Investigating Multilinguality in the Plenary Sessions of the Parliament of Finland with Automatic Language Identification

Tommi Jauhiainen, Jussi Piitulainen, Erik Axelson, Ute Dieckmann, Mietta Lennes, Jyrki Niemi, Jack Rueter, Krister Lindén
Department of Digital Humanities, University of Helsinki

Abstract

In this paper, we use automatic language identification to investigate the usage of different languages in the plenary sessions of the Parliament of Finland. Finland has two national languages, Finnish and Swedish. The plenary sessions are published as transcriptions of speeches in Parliament, reflecting the language the speaker used. In addition to charting out language use, we demonstrate how language identification can be used to audit the quality of the dataset. On the one hand, we made slight improvements to our language identifier; on the other hand, we made a list of improvement suggestions for the next version of the dataset.

Keywords: language identification, multilinguality, plenary sessions

1. Introduction

In this paper, we use automatic language identification to investigate the usage of different languages in the plenary sessions of the Parliament of Finland. The plenary sessions are published as transcriptions of speeches given in Parliament, reflecting the language the speaker actually used. Finland has two national languages, Finnish and Swedish, as well as several minority languages, such as the Sami languages and the Finnish Romani.

Language identification can be used to bring forth many kinds of problems in the corpus processing pipeline for the dataset at hand. Instead of trying to circumvent all the problems by tweaking the language identifier, we record the issues that we can correct earlier in the pipeline.

In Section 2, we introduce some work on multilingual parliamentary proceedings. The details of the corpus we are focusing on in this paper are given in Section 3. Section 4 is a detailed description of our language identification process and how it can be used to improve the quality of both the corpus and the language identifier. In Section 5, we present the results of the language identification experiments, e.g., details on the languages used in Parliament. Section 6 is dedicated to investigating the sentences tagged as written in an undetermined language. In Section 7, we discuss the process and list our improvement suggestions.

2. Previous Work

There are several state bodies similar to the Parliament of Finland where the use of several languages is permitted. One of the most prominent bodies is the Canadian Parliament, where both English and French enjoy equal status and use (Hudon, 2022). Another source for multilingual parliamentary data is the Catalan Parliament, where discussions can include Spanish and Aranese Occitan interventions in addition to Catalan (Kulebi et al., 2022). The Belgian federal Parliament uses both Dutch and French, which were automatically identified on the paragraph level for the ParlaMint corpora (Erjavec et al., 2023).

The language use in the European Parliament is on a totally different level of multilingualism, currently with 24 official languages.[1]

As far as we are aware, this is the first study where fine-grained language identification is performed and the results analyzed on any of these corpora.

We have previously done similar experiments with the Newspaper and Periodical Corpus of the National Library of Finland (NLF) [2] and the Suomi24 Sentences Corpus 2001-2017 (suomi24-2001-2017)[3] (Jauhiainen et al., 2022b).

3. Corpus

The focal dataset of this paper is the Plenary Sessions of the Parliament of Finland, Downloadable Version 1.5 (The Parliament of Finland, 2017-01-01).

The dataset is available at the Language Bank of Finland (LBF).[4] The Language Bank is a comprehensive service suite for researchers utilizing linguistic resources. It hosts an extensive collection

[1] https://european-union.europa.eu/principles-countries-history/languages_en

[2] http://urn.fi/urn:nbn:fi:lb-2021092404

[3] suomi24-2001-2017-korp-v1-1, http://urn.fi/urn:nbn:fi:lb-2020021803

[4] https://www.kielipankki.fi/language-bank/

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 48–56
May 20, 2024.

of text and speech datasets, enabling diverse use. Users can explore and process these datasets using the Language Bank's online tools or download them to their personal computers.

The services of the Language Bank are overseen by the national FIN-CLARIN consortium, which consists of Finnish universities and research organizations.[5] FIN-CLARIN is part of the international CLARIN ERIC research infrastructure.[6] Researchers and research groups can arrange with FIN-CLARIN for the storage and distribution of their own research datasets.

