The ParlaMint Tutorial

– demo of a CLARIN teaching material –

Darja Fišer* and Kristina Pahor de Maiti**

*University of Ljubljana; Jožef Stefan Institute; Institute of Contemporary History, Slovenia
**CY Cergy Paris University, France

Today's plan

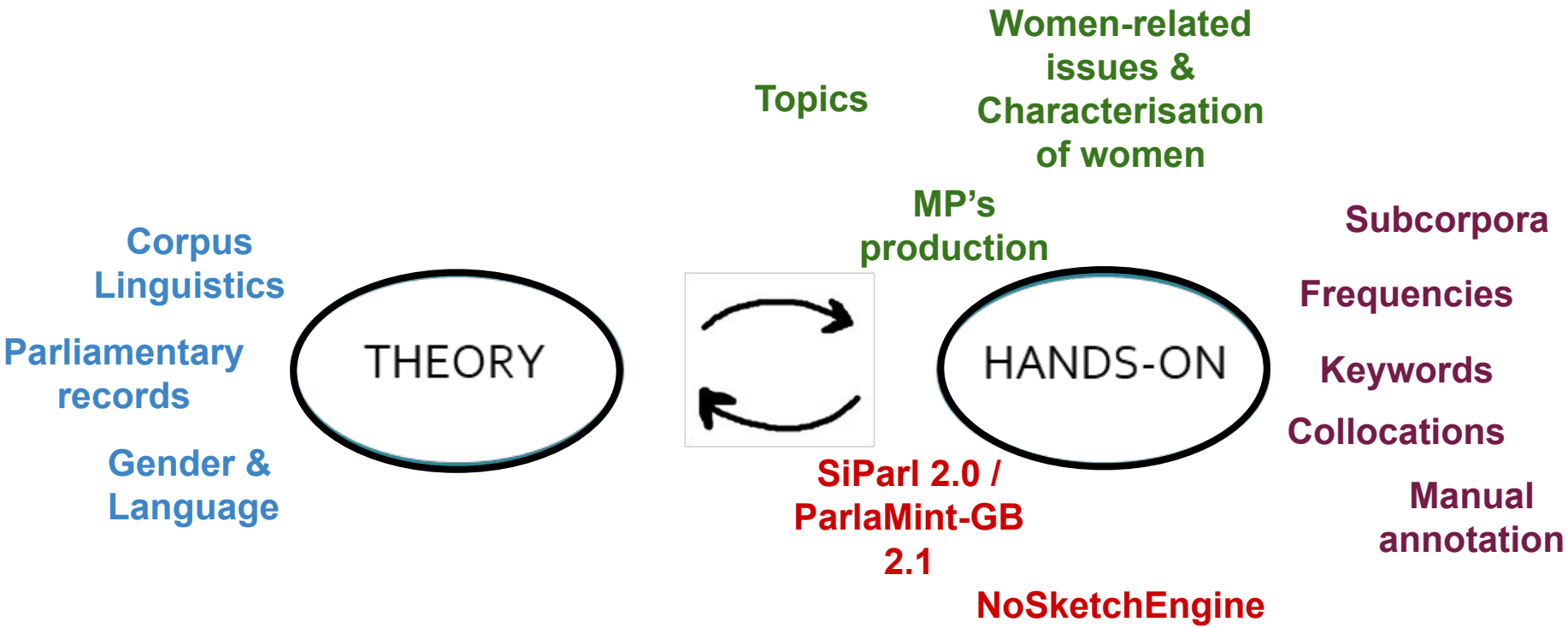
- Demo of the tutorial *Voices of the parliament*
  - tutorial overview
  - showcasing analyses adapted to the ParlaMint-GB data
    - step-by-step instructions + screen casts in the tutorial
    - special parameter tweaks for this analysis provided on the slides

- Some reflections on the tutorial creation process
  - aim & target audience
  - structure
  - sustainability & adaptability

