Comparability and interoperability of parliamentary corpora: Easier said than done

Tomaž Erjavec

Department of Knowledge Technologies
Jožef Stefan Institute

Ljubljana, Slovenia

CPSS @ KONVENS 2021
1st Workshop on Computational Linguistics for Political Text Analysis
September 6, 2021

Overview of the talk

Introduction

What is CLARIN?

- European research infrastructure for language resources and technologies
- Its goal is to support research communities from Humanities, Social Sciences and other language-related disciplines with:
  - language resources and technologies and
  - expertise and knowledge transfer.
- 22 member countries, each with at least one CLARIN centre

CLARIN work on Parliamentary corpora

CLARIN has organised a number of initiatives and events that deal with parliamentary corpora:

- CLARIN Travelling Campus "Talk of Europe": three "Creative Camps" (2014–2015) used the proceedings of the European Parliament, curated as linked open data
- CLARIN-PLUS cross-disciplinary workshop Working with parliamentary records, Sofia 2017
- CLARIN Resource Families: Parliamentary corpora, 2018-2019
- ParlaCLARIN workshop at LREC 2018
- CLARIN ParlaFormat workshop, Amersfoort, 2019
- Second ParlaCLARIN workshop at LREC 2020

The ParlaMint project

The project

- A mini-project supported by CLARIN-ERIC
- Budget: 98,000 €
- Duration: Jul 1 2020 – May 30 2021
- Motivation: Parliamentary data directly corresponds to the most recent events with global impact on human health, social life and economics such as the current COVID-19 pandemic.
- Goal: Provide resources and tools for focused observations on trends, opinions, decisions on lockdowns and restrictive measures as well as on the consequences with respect to health, medical care systems, employment, etc. during pandemic times.

Implementation

- Phase 1 (July 2020 – Sep 2020):
  Compiled the corpora of Bulgarian, Croatian, Polish and Slovene parliamentary speeches:
  Multilingual comparable corpora of parliamentary debates ParlaMint 1.0. 2020. `http://hdl.handle.net/11356/1345`
- Phase 2 (Dec 2020 – May (June) 2021:
  - call for additional corpora, 13 respondents
  - version 2.0 released towards the end of the project
  - V2.0 used in the Helsinki Digital Humanities Hackathon
  - version 2.1 built on the experience of DHH & fixed some errors
  - V2.1: 17 corpora (countries) with 16 languages and half a billion words

Results

- Downloadable sets of corpora (CC BY) @ CLARIN.SI:
  - Multilingual comparable corpora of parliamentary debates ParlaMint 2.1. 2021. `http://hdl.handle.net/11356/1432`.
  - Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1. 2021, `http://hdl.handle.net/11356/1431`
- Integrated with noSketch Engine and KonText

Data overview

ID	Lang	Houses	Ts	From	To	Yrs	Mw/Yr	Mw
BE	nl+fr	lower	2	2015-11	2020-08	4.8	6.50	31.37
BG	bg	unicameral	2	2014-10	2020-07	5.8	3.42	20.02
CZ	cs	lower	2	2013-11	2021-04	7.5	3.03	22.56
DK	da	unicameral	-	2014-10	2020-09	6.1	4.85	29.40
ES	es	lower	5	2015-01	2020-12	6.0	2.19	13.10
FR	fr	lower	1	2017-07	2020-07	3.0	10.75	32.73
GB	en	lower+upper	4	2015-01	2021-03	6.3	17.25	109.30
HR	hr	unicameral	1	2016-11	2020-05	3.6	5.81	20.65
HU	hu	unicameral	2	2014-05	2020-12	6.7	0.13	0.87
IS	is	unicameral	3	2015-01	2020-09	5.8	4.06	23.66
IT	it	upper	2	2013-03	2020-11	7.8	3.46	26.94
LT	lt	unicameral	2	2012-11	2020-11	8.1	1.82	14.78
LV	lv	unicameral	2	2014-11	2021-02	6.3	1.02	6.48
NL	nl	lower+upper	5	2014-04	2020-11	6.6	7.74	51.45
PL	pl	lower+upper	4	2015-11	2020-08	4.9	5.66	27.45
SI	sl	lower	2	2014-08	2020-07	6.0	3.34	20.19
TR	tr	unicameral	4	2009-04	2021-02	12.0	3.65	43.99

Reference vs. COVID subcorpus: November 1st, 2019.

