ParlaMint in TEITOK

Maarten Janssen and Matyáš Kopp
Charles University, Faculty of Mathematics and Physics
Prague, Czechia
janssen,kopp@ufal.mff.cuni.cz

Abstract

This paper describes the ParlaMint 4.0 parliamentary corpora as made available in TEITOK at LINDAT. The TEITOK interface makes it possible to search through the corpus, to view each session in a readable manner, and to explore the names in the corpus. The interface does not present any new data, but provides an access point to the ParlaMint corpus that is less oriented to linguistic use only, and more accessible for the general public or researchers from other fields.

Keywords: ParlaMint, TEITOK, Document visualization

1. Introduction

ParlaMint (Erjavec et al., 2022) is "a CLARIN Flagship project which focuses on the creation of comparable and uniformly annotated corpora of parliamentary debates in Europe"[1]. The current ParlaMint 4.0 (Erjavec et al., 2023) release contains parliamentary sessions from 29 European countries and autonomous regions, with over a billion words in total. All the texts have been linguistically annotated, and adorned with bibliographical information about all speakers in all the documents. All information is encoded using the TEI/XML standard, and made publicly available via the CLARIN.SI repository[2], fully following the FAIR principles and making the data accessible for research purposes.

The data in ParlaMint are relevant for more than just linguistic research, and one could even argue that linguistic investigation is only a minor use case for the data in the repository. Yet despite being fully available in theory, making use of the ParlaMint data is not trivial. The sheer amount of data makes it difficult to get started. The linguistic annotation in the source code makes it hard to grasp the structure of the data. The TEI standard allows to encode a vast array of different information. The ParlaMint schema[3] reduces the number of elements significantly, but it still may mean that many of the structures used will not be familiar to everyone. And there are many cross-links between different files, making it even more complex to figure out what all the various elements stand for.

Apart from the repository itself, the data are also made searchable via NoSketchEngine[4] (henceforth NSE), as well as Kontext[5], which makes access a lot easier. However, NSE and Kontext are both very much designed for linguistic research. There are rich metadata about the speaker, the party he/she belongs to, etc. that make it possible to search for words, and get statistical differences in language use between parties, periods, genders, etc. But it is not that easy to just read the texts, or to see which parties are in the parliament at any given time, what topics are being discussed, or who is speaking. So for much of the potential audience of ParlaMint, NSE and Kontext are not optimal interfaces.

We attempted to provide a more generally accessible version of ParlaMint by creating a corpus out of it in the TEITOK (Janssen, 2016) corpus system. TEITOK is a corpus management system that provides linguistic search options in much the same way as NSE, but furthermore provides a document visualization system that provides an easy to read version of the documents. And TEITOK is a modular system that is maintained within LINDAT[6] (the Czech node of CLARIAH), allowing us to add dedicated functions to the system designed specifically to make various data in ParlaMint accessible. In this article, we first give a short overview of TEITOK, then describe how the ParlaMint data were put into TEITOK, and finally describe the functionality of the interface of the TEITOK version of ParlaMint, as is available at LINDAT: https://lindat.mff.cuni.cz/services/teitok/parlamint-40/

[1] https://www.clarin.eu/parlamint
[2] https://www.clarin.si/repository/xmlui/handle/11356/1860
[3] https://clarin-eric.github.io/ParlaMint/
[4] https://www.clarin.si/ske/
[5] https://www.clarin.si/kontext/
[6] https://lindat.mff.cuni.cz/

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 121–126
May 20, 2024.

2. TEITOK

TEITOK is an online corpus platform that combines various corpus tasks into a single platform. Each document in TEITOK is a TEI/XML file. There are various modules that can visualize these TEI/XML files depending on their content. There are modules to edit the content of the TEI/XML files by running NLP tasks over them or performing manual annotations and corrections. And the system can create a searchable corpus out of the set of TEI/XML files. Searching the corpus will render an XML fragment, that will be linked back to the source XML.

TEITOK is an open source repository[7], that is designed to be installed locally. It is intended as a non-intrusive tool that can be customized to the style of the organization or project where it is used rather than impose its own style. Each TEITOK corpus is an independent folder, and can be fully customized. It has been used in a wide range of different projects with installations in various universities around the world[8].