3.1. Plenary Sessions of the Parliament of Finland

The proceedings of the plenary sessions of the Parliament of Finland are documented in minutes, which include information on the content of discussions, details of decisions made, and all speeches given. These minutes are prepared in both Finnish and Swedish. However, the speeches are recorded and published in the language in which they were originally delivered. The preparation of the minutes occurs in real-time during the session, and they are made available on the Parliament's website as soon as they are ready.[7]

3.2. Speech and Text Alignment

The Parliament of Finland's original written records have been synchronized with the audio from the video footage of the plenary meetings. The synchronization process involved aligning the spoken words of each individual speaker separately. This task was accomplished using automated tools developed by Aalto University.[8]

It's important to be aware that the synchronized transcripts might include inaccuracies, and unnecessary tags could have been added to the text as a result of the automated synchronization and voice recognition procedures. In instances where there was no corresponding text for the original audio in the transcripts, the speech was automatically transcribed, which could lead to unusual or incorrect entries.

3.3. eduskunta-v1.5-dl

The verticalized text (VRT) version of the Eduskunta corpus consists of one 1.9-gigabyte CWB-VRT (The IMS Open Corpus Workbench-VRT, (Evert and Team, 2022)) file comprising 28 million lines, organized into 1,009 text elements that mirror video files. These elements are further broken down into paragraphs (111,097), utterances (1,499,627, linked to specific video timestamps), and sentences, which are sequences of tokens. Each token is on its own line, together with the linguistic analysis of the sentence as token annotations. The sentence length distribution in Table 1 was computed with one of the vrt-tools developed in the Language Bank.

Total	Nobs	Mean
22,458,581	1,499,627	14.98

Min	D1	D2	LoQ	D3	D4	
1	4	6	7	9	11	
Med	**D6**	**D7**	**HiQ**	**D8**	**D9**	**Max**
13	15	18	20	22	28	406

Table 1: The distribution of sentence lengths measured in tokens, as segmented in eduskunta.vrt (v1.5). There are over 22 million tokens in 1.5 million sentences, with a mean sentence length of just below 15 tokens and a median of 13. The quantile points (deciles and the low and high quartile) are represented by the observed value at or above the point.

Table 2 shows the distribution of the videos over the time covered by the corpus.

videos	year
46	2008
120	2009
132	2010
124	2011
130	2012
134	2013
130	2014
116	2015
77	2016

Table 2: The 1,009 videos (by the attributes in the text element tags in the VRT file) counted by year.

The LBF has invested in the ability to annotate a single file format (CWB-VRT) with different tools, which is facilitated by adding *field names* to the otherwise purely positional token records. The names are declared in a comment at the beginning of the file, leaving token lines in the form of tab-separated values. The various VRT tools[9] can then refer to the input and output fields by name regardless of

[5] https://www.kielipankki.fi/organization/

[6] https://www.clarin.eu

[7] https://www.eduskunta.fi/FI/taysistunto/Sivut/Taysistuntojen-poytakirjat.aspx

[8] http://urn.fi/urn:nbn:fi:lb-2014091904

[9] https://github.com/CSCfi/Kielipankki-utilities/tree/master/vrt-tools

6,917,510	N
4,697,950	V
2,523,037	Adv
2,485,364	Pron
1,729,493	C
1,589,639	A
1,499,627	Punct
348,303	Num
315,561	Foreign
281,789	Adp
52,092	Symb
18,216	Interj

Table 3: The counts of the "parts of speech" of the tokens in the corpus file, as identified by the annotation pipeline.

their actual position on the line.