Tutorial demo

Image by Joakim Honkasalo

OVERVIEW



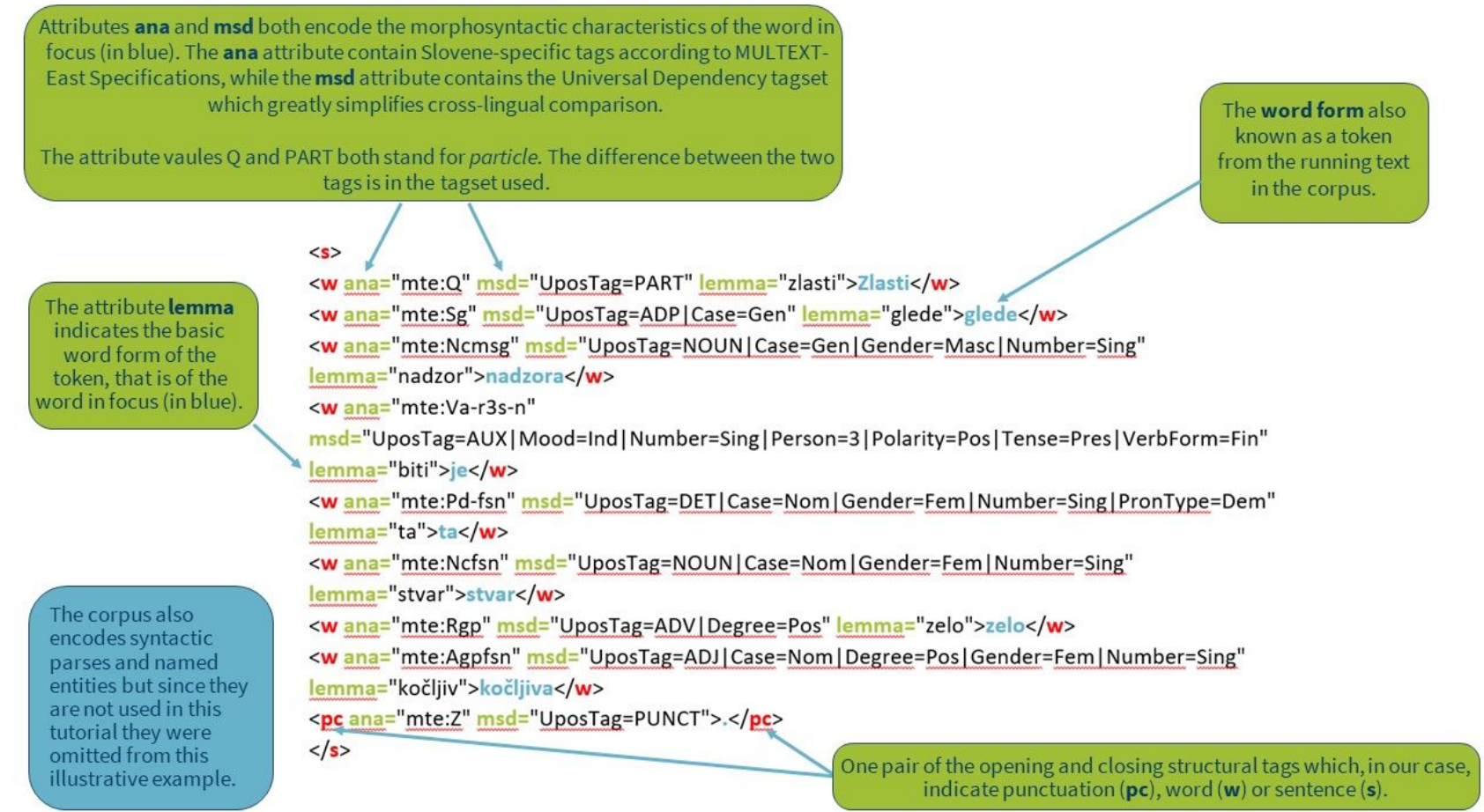
Voices of the Parliament – A Corpus Approach to Parliamentary Discourse Research

SiParl – Slovene parliamentary debates 1990–2018

THEORETICAL PART

CORPORA and CONCORDANCERS	PARLIAMENTARY DISCOURSE, RECORDS and CORPORA	GENDER, LANGUAGE and POLITICS
● What are language corpora? ● How is data encoded? ● What is a token? ● What are different levels of annotation? ● Why use linguistic annotations? ● What is lemmatization? ● What is PoS tagging? ● What is metadata? ● What is a concordancer? ● Which are different concordancers available? ● What are the most popular corpus analysis techniques?	● What are parliamentary corpora? ● What are the most often included metadata? ● What are the specifics of parliamentary discourse? ● How faithful are the records to the actual spoken word? ● Why is it important to know your data?	● Is there a women-specific debating style? ● Which topics do women and men debate in the parliaments? ● What are the dangers of performing analyses when knowing a parameter, such as gender, in advance?