TEITOK has a modular design that makes it easy to create additional modules for custom visualization of documents, or for providing additional information taken from sources other than the corpus documents. There is an ever growing number of modules to work with different types of corpus documents. There are for instance modules to work with manuscript corpora with alignment to facsimile images (Janssen, 2018a). There are modules to work with audio or video corpora with alignment between the audio and the transcription (Janssen, 2021). And there are modules to work with dependency parsed corpora that can visualize dependency trees (Janssen, 2018b).

For the corpus search, by default TEITOK uses the Corpus WorkBench (CWB) (Evert and Hardie, 2011), but it is also possible to use other search engines, including dependency based search languages such as PML-TQ[9] or Grew (Guillaume, 2019), or it is possible to let external tools like Kontext (Machálek, 2020) handle the search (Janssen, 2020). For NLP tasks, the default in TEITOK is to use UDPIPE[10].

TEITOK has been used in many different projects in different universities. At LINDAT, it is used to gradually provide a search interface to all data in the repository, and it is the primary tool for creation and deployment for many new projects. One of the projects made available in TEITOK at LINDAT is ParCzech (Kopp et al., 2021; Kopp, 2024), the Czech parliamentary corpus that forms the basis of the Czech subcorpus of ParlaMint. The experience with ParCzech was one of the main motivations for creating the version of ParlaMint in TEITOK.

3. Converting ParlaMint to TEITOK

TEITOK documents are stored in tokenized TEI/XML format, and so are the files of ParlaMint. So in principle, creating a TEITOK version of the ParlaMint corpus is easy. However, there are differences in the way TEI is used in the two projects. ParlaMint uses an adaptation of the Parla-CLARIN guidelines[11], while TEITOK is designed to work with almost any kind of TEI, but with a limited number of constructions that cannot be used, some deviation from pure TEI, and some prefered constructions that differ from those used in Parla-CLARIN. Therefore, the documents cannot be used directly, but some minor conversions are needed.

Because of ParCzech in TEITOK, much of the conversion was already in place, but still needed to be adapted for ParlaMint. Firstly, ParCzech contains not only the transcription but also the audio recording. Secondly, because there are differences between the different ParlaMint subcorpora that were not accounted for by the scripts. And thirdly, because we needed the TEITOK version of ParlaMint to follow some of the decisions made in the NSE version for consistency.

The conversion consists, apart from some trivial naming differences, in providing information locally as much as possible, rather than distributed as it is in Parla-CLARIN. ParlaMint uses a central repository of names (per subcorpus) and each utterance is linked to a person. People can have multiple affiliations over time, and even multiple names. This is very good for complex cases and for political correctness, but not very helpful for giving a unique answer about the name of a person in a search or in mouse-over information. And the same holds to a lesser extent for information like the dependency relations and the (chronological) order of the transcriptions.

For the conversion, the data corresponding to the repository record were downloaded onto the server, where a local script did all the necessary conversions, subcorpus by subcorpus and one file at a time[12]. The conversion was done with pre-final

[7] https://gitlab.com/maartenes/TEITOK

[8] http://www.teitok.org/index.php?action=projects

[9] https://ufal.mff.cuni.cz/pmltq

[10] https://lindat.mff.cuni.cz/services/udpipe/

[11] https://clarin-eric.github.io/parla-clarin/

[12] Conversions and editing in TEITOK are typically done via the interface, or via the API for larger conversions. But in the case of ParlaMint, no editing is needed since the TEITOK corpus is used just as an interface for existing data and not the primary data source itself. And given the sheer size of the ParlaMint corpus, even the API is too slow for the amount of processing needed.

releases of the corpus, so that the TEITOK corpus could be ready at the time of the launch of ParlaMint. This meant that the conversion had to be rerun due to some last minute corrections in ParlaMint, but also that inconsistencies could be communicated back to the project. And it means that the scripts are streamlined and can be easily used to convert possible future updates of the corpus.

The script compiles the person data from the tab separated text file included in the repository, which was compiled for the NSE version of ParlaMint. The reason for using this compiled file instead of the raw source data is not only that it avoids having to account for complex cases and possible inconsistencies since they have already been dealt with; but also because that way, all decisions will coincide with those taken for the NSE corpus, so that people doing the same search in the different versions of the corpus will (as far as possible) get the same answers. The script also places the dependency information directly on the tokens following the style of CoNLL-U, and introduces pagination markers to be able to display reasonably sized parts of the transcription files in the browser.