The sentences were annotated in the LBF with the old TurkuNLP Finnish dependency parser pipeline, adapted for the VRT file format.[10] The pipeline consists of two uses of a lexical transducer, OmorFI (Pirinen, 2015), first to look up all possible lemmatizations and some corresponding morpho-syntactic features for each word form in a sentence, disambiguated with a MarMot model (Mueller et al., 2013) trained by the Turku group as part of their pipeline. This is followed by another OmorFi lookup to fill in the features of the contextually selected reading of each token, and finally syntactic dependency analysis corresponding to the Turku Dependency Treebank (TDT) (Haverinen et al., 2014) with a trained model that uses MaTe tools (Björkelund et al., 2010).[11] The annotation model predates the Universal Dependencies effort (De Marneffe et al., 2021).

Further variants of the base forms were added afterward to enable certain features in the Korp platform, where the corpus is made available for the search of examples.[12]

The corpus was further annotated with FiNER (Ruokolainen et al., 2020) to annotate the tokens that were recognized to be parts of names (or some other expressions) by their classes (like person, organization, location).

Table 3 shows how many times the annotation pipeline classified a token as noun, verb, and so on. The number of "Foreign" words may or may not be an indication of the proportion of non-Finnish language in the corpus.

The sentence-per-line view of the VRT file used in the following experiments was extracted with a relatively straightforward VRT tool that, by default, lists the token forms of each sentence on the same line, separated by space characters.

4. Language Identification

Our language identification experiments were conducted on a sentence level using the HeLI-OTS language identifier (Jauhiainen et al., 2022a).[13] We are currently using this language identifier on our standard corpus creation pipeline (Jauhiainen et al., 2022c; Dieckmann et al., 2023). However, the level on which the language identification is sensible differs from one dataset to another. For example, the optical character recognition (OCR) quality of the Newspaper and Periodical Corpus of the National Library of Finland (NLF) [14] is in places so terrible that out of the box identification results for the sentences can be very exotic (Jauhiainen et al., 2022b).

For development purposes, we have an internal test set for HeLI-OTS. The test set contains more than 1.2 million lines of text written in one of the 200 languages HeLI-OTS has in its repertoire. Whenever we modify the software or its language models, we investigate the effects of these changes by considering the recall, precision, and F-scores before and after the change. We look at these scores on the overall average level for all languages as well as on the level of individual languages if needed. The test set is not an independent entity, and it has not been manually verified, so whenever we make changes that cause the error rates to increase for some languages, we may take a look at the misidentified sentences in order to check their validity and remove them from the test set. We may also add new text lines to the test set when developing the identifier system as part of a specific investigation similar to what is described in this paper.

As of the writing of this paper, the current published version of the identifier is HeLI-OTS 1.5.[15] On the internal test set, it attains a macro F1 score of 99.21% over the 200 languages and a micro F1 score of 99.62% over the c. 1.2 million lines.

4.1. Experiments with HeLI-OTS 1.5

At first glance, the quality of the sentences in the corpus at hand seems to be far superior to the one in the NLF corpus. However, sentence-level monolingual language identification still comes up with sentences in 129 different languages. Table 4

[10] https://github.com/TurkuNLP/Finnish-dep-parser

[11] https://www.ims.uni-stuttgart.de/en/research/resources/tools/matetools/

[12] https://www.kielipankki.fi/korp

[13] http://urn.fi/urn:nbn:fi:lb-2022011801

[14] http://urn.fi/urn:nbn:fi:lb-2021092404

[15] https://doi.org/10.5281/zenodo.10071264

lists the ten languages with the most identifications on the initial language identification run.

# sentences	ISO 639-3	Language
1,460,998	fin	Finnish
17,408	swe	Swedish
9,454	ido	Ido
1,840	hat	Haitian Creole
1,306	izh	Ingrian
1,044	ewe	Ewe
1,041	vot	Votic
756	und	Undetermined
512	kal	Greenlandic
468	olo	Livvi

Table 4: Top 10 identified languages with the number of sentences for each. HeLI-OTS 1.5 language identifier.

6,449 of the sentences identified as Ido were simply “Ed .”. This is an abbreviation for “Edustaja” meaning “representative”. Most of the rest of the sentences identified as Ido ended with “ ed .”. The problem here seems to be on the sentence tokenization level, as the sentences have been cut using the period after the abbreviation in a way it should not have been done.