Metadata on speakers

ID	Prts	C/O	Orgs	Spks	Gender	MP	Affill	Birth	URL
BE	63	10	2	775	548	548	548	548	0
BG	14	4	5	606	606	420	310	534	99
CZ	61	5	851	485	461	366	366	403	463
DK	19	4	2	454	454	446	454	454	0
ES	50	10	2	814	814	764	758	793	0
FR	16	0	100	670	670	609	585	664	0
GB	31	5	2	1,901	1,901	1,865	1,897	0	1,901
HR	16	2	2	322	322	182	186	168	0
HU	10	0	2	194	194	194	194	192	0
IS	10	6	2	205	205	113	201	205	0
IT	42	22	2	739	739	689	589	739	0
LT	13	20	214	799	799	247	233	247	0
LV	11	0	2	219	219	174	174	0	0
NL	29	12	3	492	492	454	457	0	0
PL	10	3	1	1,123	1,122	743	709	742	0
SI	15	8	5	377	377	167	163	193	78
TR	19	3	2	1,237	1,237	1,223	1,203	0	0

Speeches

ID	Speeches	W.Spks	W.NCs	W.MPs	Notes	Incidents
BE	148,425	147,940	116,214	141,340	140,512	865
BG	146,351	146,295	73,981	120,780	0	34,313
CZ	154,460	154,460	72,301	150,957	188,563	25,692
DK	287,144	287,144	137,210	277,835	10,544	0
ES	49,919	27,812	21,414	27,709	46,965	0
FR	481,603	465,590	421,241	437,965	12,498	62,709
GB	552,103	549,710	537,928	547,305	165,648	0
HR	124,496	124,486	62,128	116,716	9	11,842
HU	3,086	3,086	3,086	3,086	2,958	3,752
IS	74,132	74,132	71,693	71,900	99	41,405
IT	79,373	79,373	50,735	78,269	192,855	61,607
LT	244,835	244,835	126,488	229,980	35,406	30,155
LV	122,136	122,136	60,663	117,899	122,136	0
NL	474,964	474,964	351,789	463,629	191,113	0
PL	331,044	331,044	226,046	302,965	9,453	112,786
SI	75,122	75,122	37,216	70,609	85,111	2,337
TR	1,576,728	1,314,150	1,120,439	1,310,804	142,415	0

Linguistic annotation

- Tokens and sentences
- Lemmas
- UD PoS and morphological features
- UD syntactic dependencies
- Named entities (PER, LOC, ORG, MISC)

Corpus encoding

The importance of encoding

- The idea of ParlaMint was that the corpora are encoded as uniformly as possible
- This would allow the corpora to be interoperable, so that e.g. they can be converted to other formats by the same scripts
- However:
  - the corpora have very different source encoding
  - they are differently structured, contain different information, and reflect different parliamentary traditions
  - each was corpus produced by a separate partner

The definition of a rich but constrained format and the possibility to validate the corpora is crucial!

Parla-CLARIN

- "CLARIN ParlaFormat" workshop (2019)
- Introduced a "standard" format for parliamentary corpora called "Parla-CLARIN" (Erjavec and Pančur, 2019)
- Parla-CLARIN is a simple customisation of the Text Encoding Initiative Guidelines:
  `https://github.com/clarin-eric/parla-clarin`
- However, we did write extensive annotation guidelines:
  `https://clarin-eric.github.io/parla-clarin`
- First Parla-CLARIN encoded corpus:
  Pančur, Andrej; Erjavec, Tomaž; Ojsteršek, Mihael; Šorn, Mojca and Blaj Hribar, Neja, 2020, *Slovenian parliamentary corpus (1990-2018) siParl 2.0*, Slovenian language resource repository CLARIN.SI, `http://hdl.handle.net/11356/1300`.