Data encoding



Faithfulness of the records

Official records

"Madam President, I think we all agreed that equal pay is, first of all, a matter of elementary fairness, and the fact that we still have this average pay gap of 16-17% also shows, among other things, the lack of respect for women, which is so deeply rooted in traditions and stereotypes. /.../."

(Věra Jourová, Member of the Commission)

Verbatim transcription

*"**Ladies and Gentlemen, yes gentlemen also are here still. Am**... I think we all agreed that equal pay is, first of all, a matter of elementary fairness, and **that** the fact that still **we** have this **pay gap am**... average 16-17% **this** also shows, among other things, the lack of respect to women, which is so deeply rooted in traditions and stereotypes. /.../."*

(Věra Jourová, Member of the Commission)

ParlaMint

- The ParlaMint corpora
  - comparable and uniformly annotated corpora of parliamentary sessions
    - more about the ParlaMint project
    - half a billion words or 5 million speeches produced by around 11 thousand speakers
  - records of 17 national parliaments
    - Belgium, Bulgaria, Croatia, the Czech Republic, Denmark, France, Great Britain, Hungary, Iceland, Italy, Latvia, Lithuania, the Netherlands, Poland, Romania, Slovenia, and Turkey
  - (at least) 2015–mid 2020
    - all corpora divided into *Reference* (until 2019-10) and *COVID* (from 2019-11) subcorpora
  - richly annotated
    - linguistic annotations (lemma, PoS, morphological and syntactic features, NE)
    - metadata (SPEECH: date, session, title …; SPEAKER: name, party, role, type, gender, birth)

ParlaMint-GB*

ID	Lang	Houses	Terms	From	To	Yrs	Mill. words/Yr	Mill. Words
GB	en	Lower+Upper	4	2015–01	2021–03	6.3y	17.25	109.30

ID	Parties	C/O	Orgs	Prsns	Gender	MP	URL	IMG
GB	31	5	2	1901	1901	1865	1901	1029

ID	Spchs	By KnwnSpks	By NChairs	ByMPs	Heads	Notes	COVID (2019-11)	REF
GB	552,103	549,71	537,928	547,305	31,389	165,65	109940	442,163

Erjavec et al. 2022

*Now available at the new noSketchEngine concordancer as well as in the SketchEngine concordancer

PRACTICAL PART

- Pre-task: creation of subcorpora (CQL)
- Task 1: production of speeches by M and F MPs (frequencies)
- Task 2: topics debated by F and M MPs (keywords)
- Task 3: women-related issues by F and M MPs and characterisation of women and men (collocations, UD tags)

Pre-task: subcorpora creation

- <speech speaker_gender="F|M" & house="Lower house" & subcorpus="COVID"/>
- <speech speaker_gender="F|M" & house="Lower house" & from>="2018-07-03" & from<="2019-10-31"/>