The 576 sentences identified as Haitian Creole were “Värderade talman .”, meaning "Honored Speaker" in Swedish. Most of the other sentences identified as Haitian Creole included the word “talman” as well. “talman” seems to be a common word ending in the HeLI-OTS training corpus for Haitian Creole, whereas the word “talman” is so rare in the Swedish training corpus that the word has not made it to the word level language model for Swedish. Both training corpora are based on web crawls and originate from the Leipzig corpora collection (Goldhahn et al., 2012).[16] As this is a clear language identification error on a correctly tokenized sentence, we decided to switch to our development version of the HeLI-OTS language identifier featuring individual confidence thresholds for each language. We expected that sentences like “Värderade talman .” and other short sentences in Swedish identified as Haitian Creole would not have high confidence scores.

The unpublished version of the HeLI-OTS used in these experiments had been modified from the 1.5 version in the context of performing language identification on an excerpt of 10,000 Tweets from the Sydney area. In addition to the new confidence thresholds, the modifications included cleaning English material from the training corpora of other languages. On the internal test set, this version attained a macro F1 score of 99.59% and a micro F1 score of 99.66%.

4.2. Experiments with Confidence Thresholds

The development version came up with a slightly lower number of languages for the dataset: 112. The renewed top 10 language list can be seen in Table 5.

# sentences	ISO 639-3	Language
1,462,709	fin	Finnish
17,523	swe	Swedish
13,802	und	Undetermined
1,514	hat	Haitian Creole
711	izh	Ingrian
666	vot	Votic
331	ido	Ido
275	lud	Ludic
194	est	Estonian
99	pol	Polish

Table 5: Top 10 identified languages with the number of sentences for each. Development version of the HeLI-OTS language identifier.

The number of sentences in the undetermined category rose drastically due to the introduction of the confidence thresholds. Some of the language models in the development version have also been improved, so the number of sentences identified as Finnish and Swedish also rose slightly. Surprisingly, the number of sentences identified as Haitian Creole did not decrease as much as expected. “Värderade talman .”, “Ärade talman .”, and “Herr talman .” were still identified as Haitian Creole.

4.3. Increasing Swedish Vocabulary

Developing a general-purpose language identification system is always a compromise between the compactness of the system and the number of features retained for each language. At this point, the 10,000 most common features were retained in each feature category for Swedish.[17] The number of features retained is an individual setting for each language, currently spanning from 5,000 to 50,000 features. Based on these perfectly Swedish, and surely not Haitian Creole, sentences being misidentified, we increased the number of retained features to 30,000 for Swedish. The updated list of the top 10 languages and the number of sentences identified as each is shown in Table 6.

The number of sentences identified as Haitian Creole decreased so much that the language

[16] The corpora are “hat-ht_web_2015_30K” for Haitian Creole and the “swe_web_2002_1M” for Swedish.

[17] The feature categories in the off-the-shelf HeLI-OTS are words and character n-grams from one to six.

# sentences	ISO 639-3	Language
1,462,451	fin	Finnish
19,303	swe	Swedish
13,787	und	Undetermined
711	izh	Ingrian
666	vot	Votic
331	ido	Ido
275	lud	Ludic
200	est	Estonian
98	pol	Polish
79	eng	English

Table 6: Top 10 identified languages with the number of sentences for each on the third LI round.

dropped out of the top 10, with the number of sentences identified as Swedish increasing accordingly. On the internal test set, this version attains a macro F1 score of 99.21% and a micro F1 score of 99.64%. Increasing the Swedish vocabulary resulted in more of the sentences marked as Norwegian or Danish in the internal test set to be identified as Swedish. However, we considered it less of an error to confuse between these close Scandinavian languages than between Scandinavian languages and Haitian Creole. We should also be able to rectify this problem later by increasing the size of the Danish and Norwegian language models similarly.

