.ca/ (Last accessed 04 September 2025)

Fishman, J. A. (1991). *Reversing language shift*. Bristol, Blue Ridge Summit: Multilingual Matters. Retrieved 2025-09-11, from https://doi.org/10.21832/9781800418097 doi: doi:10.21832/9781800418097

Foot, R. (2020). *Canadian charter of rights and freedoms.* Historica Canada. Retrieved from https://www.thecanadianencyclopedia.ca/en/article/canadian-charter-of-righ (Last accessed 07 September 2025)

Gallant, D. J. (2022). *Indigenous languages in canada.* Historica Canada. Retrieved from https://thecanadianencyclopedia.ca/en/article/aboriginal-people-languages (Last accessed 04 September 2025)

Gavin, M. C., Botero, C. A., Bowern, C., Colwell, R. K., Dunn, M., Dunn, R. R., … others (2013). Toward a mechanistic understanding of linguistic diversity. *BioScience*, *63*(7), 524–535.

Gavin, M. C., & Stepp, J. R. (2014). Rapoport's rule revisited: geographical distributions of human languages. *PloS one*, *9*(9), e107623.

Genolini, C., Ecochard, R., & Jacqmin-Gadda, H. (2013, 01). Copy mean: A new method to impute intermittent missing values in longitudinal studies. *Open Journal of Statistics*, *03*, 26-40. doi: 10.4236/ojs.2013.34A004

Gordon, J., Raymond G. (Ed.). (2005). *Ethnologue: Languages of the world* (15th ed.). SIL International.

Greenberg, J. H. (1956). The measurement of linguistic diversity. *Language*, *32*(1), 109–115.

Grimes, B. F. (Ed.). (1996). *Ethnologue: Languages of the world* (13th ed.). Summer Institute of Linguistics.

Grimes, B. F. (Ed.). (2000). *Ethnologue: Languages of the world* (14th ed.). SIL International.

Grin, F., & Fürst, G. (2022). Measuring linguistic diversity: A multi-level metric. *Social indicators research*, *164*(2), 601–621.

Gullifer, J. W., & Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and cognition*, *23*(2), 283–294.

Hammarström, H. (2015). Ethnologue 16/17/18th editions: A comprehensive review. *Language*, *91*(3), 723–737.

Hammarström, H. (2016). Linguistic diversity and language evolution. *Journal of Language Evolution*, *1*(1), 19–29.

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2025). *Glottolog 5.2.* Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://doi.org/10.5281/zenodo.15525265` (Last accessed 07 October 2025)

Harmon, D., & Loh, J. (2010). The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages. *Language documentation and conservation*, *4*, 97–151.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020, September). Array programming with NumPy. *Nature*, *585*(7825), 357–362. Retrieved from `https://doi.org/10.1038/s41586-020-2649-2` doi: 10.1038/s41586-020-2649-2

Henderson, W. B., & Bell, C. (2019). *Rights of indigenous peoples in canada.* Historica Canada. Retrieved from `https://thecanadianencyclopedia.ca/en/article/aboriginal-rights` (Last accessed 03 September 2025)

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*(2), 427–432. Retrieved 2025-08-27, from `http://www.jstor.org/stable/1934352`

Hua, X., Greenhill, S. J., Cardillo, M., Schneemann, H., & Bromham, L. (2019). The ecological drivers of variation in global language diversity. *Nature communications*, *10*(1), 2047.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

Hutson, J., Ellsworth, P., & Ellsworth, M. (2024). Preserving linguistic diversity in the digital age: a scalable model for cultural heritage continuity. *Journal of Contemporary Language Research*, *3*(1).

Hyndman, R. (2010, 01). Moving averages. In (p. 866-869). Springer. doi: 10.1007/978-3-642-04898-2$_3$80

Irwin, R. (2022). *Reserves in canada.* Historica Canada. Retrieved from `https://www.thecanadianencyclopedia.ca/en/article/aboriginal-reserves`

(Last accessed 02 September 2025)

IUCN. (2012). *Iucn red list categories and criteria: Version 3.1* (2nd ed.). Gland, Switzerland and Cambridge, UK: IUCN.

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*(2), 363-375. Retrieved from https://nsojournals-onlinelibrary-wiley-com.uaccess.univie.ac.at/doi/abs/1(
doi: https://doi-org.uaccess.univie.ac.at/10.1111/j.2006.0030-1299.14714.x

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, *88*(10), 2427-2439. Retrieved from https://esajournals-onlinelibrary-wiley-com.uaccess.univie.ac.at/doi/abs/1(
doi: https://doi-org.uaccess.univie.ac.at/10.1890/06-1736.1

Kalbach, W., James-Abra, E., & Edmonston, B. (2021). Canadian census. In *The canadian encyclopedia.* Historica Canada. Retrieved from https://thecanadianencyclopedia.ca/en/article/demographic-data-collection

Kandler, A., & Unger, R. (2023). Modeling language shift. In *Diffusive spreading in nature, technology and society* (pp. 365–387). Springer.