Text Encoding Initiative

- Aim: enabling annotation of digital documents or any type and in any language for the purposes of scholarly analysis
- The TEI Guidelines define and name several hundred useful textual distinctions
- The TEI provides a framework for the definition of multiple schemas
- Probably the oldest still active standardisation effort for text
- Converters to and from TEI: Word, HTML, etc.
- Overview of parliamentary corpora in the scope of the CLARIN resource families showed that the majority of parliamentary corpora are encoded in TEI or in a "TEI-like encoding"

Corpus header

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0"
 xml:id="ParlaMint-CZ" xml:lang="cs">
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title type="main" xml:lang="cs">Český parlamentní korpus
      ParlaMint-CZ [ParlaMint]</title>
    <title type="main" xml:lang="en">Czech parliamentary corpus
      ParlaMint-CZ [ParlaMint]</title>
    <title type="sub" xml:lang="cs">Parlament České republiky,
      Poslanecká sněmovna</title>
    <title type="sub" xml:lang="en">Parliament of the Czech
      Republic, Chamber of Deputies</title>
    <meeting ana="#parla.term #parla.lower #parliament.PSP7"
       n="ps2013">ps2013</meeting>
```

Encoding of political parties

```
<org xml:id="party.BM_365_NZ" role="politicalParty">
 <orgName full="yes" xml:lang="hr">Klub Stranke rada i
   solidarnosti i nezavisnih zastupnika</orgName>
 <orgName full="init">BM 365 i NZ</orgName>
</org>
...
<listRelation>
 <relation name="coalition" mutual="#party.HDZ #party.HNS
 #party.BM_365_NZ #party.SDSS #party.HDS_HSLS_HDSSB"
 from="2016-10-14" to="2020-07-21" ana="#HS.9"/>
 <relation name="opposition" active="#party.GLAS #party.HS
 #party.HSS_Demokrati ..." passive="#government.HR"
 from="2016-10-14" to="2020-07-21" ana="#HS.9"/>
</listRelation>
```

Encoding of speakers

```
<listPerson>
 <head>List of speakers</head>
 <person xml:id="SayeedaWarsi">
  <persName>
   <forename>Sayeeda</forename>
   <surname>Warsi</surname>
  </persName>
  <sex value="F">Female</sex>
  <affiliation from="2007-10-11" ref="#parla.lower" role="MP"/>
  <affiliation from="2007-10-11" ref="#party.CON" role="member"/>
  <affiliation from="2010-05-12" to="2012-09-06" ref="#PoGB"
    role="minister"/>
  <affiliation from="2012-09-06" to="2014-08-05" ref="#PoGB"
    role="minister"/>
  <idno subtype="contact"
   type="URI">https://members.parliament.uk/member/3839/contact</idno>
  <figure>
   <graphic url="https://api.parliament.uk/photo/Paa3j0vS.jpg"/>
  </figure>
 </person>
```

Endcoding of transcriptions

```
<text ana="#reference">
 <body>
  <div type="debateSection">
   <head type="session">1. seja</head>
   <note type="time">Seja se je začela ob 10. uri.</note>
   <note type="speaker">PREDSEDUJOČA MARJANA KOTNIK POROPAT:</note>
   <u who="#KotnikPoropatMarjana"
      xml:id="ParlaMint-SI_2014-08-01-SDZ7-Redna-01.u1" ana="#chair">
    <seg xml:id="ParlaMint-SI_2014-08-01-SDZ7-Redna-01.seg1">
      Spoštovani, prosim, da zasedete svoja mesta.</seg>
    <seg xml:id="ParlaMint-SI_2014-08-01-SDZ7-Redna-01.seg2">V naši
      sredini pozdravljam predsednika države, gospoda Boruta Pahorja.
    </seg>
    <kinesic type="applause">
     <desc>aplavz</desc>
    </kinesic>
    <gap reason="editorial">
     <desc>himna</desc>
    </gap>
   ...
```

Encoding of the linguistic annotation