4. The TEITOK ParlaMint Interface

4.1. Subcorpora

The ParlaMint corpus in TEITOK is divided into a separate subcorpus for each of the parliamentary sub-parts of ParlaMint. Therefore, the user first has to select the subcorpus he wants to consult. For convenience, the sub-corpora are not only presented as a list, but also shown on a map of Europe, following the CLARIN map style. The subcorpus select page is shown in Figure 1, which is not only convenient, but also gives a quick view of which European countries are included in the ParlaMint release.

Selecting a subcorpus brings you to the landing page of that subcorpus, which is a static page that combines a number of different data about the subcorpus. Let us use CZ as an example:

- It provides the description of the subcorpus, as included in the repository as the README for that subcorpus, converted from MD to HTML - so the contents on README-CZ.md
- It lists the source(s) used for the compilation of the subcorpus, as listed in the `<sourceDesc>` of ParlaMint-CZ.ana.xml, with the title, the link to the source, and the begin and end date used for ParlaMint
- It lists the location(s) where the parliamentary sittings took place, as listed in the `<settingsDesc>` of ParlaMint-CZ.ana.xml
- It lists all the people listed as responsible for the creation of the subcorpus with their role, as listed in the `<respStmt>` of ParlaMint-CZ.ana.xml

This way, the interface makes various types of information that should be pertinent for the use and attribution of the corpus more visible than they are in the repository or the project site.

4.2. Search and Browse

The search provided in the ParlaMint project in TEITOK uses CWB. The search and statistics options are very similar to other interfaces based on the CWB Query Language (CQL) with mostly visualation differences. That inludes the NSE and Kontext interfaces to ParlaMint mentioned earlier, but also the Polish Parliamentary Corpus (PPC)[13], the Plenary Sessions of the Parliament of Finland[14] and many others.

Where the TEITOK interface differs is that from the search result (KWIC list) it brings you to the full document visualization with the matching word highlighted. Many search engine, like Kontext, provide a limited, mostly plain text context or do not provide any larger context at all (often intentionally). And for instance PPC does provide a link to the full context for each search result, but the context is provided as a PDF document with no indication where in the text the matching segment of text can be found.

Another option in the TEITOK interface of ParlaMint is that it allows visitors to browse through the transcriptions, and find transcriptions based by sitting or by date. Browsing by date uses a feature that was introduced for the ParCzech corpus in which a calendar is presented with dates for which transcriptions are available highlighted, and those for which there are not greyed out.

4.3. Document Visualization

The interface can display individual transcription files with various types of information assembled in the interface, as can be seen in Figure 2.

The header of the page contains the pertinent metadata of the file: the country of the parliament; the source it was taken from with a direct link where available; the information about the session - the date, term, meeting, sitting and agenda; and a link to the previous and the next document in the corpus.