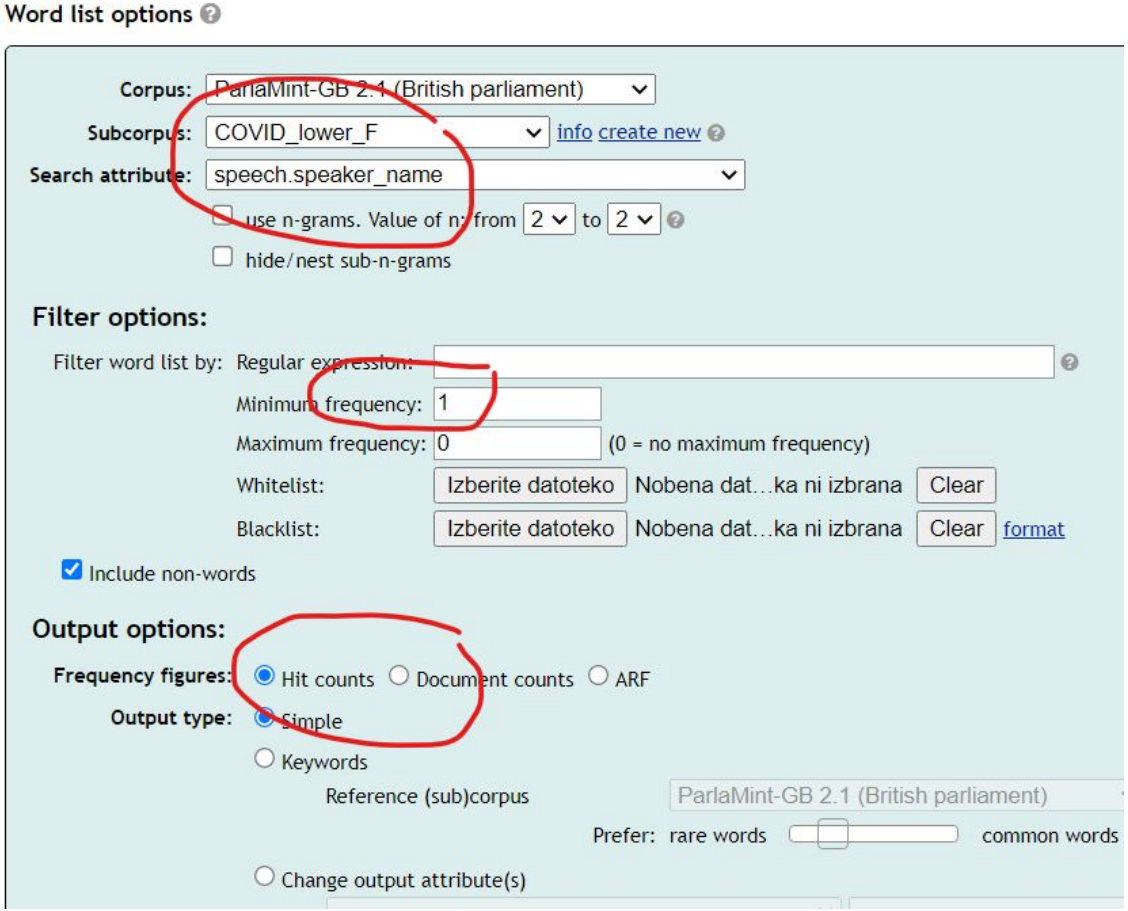
	REF (tokens)	COVID
lower_F	4,950,325	4,023,907
lower_M	10,033,622	9,243,264
lower_FM	14,983,947	13,267,171

TASK 1: Production of speeches

SPKRS	M	%	F	%	TOTAL
COV	453	66.4	229	33.6	682
REF	435	67.7	208	32.3	643
SPCHS					
COV	50,966	72.3	19,517	27.7	70,483
REF	66,788	69.3	29,638	30.7	96,426
TKNS					
COV	9,243,264	69.7	4,023,907	30.3	13,267,171
REF	10,033,622	67.0	4,950,325	33.0	14,983,947

- What is the proportion of MPs per gender?
- How often the MPs take the floor?
- How much MPs' talk?

TASK 1: frequencies



http://www.clarin.si/noske/parlamint21.cgi/wordlist_form?corpname=parlamint21_gb

TASK 1: Production of speeches

AVG SPCH LENGTH	M	F	TOTAL
COV	181.36	206.17	188.23
REF	150.23	167.02	155.39
AVG No. of SPCH/DAY			
COV	278.50	106.65	385.15
REF	335.62	148.93	484.55

- How long are the MPs' speeches?
- How active are the MPs in both periods on a daily basis?