Khawaja, M. (2021). Consequences and remedies of indigenous language loss in canada. *Societies*, *11*(3), 89.

Kosten, I. L. (1967). Lineare interpolation. In *Handbuch der mathematik* (Reprint 2015 ed.). Germany: Walter de Gruyter GmbH.

Laserson, U. (2024, July). *squarify.* Retrieved from https://github.com/laserson/squarify

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (16th ed.). SIL International.

Lewis, M. P., & Simons, G. F. (2010). Assessing endangerment: Expanding fishman's gids. *Revue roumaine de linguistique*, *55*(2), 103–120.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (Eds.). (2013). *Ethnologue: Languages of the world* (17th ed.). SIL International.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (Eds.). (2015). *Ethnologue: Languages of the world* (18th ed.). SIL International.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (Eds.). (2016). *Ethnologue: Languages of the world* (19th ed.). SIL International.

Lieberson, S. (1964). An extension of greenberg's linguistic diversity measures. *Language*, *40*(4), 526–531.

Loh, J., & Harmon, D. (2014). *Biocultural diversity: Threatened species, endangered languages.* WWF Netherlands Zeist, The Netherlands.

Mace, R., & Pagel, M. (1995). A latitudinal gradient in the density of human languages in north america. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *261*(1360), 117–121.

Mahanty, A. (2022, March). *waybackpy.* Retrieved from `https://github.com/akamhy/waybackpy` doi: 10.5281/ZENODO.3977276

McCardle, B. (2021). *Enfranchisement.* Historica Canada. Retrieved from `https://thecanadianencyclopedia.ca/en/article/enfranchisement` (Last accessed 02 September 2025)

McGhee, R. (2025). *History of early indigenous peoples in canada.* Historica Canada. Retrieved from `https://thecanadianencyclopedia.ca/en/article/prehistory` (Last accessed 02 September 2025)

McIvor, O. (2018). Indigenous languages in canada: What you need to know.

Meijering, E. (2002). A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*, *90*(3), 319–342.

Miller, J. (2024). *Residential schools in canada.* Historica Canada. Retrieved from `https://thecanadianencyclopedia.ca/en/article/residential-schools` (Last accessed 02 September 2025)

Minister of Citizenship and Immigration Canada. (2021). *Discover canada: The rights and responsibilities of citizenship.* Retrieved from `https://www.canada.ca/content/dam/ircc/migration/ircc/english/pdf/pub/discover.pdf` (Last accessed 29 August 2025)

Mougeon, R. (2015). *French language in canada.* Historica Canada. Retrieved from `https://thecanadianencyclopedia.ca/en/article/french-language` (Last accessed 07 September 2025)

Nagy, N. (2021). Heritage languages in canada. *The Cambridge handbook of heritage languages and linguistics*, 178–204.

Nestor, R. (2018). *Pass system in canada.* Historica Canada. Retrieved from `https://thecanadianencyclopedia.ca/en/article/pass-system-in-canada` (Last accessed 02 September 2025)

Nettle, D. (1999). *Linguistic diversity* (1. publ. ed.). Oxford: Oxford Univ. Press.

Norris, M. J. (2007). Aboriginal languages in canada: Emerging trends and perspectives on second language acquisition. *Canadian Social Trends*, *83*(20), 19–27.

Office Of The Commissioner Of Official Languages, & Burnaby, B. J. (2020). *Language policy in canada.* Historica Canada. Retrieved from `https://www.thecanadianencyclopedia.ca/en/article/language-policy`

(Last accessed 04 September 2025)

OpenAI. (2025, September). *openai.* Retrieved from https://github.com/openai/openai-python

Oxford English Dictionary. (2025). *Diversity, n., 5.* Retrieved from https://doi.org/10.1093/OED/1528061826 (Last accessed 09 september 2025)

Pallets. (2025, March). *Jinja2.* Retrieved from https://github.com/pallets/jinja/

pandas development team, T. (2025, July). *pandas-dev/pandas: Pandas.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.15831829 doi: 10.5281/zenodo.15831829

Paolillo, J. C., & Das, A. (2006). Evaluating language statistics: The ethnologue and beyond. *Contract report for UNESCO Institute for Statistics*.

Paquet, R. G., & Levasseur, C. (2019). When bilingualism isn't enough: perspectives of new speakers of french on multilingualism in montreal. *Journal of multilingual and multicultural development*, *40*(5), 375–391.