The next language on the list is Ingrian, an underresourced Finnic language that is rather similar to Finnish. The most common sentences that had been identified as Ingrian were: “Otan esimerkin .”, “Minä kysyn .”, and “Pulliaiselle .”, in English “I take an example.”, “I ask.”, and “To Pulliainen.” These are perfectly all-right sentences in spoken Finnish, but the problem is that they could also be so in Ingrian.

4.4. Confusion between Ingrian Dialects and Ingrian

After closer examination of the sentences identified as Ingrian in the dataset as well as the Ingrian training corpus for HeLI-OTS, we came to the conclusion that, unfortunately, a long transcribed interview of a Finnish Ingrian dialect speaker had ended up in the Ingrian corpus. The Ingrian dialects[18] are considered Finnish, whereas Ingrian[19] itself is a separate language by the ISO 639-3 standard. After cleaning the Ingrian training corpus and recalculating its language models, we arrived at a list shown in Table 7.

At this point, we also audited the results on the internal test set and produced an updated set (already version 30 for the 200 languages). This was our last modification of the HeLI-OTS in the experiments described in this paper. On the new internal test set, the macro F1 over the 200 languages was 99.61%, and the micro F1 over the c. 1.2 million sentences was 99.68%.

# sentences	ISO 639-3	Language
1,462,016	fin	Finnish
19,304	swe	Swedish
14,883	und	Undetermined
668	vot	Votic
331	ido	Ido
274	lud	Ludic
203	est	Estonian
99	pol	Polish
79	eng	English
76	gsw	Swiss German

Table 7: Top 10 identified languages with the number of sentences for each on the fourth LI round.

Closer inspection of the sentences identified as Votic and Ludic revealed that most of them had the same “ ed .” abbreviation problem as the sentences identified as Ido.

4.5. Common Abbreviation Handling

As the “ ed .” abbreviation seemed to be responsible for the majority of remaining incorrect language identification, we decided to simulate the situation where the problem would have been corrected earlier in the pipeline. We rejoined the sentences where they had been cut off after the abbreviation. We also corrected an encoding issue, which was observed on c. 400-500 lines. The total number of sentences dropped from 1,499,627 to 1,474,286, which meant that 25,341 additional sentences had been created due to the abbreviation.

With this change, the number of different languages dropped from 112 to 111, and the top ten languages with the number of sentences can be seen in Table 8.

The number of sentences with undetermined language was more than halved, and the number of sentences identified as Votic, Ido, and Ludic was drastically reduced.

The corpus description[20] of the Korp version declares, “For portions where the original audio track did not have matching text in the transcript, the speech signal was recognized automatically using a Finnish language model, and such portions may contain strange or erroneous content.” This declaration is missing from the metadata of the

[18] https://en.wikipedia.org/wiki/Ingrian_dialects

[19] https://en.wikipedia.org/wiki/Ingrian_language

[20] http://urn.fi/urn:nbn:fi:lb-2019101621

# sentences	ISO 639-3	Language
1,446,338	fin	Finnish
19,286	swe	Swedish
6,626	und	Undetermined
187	est	Estonian
111	vot	Votic
78	eng	English
75	gsw	Swiss German
57	kal	Greenlandic
56	pol	Polish
54	roh	Romansh

Table 8: Top 10 identified languages with the number of sentences for each on the fifth LI round.

downloadable version.[21] It is an important piece of information as this behavior seems to be the cause of most of the “sentences” erroneously identified as Estonian. Unfortunately, this information is not currently provided on the utterance or word level.

4.6. Automatic Speech Recognition

Even though the metadata did not indicate whether the utterances were transcribed using Finnish automatic speech recognition (ASR), we observed that in all those cases we encountered, the sentences started with a lowercase letter. Identifying the origin language of the ASR-generated sentences is a very different task from general language identification and would require the use of other kinds of tools. As most of the observed identification errors in the top 10 languages seemed to originate from exotic utterances created by ASR, we decided to filter out all those sentences starting with a lowercase letter. This operation reduced the number of “sentences” by 6.4% and the number of tokens by 5.3%, indicating that the ASR-generated texts were shorter than average.