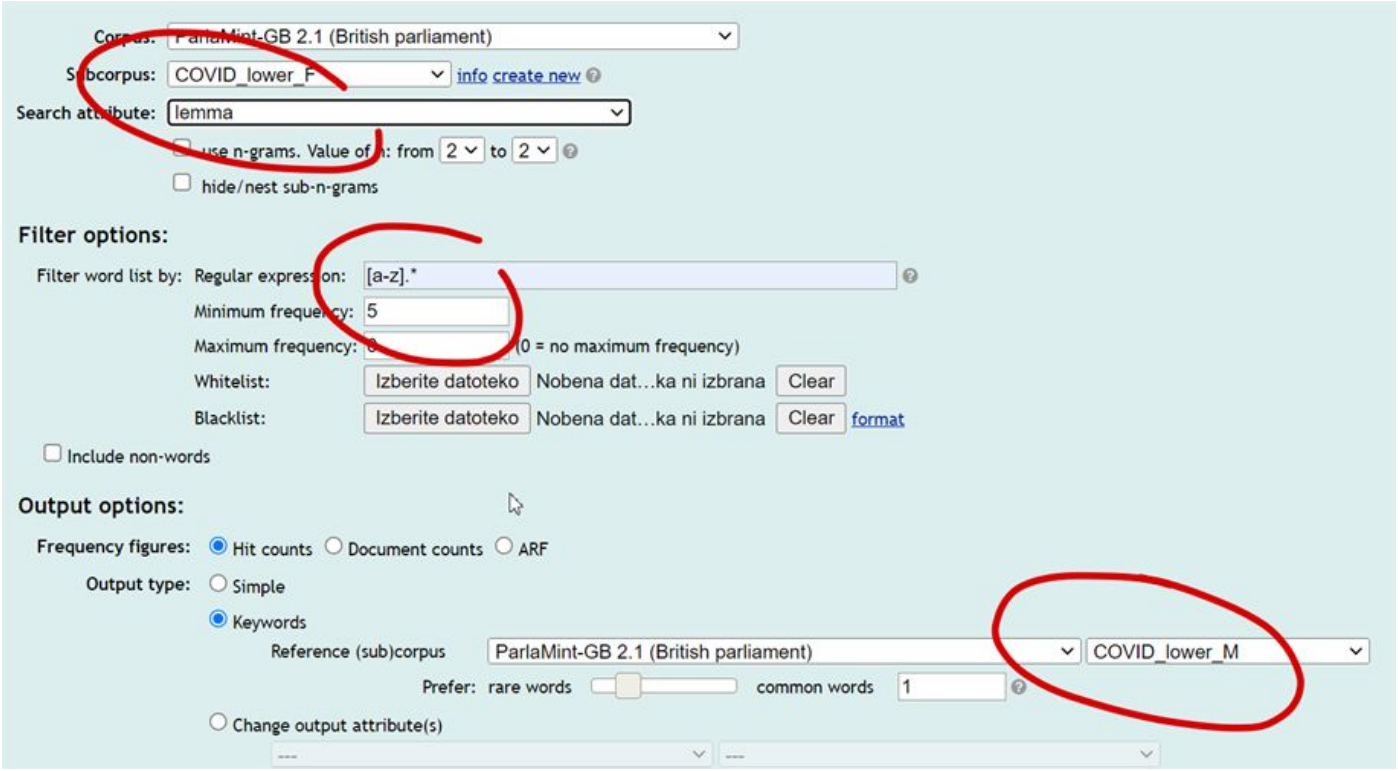
TASK 2: Topic preference

parlamint21_gb COVID_lower_F	parlamint21_gb COVID_lower_M	parlamint21_gb REF_last15mths_lower_F	parlamint21_gb REF_last15mths_lower_M
hysteroscopy	interconnector	nda	journalistic
moat	highland	pml	seahorse
pesticide	courtroom	lantern	ofs
domicile	orthopaedic	defibrillator	a40
pregnant	alas	abnormality	proposition
maternity	forbearance	**pornography**	beetle
sibling	le	**asc**	highland
pedicab	proposition	lara	borrower
priory	distinguished	**baby**	motorsport
explicable	flatten	bladed	unauthorised
loo	characteristically	**post-natal**	m26
bame	evidential	**incest**	manifestly
vascular	chairman	**pornographic**	gtr
marble	magistrate	moorland	alacrity
anti-racist	cyber	**menopause**	fertiliser
slp	diocese	**underpaid**	fella
midwifery	e3	**upskirting**	sharia
unequal	agility	**autistic**	mellifluous
pregnancy	coalfield	**cervical**	nationalisation
transgender	occasionally	crossbow	encampment