Parrott, Z. (2023). *Indigenous peoples in canada.* Historica Canada. Retrieved from https://thecanadianencyclopedia.ca/en/article/aboriginal-people (Last accessed 02 September 2025)

Pendakur, R. (1990). Speaking in tongues: Heritage language maintenance and transfer in canada.

Phoenix, J., & Taylor, M. (2024). *Prompt engineering for generative ai*. " O'Reilly Media, Inc.".

Plotly Technologies Inc. (2015). *Collaborative data science.* Montreal, QC: Author Retrieved from https://plot.ly

Prévost, J.-G. (2020). Past, present and future of canadian statistics. *Statistical Journal of the IAOS*, *36*(2), 355–363.

Reitz, K. (2025, August). *requests.* Retrieved from https://requests.readthedocs.io/en/latest/

Rice, K. (2023). *Indigenous language revitalization in canada.* Historica Canada. Retrieved from https://thecanadianencyclopedia.ca/en/article/indigenous-language-revitali (Last accessed 04 September 2025)

Richardson, L. (2007). Beautiful soup documentation. *April*.

Saint-Jacques, B., Chambers, J., & Cooper, C. (2019). *Immigrant languages in canada.* Historica Canada. Retrieved from https://thecanadianencyclopedia.ca/en/article/ethnic-languages (Last accessed 04 September 2025)

Sallabank, J. (2024). Linguistic diversity. *Global Perspectives*, *5*(1), 117328.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., … others (2024). The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Sil. (2025). Retrieved from `https://www.sil.org/`

Simons, G. F. (2019). Two centuries of spreading language loss. *Proceedings of the Linguistic Society of America*, *4*, 27–1.

Simons, G. F., & Fennig, C. D. (Eds.). (2017). *Ethnologue: Languages of the world* (20th ed.). SIL International.

Simons, G. F., & Fennig, C. D. (Eds.). (2018). *Ethnologue: Languages of the world* (21st ed.). SIL International.

Simpson, E. H. (1949). Measurement of diversity. *nature*, *163*(4148), 688–688.

Speer, E. R. (2024, November). *langcodes.* Retrieved from `https://github.com/georgkrause/langcodes`

Statistics Canada. (1997). *1996 census handbook.* Retrieved from `publications.gc.ca/pub?id=9.564472sl=0` (Last accessed 13 August 2025)

Statistics Canada. (1998a). *Population able to speak various non-official languages (73), showing age groups (13a) and sex (3), for canada, provinces, territories, census divisions and census subdivisions, 1996 census (20% sample data).* Retrieved from `https://www12.statcan.gc.ca/English/census96/data/tables/Rp-eng.cfm?LANG=EAPATH=3D` (Last accessed 13 August 2025)

Statistics Canada. (1998b). *Population by knowledge of official languages (5), showing age groups (13a) and sex (3), for canada, provinces, territories, census divisions and census subdivisions, 1996 census (20% sample data).* Retrieved from `https://www12.statcan.gc.ca/English/census96/data/tables/Rp-eng.cfm?LANG=EAPATH=3D` (Last accessed 13 August 2025)

Statistics Canada. (1998c). *Population by mother tongue (77), showing age groups (13a) and sex (3), for canada, provinces, territories, census divisions and census subdivisions, 1996 census (20% sample data).* Retrieved from `https://www12.statcan.gc.ca/English/census96/data/tables/Rp-eng.cfm?LANG=EAPATH=3D` (Last accessed 13 August 2025)

Statistics Canada. (1999). *1996 census dictionary.* Retrieved from `publications.gc.ca/pub?id=9.564473sl=0` (Last accessed 13 August 2025)

Statistics Canada. (2003). *Detailed mother tongue (80), age groups (13)*

and sex (3) for population, for canada, provinces, territories and forward sortation areas, 2001 census - 20% sample data ©. Retrieved from https://www12.statcan.gc.ca/English/census01/products/standard/themes/Rp-en (Last accessed 13 August 2025)

Statistics Canada. (2004a). *2001 census dictionary.* Retrieved from publications.gc.ca/pub?id=9.557970sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2004b). *Languages : 2001 census technical report.* Retrieved from publications.gc.ca/pub?id=9.557975sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2007). *Detailed mother tongue (148), single and multiple language responses (3) and sex (3) for the population of canada, provinces, territories, census metropolitan areas and census agglomerations, 2006 census - 20% sample data.* Retrieved from https://www12.statcan.gc.ca/census-recensement/2006/dp-pd/tbt/Rp-eng.cfm?LA (Last accessed 13 August 2025)

Statistics Canada. (2010a). *2006 census dictionary.* Retrieved from publications.gc.ca/pub?id=9.692922sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2010b). *Census 2006 - 2b (long form).* Retrieved from https://www12.statcan.gc.ca/census-recensement/2006/ref/about-apropos/versi (Last accessed 13 August 2025)