The total number of identified languages went down from 111 to 77. The updated list of sentences per language is shown in Table 9.

# sentences	ISO 639-3	Language
1,358,878	fin	Finnish
18,030	swe	Swedish
3,053	und	Undetermined
84	vot	Votic
43	est	Estonian
36	kal	Greenlandic
36	eng	English
35	roh	Romansh
24	lat	Latin
23	ido	Ido

Table 9: Top 10 identified languages with the number of sentences for each on the sixth LI round.

The final list of languages is missing the lone Northern Sami utterance we discovered during the described process. It came from Oras Tynkkynen[22] in 2013. In the middle of his speech, he re-saluted the speaker of the house in four languages. Finnish, Swedish, Northern Sami, and Russian: “Arvoisa puhemies. Ärade talman. arvvus adnon sagadoalli. Uvazhajemyi predsedatel.” The Russian version was transcribed using Latin characters and was thus not identified as Russian but as Slovakian. The Sami version was lost when we discarded all sentences beginning with a lowercase letter. We inspected the transcript on the Parliament site and found that for that sentence, it reads: “Árvvus adnon ságadoalli!”. It seems our corpus preparation process has dismissed the accents in this case. On this occasion, we noticed that all punctuations other than periods had also been either removed or transformed into periods.

5. Results

The actual languages attested in the dataset were very few: Finnish, Swedish, English, Latin, French, German, Spanish, Italian, and Northern Sami. Table 10 gives the number of “sentences” containing languages other than Finnish or Swedish observed in the dataset. Some sentences were well formed, but others were only single-word or partial sentences, as well as multilingual sentences containing Finnish and the indicated language.

# sentences	ISO 639-3	Language
34	eng	English
21	lat	Latin
2	fra	French
2	deu	German
2	spa	Spanish
1	ita	Italian

Table 10: The number of sentences in languages other than Finnish or Swedish that were actually observed and correctly identified in the dataset.

The longest English sentence we have found was uttered by Jacob Söderman[23] in 2011: “Whistle-blowing is the popular term used when someone who works in or for an organisation raises a concern about a possible fraud crime danger or other serious risk that could threaten customers colleagues shareholders the public or the organisations own reputation.”. He used this definition when explain-

[21] http://urn.fi/urn:nbn:fi:lb-2019101721

[22] https://www.eduskunta.fi/EN/kansanedustajat/Pages/846.aspx

[23] https://www.eduskunta.fi/FI/kansanedustajat/Sivut/311.aspx

ing the concept of whistleblowing to the Parliament in an otherwise Finnish speech.

The only real sentence in Latin we found was "Navigare necesse est.", which was said by Astrid Thors[24] in 2012. Her speech about the state of Finnish seafaring was mostly in Swedish but contained two longer passages in Finnish as well.

Our only French sentence comes from Timo Soini[25] in 2014: "Un pere une mere cest elementaire.". He was talking about participating in protests in France in the context of a discussion about same-sex marriage in Finland. The only lone (real) sentence identified as German comes from the same political party, the Finns: "Kein Geld fur Merkel nicht mehr.". It was uttered by Juha Väätäinen[26] in December 2011 in the context of European monetary policy. His previous sentence in the same speech is one of the two Spanish sentences we found in the corpus: "No mas dinero para Espana no mas euros para Italia.". A week later, he said the only more than one-word sentence identified as Italian: "Bravo bravissimo.".

6. Undetermined Languages

During the language identification experiments, we were especially focused on minimizing the number of false positives in languages that were not actually attested in the dataset. In the final list, shown in Table 9, we additionally had a little over three thousand sentences tagged as written in an undetermined language.

745 of these did not contain any alphabetical characters but consisted only of number characters or a single dot. Furthermore, c. 200 "sentences" consisted only of a personal name. These we consider to be correctly identified when tagged with an "und" label.