```
<s xml:id="ParlaMint-CZ_...s2">
 <w xml:id="ParlaMint-CZ_...w1" lemma="dovolit"  msd="UPosTag=VERB|
   Aspect=Perf|Mood=Imp|Number=Plur|Person=2|Polarity=Pos|
   VerbForm=Fin">Dovolte</w>
 <w xml:id="ParlaMint-CZ_...w2" lemma="já" msd="UPosTag=PRON|
   Case=Dat|Number=Sing|Person=1|PronType=Prs|Variant=Short"
   join="right">mi</w>
 <pc xml:id="ParlaMint-CZ_...w3" msd="UPosTag=PUNCT">,</pc>
 <w xml:id="ParlaMint-CZ_...w4">abych
   <w xml:id="ParlaMint-CZ_...w5" lemma="aby" msd="UPosTag=SCONJ"
   norm="aby"/>
   <w xml:id="ParlaMint-CZ_...w6" lemma="být" msd="UPosTag=AUX|
     Mood=Cnd|Number=Sing|Person=1|VerbForm=Fin" norm="bych"/>
 </w>
 ...
 <pc xml:id="ParlaMint-CZ_...w9" msd="UPosTag=PUNCT">.</pc>
 <linkGrp targFunc="head argument" type="UD-SYN">
   <link ana="ud-syn:root"
     target="#ParlaMint-CZ_...s2 #ParlaMint-CZ_...w1"/>
   ...
```

Project work-flow

Workflow for Version 1

- We started off with encoding the 4 ParlaMint corpora for Version 1 according to the Parla-CLARIN schema
- During this process we unified the encoding and produced much stricter RelaxNG XML schemas just for validating ParlaMint corpora
- We also wrote XSLT scripts to convert the corpora to other formats:
  - into vertical files for mounting on the concordancers
  - into CoNLL-U files
- The process was not too complicated: few partners, also with previous collaboration

Workflow for Version 2

- Much more complicated: 13 new partners, most unknown
- Also new types of phenomena to encode: extensions and changes to schema (sometimes necessary to fix V1 corpora)
- Huge scope for errors (x 13!)
- The idea was for each partner to fully prepare their corpus and only submit it for validation
- Communication: GitHub issues, but also *lots* of emails

Validation

For V2 validation was crucial:

- XML schemas
- XSLT scripts to check more complicated constraints
- Derived CoNLL-U files checked with Universal Dependencies validations scripts:
  a number of formal errors found in linguistic encoding
- Functional validation:
  - using the corpora in the concordancers
  - preparing overview tables of the corpora
- Despite all the validation, we had to write a XSLT script to "polish" the corpora

GitHub

- Using Git was quite helpful for the project
  (but could've been even more so)
- **`https://github.com/clarin-eric/ParlaMint`**
  - XML schemas
  - samples of all corpora in XML:
  - also in the derived formats
    (plain text, TSV metadata files, CoNLL-U format, vertical format)
  - XSLT and Perl scripts for validation and conversion
  - Some derived metadata information

Conclusions

Conclusions

- Presented the ParlaMint project, corpora and encoding
- Work-flow, in retrospect:
  - develop encoding guidelines and strict validation from the outset
  - allow partners to submit only non-redundant information in their corpora
    (rather than checking/correcting such info later)
  - insist on using GitHub issues?

Further work

We will apply to a continuation of the project:

- Better documentation, validation
- Better Git(Hub) rules and control
- More corpora
- Extend current corpora in time and with metadata
- MT all the corpora to English
- Experiment with adding speech data
- Using the corpora: DHH 2021, shared task, tutorial

ParlaMint Compilers

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 Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer.
+ Henk van der Pol, Griet Depoorter (BE); Vladislava Grigorova (BG); Barbora Hladká (CZ); Bart Jongejan, Dorte Haltrup Hansen (DK); Sascha Diwersy (FR); Miklós Sebők (HU); Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Andrea Cimino, Roberto Bartolini (IT); Andrius Utka, Mindaugas Petkevičius, Monika Briedienė (LT); Michał Lenart, Daniel Janus, Bartłomiej Nitoń (PL).

Comparability and interoperability of parliamentary corpora: Easier said than done

Tomaž Erjavec

Department of Knowledge Technologies
Jožef Stefan Institute

Ljubljana, Slovenia

CPSS @ KONVENS 2021
1st Workshop on Computational Linguistics for Political Text Analysis
September 6, 2021