[13] https://kdp.nlp.ipipan.waw.pl/query_corpus/

[14] https://www.kielipankki.fi/korp/#corpus=eduskunta&cqp=%5B%5D

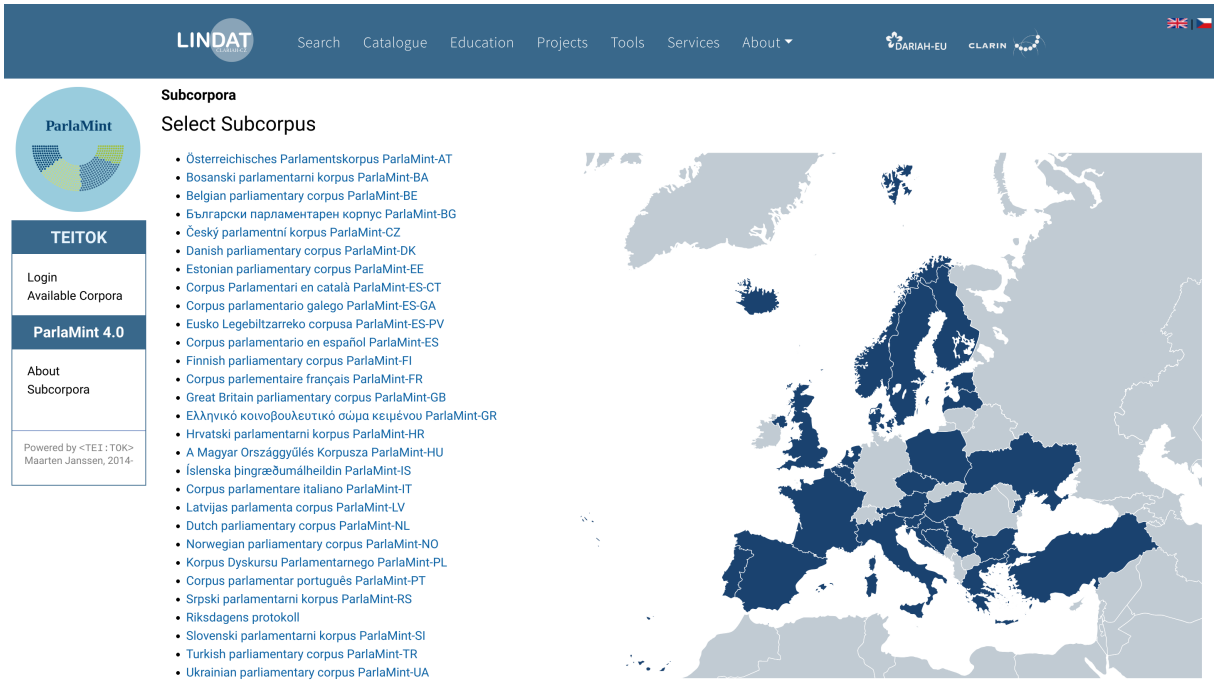


Figure 1: The subcorpus select page

Figure 2: An indidivual transcription

The text itself is a direct visualization of the source TEI/XML file, and hence does not only contain the utterances in the transcription, but also all the comments and other information in the source, which are not in the searchable corpus.

The default visualization in TEITOK is a linguistic view that shows all linguistic annotations about the tokens on mouse-over. Since the expectation is that the majority of visitors of ParlaMint in TEITOK does not have a linguistic background, in the ParlaMint project the linguistic view is instead linked on the bottom of the text, with the default view highlighting all named entities in the text, with information about the type of entity on mouse-over.

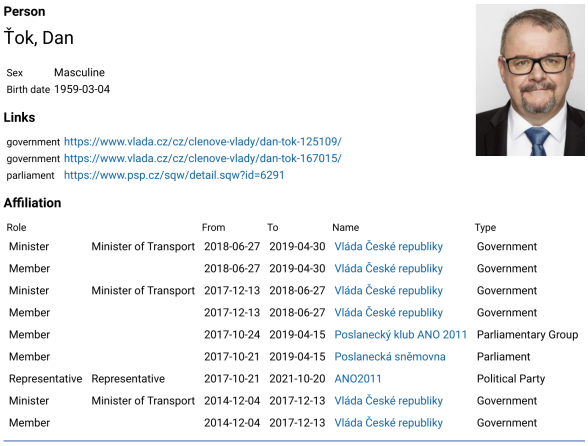
And each speaker is identified on top of the transcription of their speech act, with all available information about the speaker shown on mouse-over: full name, gender, birth year, and the name and political orientation of the party he/she belongs to. As mentioned in the previous section, this information is taken from the tabular data compiled for NSE to make sure that the data are consistent across the different interfaces, and all data should reflect the status of the person and the party at the time of the sitting.

4.4. People and Organizations

Apart from the transcriptions themselves, the TEITOK interface also provides a visualization of the people and organizations in the ParlaMint sources. For each subcorpus, it presents a searchable list of all people in the metadata file. Each person is listed along with its sex and birth date. Clicking on a name will open up a window about that person, with on top all biographical data available, such as name, sex, birth date or photographs. Below that all links to external pages related to that person as present in the ParlaMint sources. And then a list of all the organizations that that person has been a member of, with the name, the period,

and the role of the person in the organization. An example is given in Figure 3.

Similarly, you can also start from the organizations, where each organization provides all available information such as political orientation of the organization, external links, and a list of all people that were a member of that organization, with name, period, and sex.