We collected the top 10 Finnish sentences left undetermined in Table 11. Most of these gain similar scores for other close Finnic languages as they do for Finnish. In cases like "Ei." e.g., "No", or "On.", e.g., "is", they could correctly be tagged with several languages such as Finnish and Estonian. A more advanced way of handling multi-lingual words would be to tag them with several language labels or with a label of the language group the languages belong to. A notable difference in the list is the last example, which contains the abbreviation "Ed." again favoring the identification as the Ido language. The "Ed." abbreviation followed by an inflected personal name is found in a further 250 sentences.

#	**Sentence**
273	Hyvät kollegat.
133	Hyvät edustajat.
88	Hyvät edustajakollegat.
62	Ei ole.
50	Ei.
28	Näin ei voi jatkua.
20	Kysynkin ministeri Risikolta.
17	On.
14	Hyvät ystävät.
12	Ed. Ukkolalle.

Table 11: The counts of the top 10 Finnish sentences tagged with undetermined language.

HeLI-OTS has the option to perform language set identification, which means that in the case of multilingual sentences, it can give several tags to the sentence. We have not yet experimented with this feature on this corpus, but there is a clear need for it as the next most common sentences left undetermined were multilingual Finnish-Swedish sentences. We give the top eight multilingual sentences with their counts in Table 12. The rest of the multilingual sentences did not occur more than once. Seven out of the eight repeating sentences are multilingual only due to the decision made by the transcriber. They could as well have been transcribed as two separate sentences, e.g., the first part of the most common multilingual sentence "Värderade herr talman" occurs 80 times as a lone sentence and the latter part "Arvoisa herra puhemies" 18,552 times. The word "Eli", e.g., "So", followed by the Latin phrase "summa summarum", could perhaps be considered a real code-switched sentence.

#	**Sentence**
36	**Värderade herr talman** arvoisa herra puhemies.
34	Arvoisa puhemies **herr talman**.
32	**Herr talman** arvoisa puhemies.
11	Arvoisa herra puhemies **värderade herr talman**.
8	Eli **summa summarum**.
4	**Fru talman** rouva puhemies.
2	**Värderade herr talman** ar voisa herra puhemies.
2	Hyvät edustajat **bästa riksdagsledamöter**.

Table 12: The counts of the top eight multilingual sentences tagged with undetermined language. The non-Finnish parts are indicated by boldface type.

The next notable group of sentences with undetermined languages consists of two to three-

[24] https://www.eduskunta.fi/FI/kansanedustajat/Sivut/770.aspx

[25] https://www.eduskunta.fi/FI/kansanedustajat/Sivut/767.aspx

[26] https://www.eduskunta.fi/FI/kansanedustajat/Sivut/1139.aspx

letter sentences containing the word “ministeri”, e.g., “minister”. For some reason, “ministeri” is a very common word in the Greenlandic training corpus, which resulted in a high number of short sentences being identified as Greenlandic, as can be seen in Table 4. Now, 231 of these sentences containing “Ministeri” or “ministeri” are tagged with an undetermined language.

Additionally, there were still a few long sentences containing Finnish and Swedish words that were clearly produced by the ASR that we had not managed to filter out. While perusing the three thousand undetermined sentences, we did not notice any written in languages not already mentioned.

7. Discussion and Conclusions

After the modifications during the described experiments, the general results on the internal test set of the development version of the HeLI-OTS remained at the same level. However, the identification accuracy on the dataset at hand was clearly improved.[27]

The following is a list of improvement ideas specific to the corpus at hand, which we noticed while inspecting the results of the language identification process. In addition to guiding us in preparing the next edition of the corpus, it functions as a general example of what kind of issues can be brought to light when fine-grained language identification is performed on this kind of corpora.

