Overview of ParlaMint corpora

Common characteristics of the corpora

ParlaMint is a multilingual set of comparable corpora containing parliamentary debates mostly starting at the end of 2015 and extending to mid-2020, with each corpus being about 20 million words in size. The sessions in the corpora are marked as belonging to the COVID-19 period (starting with Nov. 1st 2019), or being "reference" (before that date).

The corpora have extensive meta-data about the speakers (speaker name, gender, party affiliation, MP status), are structured into time-stamped terms, sessions and meetings, with each speech being marked by its speaker and their role (chair, regular speaker). The speeches also contain marked-up transcriber comments, such as gaps in the transcription, interruptions, applause, etc.

The corpora are encoded according to the Parla-CLARIN TEI recommendation (https://clarin-eric.github.io/Parla-CLARIN/) but have been validated to the conformant with the much stricter ParlaMint schemas, available from the ParlaMint GitHub repository (https://github.com/clarin-eric/ParlaMint). This repository includes, apart from the XML schemas, also content validation scripts, scripts to convert the corpora into other formats, as well as samples from all the available corpora.

In addition to the ParlaMint TEI encoded “plain text” corpora, their linguistically encoded variants (distinguished by the ".ana" suffix) are also available. This annotation includes named entities, lemmatisation, morphological features and syntactic parses according to the Universal Dependencies recommendations. State-of-the-art tools have been used to perform the annotations. Samples of the “plain text” and linguistically annotated corpora, as well as the samples in several derived formats are also available from the ParlaMint GitHub repository.

Metadata

The main metadata types are *Speaker metadata* and *Speech metadata*:

The following static metadata is available for each speaker:

- speaker’s ID
- name and surname(s)
- gender

The following static metadata **may** be available for a speaker:

- birth date
- birth place
- death date
- death place
- education
- employment (although in principle dynamic, this information is typically available only for the employment when the person becomes an MP)
- link to Wikipedia, VIAF or other authoritative records

The following dynamic metadata **must** be available for each speaker:

- status of the speaker (MP, invited speaker)
- political affiliation (only necessary for MPs, not for occasional speakers)

Furthemore, all the political parties are placed into a taxonomy, which gives the abbreviation of the party, its full name in the local language and in English, and the party ID. Then the person metadata only refers to the ID of the party.

Speech metadata

The following metadata **must** be available for each speech:

- date of the speech (but can be approximate, if exact date is unknown)
- speaker ID
- role of the speaker (one of: chair (required), regular (required), irregular (optional)).

The following metadata **may** be available for a speech:

- time of speech
- type of speech (question, answer)
- to whom the speech is addressed
- summary of the speech

Linguistic annotation

All the corpora are automatically linguistically processed on the following levels:

- tokenization and sentence splitting
- morphosyntactic annotation
- lemmatization
- dependency syntax
- named entity recognition

For the morphosyntactic and dependency syntax annotation the Universal Dependencies formalism (https://universaldependencies.org) was used. Named entities were annotated for the following categories: person, location, organization and miscellaneous.

ParlaMint corpora: Polish

Corpus information

The corpus contains the stenographic record of plenary sittings of the Sejm – the lower chamber of the parliament of the Republic of Poland (8th and 9th term of office) and Senate – the upper chamber (9th and 10th term of office). It is composed of two subcorpora: the reference subcorpus, with utterances between 2015-11-12 and 2019-10-31 and COVID subcorpus, between 2019-11-01 and 2020-08-18. Both subcorpora contain 516 files representing individual session days, 330k utterances and 27M words (additional statistics are available at the NoSketch Engine corpus info page).

Data source and acquisition

The data and linguistic annotation was retrieved from the Polish Parliamentary Corpus.

MP metadata (gender, birth date, political affiliations) was retrieved from the websites of Sejm and Senate. The speakers were assigned a role of chairman, regular or guest. All MPs were given the role of regular, even if they are speaking as PM, minister or someone else.

Data encoding process

The data was converted to Parla-CLARIN format from its internal TEI P5 XML representation following the format of the National Corpus of Polish. The conversion was performed with a set of Python scripts. Some errors in the original corpus were automatically corrected during conversion.

Structure

Heuristics were used to convert event descriptions and comments into Parla-CLARIN types, mostly based on typical phrases used in the text:

Event	Type	Typical phrases
note	vote	głosowanie nr
	time	przerwa, początek, koniec, wznowienie
	debate	na posiedzeniu, przewodnictwo, chwila ciszy, chwila przerwy
kinesic	applause	oklaski
	ringing	dzwonek, sygnał telefonu
	laughter	wesołość, śmiech
	signal	uderza laską, pokazuje
	playback	wyświetla, odtwarza, projekcja
vocal	noise	gwar, poruszenie, rozmowy na sali, uderza w pulpity
	shouting	skanduje
gap	reason: inaudible	poza mikrofonem, zakłócenia wypowiedzi, w tle, poza nagraniem, brak nagrania
incident	entering	wchodzi, przybywa, przybycie

Event	Type	Typical phrases
	leaving	wychodzą
	action	wstają, włącza, wyłącza, wręcza, otrzymuje, odczytuje, trzyma, prezentuje, podaje, składa gratulacje

Linguistic annotation

The resource contains automatically created annotation of:

- utterance-level segmentation, tokenization and lemmatization produced with Morfeusz2
- disambiguated morphosyntactic description produced with Concraft2
- named entities produced with Liner2
- dependency structures produced with COMBO parser.

Named entities, originally following the model used by the National Corpus of Polish (NKJP), were converted from PPC annotation in the following way:

NKJP	Parla-CLARIN	Comment
date	–	ignored
geogName	LOC	
orgName	ORG	
persName	PER	subtypes (forename, surname, addName) ignored
placeName	LOC	subtypes (district, settlement, region, country, bloc) ignored
time	–	ignored