Person

Ťok, Dan

Sex Masculine
Birth date 1959-03-04

Links

government https://www.vlada.cz/cz/clenove-vlady/dan-tok-125109/
government https://www.vlada.cz/cz/clenove-vlady/dan-tok-167015/
parliament https://www.psp.cz/sqw/detail.sqw?id=6291

Affiliation

Role		From	To	Name	Type
Minister	Minister of Transport	2018-06-27	2019-04-30	Vláda České republiky	Government
Member		2018-06-27	2019-04-30	Vláda České republiky	Government
Minister	Minister of Transport	2017-12-13	2018-06-27	Vláda České republiky	Government
Member		2017-12-13	2018-06-27	Vláda České republiky	Government
Member		2017-10-24	2019-04-15	Poslanecký klub ANO 2011	Parliamentary Group
Member		2017-10-21	2019-04-15	Poslanecká sněmovna	Parliament
Representative	Representative	2017-10-21	2021-10-20	ANO2011	Political Party
Minister	Minister of Transport	2014-12-04	2017-12-13	Vláda České republiky	Government
Member		2014-12-04	2017-12-13	Vláda České republiky	Government

Figure 3: The person record visualization

5. Conclusion

The TEITOK interface for the ParlaMint data provides a more accessible entry point for people with a background other than linguistics. It attempts to focus on those data that the average user is expected to be most interested in, accessing data by people, dates, and organizations. The interface makes all the information in the ParlaMint corpus easy to view and browse. Of course the interface still presents the data as a corpus of text with additional data - so dedicated research for instance in the field of political sciences would likely still need to start from the raw data, but for more cursory access, we believe the interface makes the data accessible to a much wider audience.

The interface could be improved in future versions of ParlaMint. Due to the modular set-up of TEITOK it is easy to add more dedicated functionality over time. For instance, if apart from named entity recognition the names in the transcriptions would also be entity linked. This would make it possible to create an interface for all topics discussed in the parliamentary sitting, which is probably one of the most interesting issues for many people. But named entity recognition alone, especially with many languages in the corpus where names are inflected, does not give satisfactory results.

We are currently working on creating a live version of ParlaMint alongside the static version of ParlaMint 4.0 currently provided. New versions of subcorpora are sometimes released, and for most users, the most pertinent version is the most recent version of the documents. But replacing the searchable corpus would break the reproducibility of published results based on ParlaMint 4.0. So the intention is to keep the version of ParlaMint 4.0 unmodified, while at the same time having a separate version that always contains the most recent version of all subcorpora.

Another direction of work is to make the english version of the corpus accessible[15]. The idea is to leverage those translations in a number of different ways – to provide searches in both the original and the translation, or combinations of those. To display all metadata both in the original and the translations, and to display the two versions next to each other. At this point in time, there are still several issues to be resolved before this can be fully implemented.

6. Acknowledgements

This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ and CLARIN ERIC, ParlaMint: Towards Comparable Parliamentary Corpora.

7. Bibliographical References

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darundefinedis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The parlamint corpora of parliamentary proceedings. *Lang. Resour. Eval.*, 57(1):415–448.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Rodrigo Agerri, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Dargis, Jesse de Does, Ruben de Libano, Griet Depoorter, Katrien Depuydt, Sascha Diwersy, Réka Dodé,

[15] http://hdl.handle.net/11356/1864

Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Nikola Ljubešić, Giancarlo Luxardo, Carmen Magariños, Måns Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Valeria Quochi, Paul Rayson, Xosé Luís Regueira, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Minna Tamper, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer. 2023. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0. http://hdl.handle.net/11356/1860.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics 2011*.

Bruno Guillaume. 2019. Graph Matching for Corpora Exploration. In *JLC 2019 - 10èmes Journées Internationales de la Linguistique de corpus*, Grenoble, France.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4037–4043.

Maarten Janssen. 2018a. Adding words to manuscripts: From pagesxml to TEITOK. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11057 LNCS:152–157.

Maarten Janssen. 2018b. TEITOK as a tool for dependency grammar. *Procesamiento del Lenguaje Natural*, 61:185–188.

Maarten Janssen. 2020. Integrating TEITOK and Kontext at LINDAT. In *Proceedings of CLARIN Annual Conference 2020*, Madrid, Spain. CLARIN.

Maarten Janssen. 2021. A corpus with Wavesurfer and TEI: Speech and video in TEITOK. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*, page 261–268, Berlin, Heidelberg. Springer-Verlag.

Matyáš Kopp. 2024. ParCzech 4.0. http://hdl.handle.net/11234/1-5360.

Matyáš Kopp, Vladislav Stankov, Jan Oldřich Krůza, Pavel Straňák, and Ondřej Bojar. 2021. Parczech 3.0: A large czech speech corpus with rich metadata. In *24th International Conference on Text, Speech and Dialogue*, pages 293–304, Cham, Switzerland. Springer.

Tomáš Machálek. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.