- Add “ed.” to the list of known abbreviations after which the sentence should not be cut. More generally, any domain-specific text corpus can contain a disproportionate number of abbreviations not attested in a more general text corpus for the same language.
- Some of the parliamentary sessions are very long. At least one observed session (on the 20th of December 2011) lasted for more than 12 hours. However, the metadata for that session in Korp says it is 10 hours longer. This might be a systematic error when the metadata is created.
- In some cases, the metadata indicated that the utterance happened later than the end of the session, even though the metadata reflected the correct duration for the session.
- The encoding for common Scandinavian characters “ä” and “ö” was messed up in some of the sentences (less than 500). For example, “käy myöskin” had changed to: “kÃ€y myÃ¶skin”.
- Add metadata indicating whether the utterances, sentences, and tokens were automatically generated by ASR during the text alignment process.
- Consider retaining manually transcribed accents and punctuation.
- Use language identifier with confidence thresholds.
- Add a Latinized version of Russian as one of the languages in order to detect further use of Russian.

The problems we encountered pertaining to the ASR-generated texts were similar in nature to the OCR problems we encountered with the NLF corpora, albeit less severe (Jauhiainen et al., 2022b). In both cases, language identification brought to light parts where OCR and ASR had been especially underperforming. With the Suomi24 corpus, we suggested leaving close Finnish-related languages out of the language repertoire when performing the language identification (Jauhiainen et al., 2022b). In this work, we were able to improve the quality of the Ingrian training corpus and use confidence thresholds to bring down the number of sentences that needed to be manually verified.

In this paper, we have demonstrated how a fine-grained language identification system can be used to find rare usage of foreign languages amongst a large number of sentences. We have also demonstrated how inspecting the language identification results with unexpected languages can bring forth problems in the corpus.

8. Acknowledgements

This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Research Council of Finland under grant number 358720 (FIN-CLARIAH – Developing a Common RI for CLARIAH Finland).

We thank the anonymous reviewers, especially for pointing out the need to examine the sentences tagged as written with an undetermined language.

9. Bibliographical References

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China. Coling 2010 Organizing Committee.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021.

[27] The newest version of HeLI-OTS including changes described in this article is available at `https://doi.org/10.5281/zenodo.10907468`.

Universal dependencies. *Computational linguistics*, 47(2):255–308.

Ute Dieckmann, Mietta Lennes, Jussi Piitulainen, Jyrki Niemi, Erik Axelson, Tommi Jauhiainen, and Krister Lindén. 2023. The pipeline for publishing resources in the Language Bank of Finland. In *CLARIN Annual Conference*, pages 33–43.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, et al. 2023. The ParlaMint corpora of parliamentary proceedings. *Language resources and evaluation*, 57(1):415–448.

Stephanie Evert and The CWB Development Team. 2022. *The IMS Open Corpus Workbench (CWB) Corpus Encoding and Management Manual*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48:493–531.

Marie-Ève Hudon. 2022. Official languages and parliament. *Ottawa, Canada: Library of Parliament.*

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022a. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3912–3922, Marseille, France. European Language Resources Association.

Tommi Jauhiainen, Jussi Piitulainen, Erik Axelson, and Krister Linden. 2022b. Language diversity in the newspaper and periodical corpus of the National Library of Finland. Digital Research Data and Human Sciences (DRDHum) ; Conference date: 01-12-2022 Through 03-12-2022.

Tommi Jauhiainen, Jussi Piitulainen, Erik Axelson, and Krister Lindén. 2022c. Language identification as part of the text corpus creation pipeline at the Language Bank of Finland. In *The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, pages 251–259, Uppsala, Sweden.

Baybars Kulebi, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2022. ParlamentParla: A speech corpus of Catalan parliamentary sessions. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 125–130, Marseille, France. European Language Resources Association.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Tommi A Pirinen. 2015. Omorfi — free and open source morphological lexical database for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 313–315, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2020. A Finnish news corpus for named entity recognition. *Language Resources and Evaluation*, 54:247–272.

10. Language Resource References

The Parliament of Finland. 2017-01-01. *Plenary Sessions of the Parliament of Finland, Downloadable Version 1.5*. Kielipankki.