Statistics Canada. (2012a). *Census dictionary, census year, 2011.* Retrieved from publications.gc.ca/pub?id=9.695208sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2012b). *Census profile. 2011 census.* Retrieved from http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/index.cfm?Lan (Last accessed 13 August 2025)

Statistics Canada. (2012c). *Languages reference guide : 2011 census.* Retrieved from publications.gc.ca/pub?id=9.696794sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2017a). *Census profile. 2016 census.* Retrieved from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?La (Last accessed 13 August 2025)

Statistics Canada. (2017b). *Languages reference guide : Census of population, 2016.* Retrieved from publications.gc.ca/pub?id=9.818283sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2018a). *Dictionary, census of population, 2016.* Retrieved from publications.gc.ca/pub?id=9.826574sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2018b). *The evolution of language populations in canada, by mother tongue, from 1901 to 2016.* Retrieved from publications.gc.ca/pub?id=9.851551sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2018c). *Standing on the shoulders of giants : history of statistics canada, 1970 to 2008.* Retrieved from publications.gc.ca/pub?id=9.862841sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2023a). *Census profile. 2021 census of population.* Retrieved from https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E (Last accessed 13 August 2025)

Statistics Canada. (2023b). *Dictionary, census of population, 2021.* Retrieved from publications.gc.ca/pub?id=9.928275sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2023c). *Languages reference guide : Census of population, 2021.* Retrieved from publications.gc.ca/pub?id=9.908897sl=0 (Last accessed 13 August 2025)

Statistics Canada. (2025a). *Language diversity in canada.* Retrieved from https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2025007-eng.htm (Last accessed 15 september 2025)

Statistics Canada. (2025b). *Population estimates on july 1, by age and gender.* Retrieved from https://doi.org/10.25318/1710000501-eng (Last accessed 25 August 2025)

Sutherland, W. J. (2003). Parallel extinction risk and global distribution of languages and species. *Nature (London)*, *423*(6937), 276–279.

The Canadian Encyclopedia. (2022). *Indian act.* Historica Canada. Retrieved from https://thecanadianencyclopedia.ca/en/article/indian-act (Last accessed 02 September 2025)

The Canadian Encyclopedia. (2025a). *Timeline: Languages policy.* Historica Canada. Retrieved from https://thecanadianencyclopedia.ca/en/timeline/languages-policy (Last accessed 07 September 2025)

The Canadian Encyclopedia. (2025b). *Timeline: Significant events in canadian history.* Historica Canada. Retrieved from https://thecanadianencyclopedia.ca/en/timeline/100-great-events-in-canadian-histor (Last accessed 02 September 2025)

Theune, C. (2024, June). *pycountry.* Retrieved from https://github.com/pycountry/pycountry

ThoughtWorks. (2025, August). *selenium.* Retrieved from https://selenium-python-api-docs.readthedocs.io/en/latest/

Troper, H. (2024). *Immigration to canada.* Historica Canada. Retrieved from https://thecanadianencyclopedia.ca/en/article/immigration (Last accessed 03 September 2025)

Truth and Reconciliation Commission of Canada. (2015). *Honouring the truth, reconciling for the future : summary of the final report of the truth and reconciliation commission of canada*. Truth and Reconciliation Commission of Canada. Retrieved from https://nctr.ca/wp-content/uploads/2021/01/Executive$_S$ummary$_E$nglish$_W$eb.pdf (Last accessed 02 September 2025)

UNESCO. (2024). *Languages, cultures, knowledge: Unesco's action for indigenous peoples.* Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000392089 (Last accessed 04 September 2025)

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). *Language vitality and endangerment.* Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000183699 (Last accessed 11 September 2025)

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Walsh, M. (2005). Will indigenous languages survive? *Annu. Rev. Anthropol.*, *34*(1), 293–315.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. Retrieved from https://doi.org/10.21105/joss.03021 doi: 10.21105/joss.03021

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, *21*(2-3), 213–251.

Wittmann, G. (2019). Lineare funktionen. In *Elementare funktionen und ihre anwendungen* (pp. 43–66). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-662-58060-8$_3$ doi: 10.1007/978-3-662-58060-8$_3$

Zariquiey, R., Arakaki, M., Vera, J., Torres-Orihuela, G., Cuba-Raime, C., Barrientos, C., … Hammarström, H. (2022). Linking endangerment databases and descriptive linguistics: An assessment of the use of terms relating to language endangerment in grammars. *Language Documentation Conservation*.

Zeh, K. (2025). *Quantifying the relationship between digitization and linguistic diversity: a multilevel statistical analysis using r* (Master's thesis, Universität Wien). doi: 10.25365/thesis.78469