- Topic analysis of top 50 keywords
  - [a–z.*]: to skip NE

- **Reproduction**
- **Equality**

TASK 2: keywords



http://www.clarin.si/noske/parlamint21.cgi/wordlist_form?corpname=parlamint21_gb

COVID Key

https://www.clarin.si/noske/run.cgi/wordlist?corpname=parlamint21_gb&viewmode=kwic&attrs=word%2Cdep&ctxattrs=word&structs=speech&refs=%3Dspeech.from%2C%3Dspeech.speaker_name&pagesize=50&gdexconf=&attr_tooltip=nott&wlmaxitems=100&wlsort=f&subcnorm=freq&corpname=parlamint21_gb&reload=&usesubcorp=COVID_lower_F&wlattr=lemma&usengrams=0&ngrams_n=2&ngrams_max_n=2&nest_ngrams=0&wlpat=%5Ba-z%5D.*&wlminfreq=5&wlmaxfreq=0&wlfile=&wlblacklist=&wlnums=frq&wltype=keywords&ref_corpname=parlamint21_gb&ref_usesubcorp=COVID_lower_M&simple_n=1

REF Key

https://www.clarin.si/noske/run.cgi/wordlist?corpname=parlamint21_gb&viewmode=kwic&attrs=word%2Cdep&ctxattrs=word&structs=speech&refs=%3Dspeech.from%2C%3Dspeech.speaker_name&pagesize=50&gdexconf=&attr_tooltip=nott&wlmaxitems=100&wlsort=f&subcnorm=freq&corpname=parlamint21_gb&reload=&usesubcorp=REF_last15mths_lower_F&wlattr=lemma&usengrams=0&ngrams_n=2&ngrams_max_n=2&nest_ngrams=0&wlpat=%5Ba-z%5D.*&wlminfreq=5&wlmaxfreq=0&wlfile=&wlblacklist=&wlnums=frq&wltype=keywords&ref_corpname=parlamint21_gb&ref_usesubcorp=REF_last15mths_lower_M&simple_n=1

TASK 3.1: Women-related issues

parlamint21_gb	parlamint21_gb
COVID_lower_F	COVID_lower_M
771.14/mill mentions	169.85/mill mentions
pregnant	man
man	girl
girl	servicemen
young	pregnant
against	sport
equalities	bisexual
who	brave
's	armed
violence	lbt
migrant	young
black	trans
committee	equalities
many	violence
rights	waspi
more	serviceman
face	lesbian
affect	1950
woman	against
particularly	child
and	black

	F (%)	M (%)
collective reference	5.88	**17.95**
equality & representation	**55.88**	**53.85**
reproduction	5.88	**10.26**
violence & other problems	**32.35**	17.95

F	M
face, experience, encourage, work, affect, protect	suffer, encourage

F	M
continue training and take up career, run for office, other	participate in sport

- Topic analysis of top 50 collocations for [lemma = "woman"], win.3
- verbs → *encourage*
- content analysis of concordances for *encourage*

TASK 3.1: collocations

	CONCORDANCES lemma *woman*	**COLLOCATIONS** lemma (lc); -3/+3
COVID F	https://www.clarin.si/noske/run.cgi/first?corpname=parlamint21_gb&reload=&iquery=&queryselector=lemmarow&lemma=woman&phrase=&word=&char=&cql=&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&usesubcorp=COVID_lower_F&fsca_speech.from=&fsca_speech.to=&fsca_speech.speaker_name=	https://www.clarin.si/noske/run.cgi/collx?q=aword%2C%5Blemma%3D%22woman%22%5D&corpname=parlamint21_gb&viewmode=kwic&attrs=word%2Clemma%2Cdep&ctxattrs=word&structs=speech&refs=%3Dspeech.from%2C%3Dspeech.speaker_name&lemma=woman&pagesize=50&usesubcorp=COVID_lower_F&gdexconf=&attr_tooltip=nott&cattr=lemma_lc&cfromw=-3&ctow=3&cminfreq=5&cminbgr=3&cmaxitems=50&cbgrfns=t&cbgrfns=m&cbgrfns=d&csortfn=d
COVID M	https://www.clarin.si/noske/run.cgi/first?corpname=parlamint21_gb&reload=&iquery=&queryselector=lemmarow&lemma=woman&phrase=&word=&char=&cql=&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&usesubcorp=COVID_lower_M&fsca_speech.from=&fsca_speech.to=&fsca_speech.speaker_name=	https://www.clarin.si/noske/run.cgi/collx?q=aword%2C%5Blemma%3D%22woman%22%5D&corpname=parlamint21_gb&viewmode=kwic&attrs=word%2Clemma%2Cdep&ctxattrs=word&structs=speech&refs=%3Dspeech.from%2C%3Dspeech.speaker_name&lemma=woman&pagesize=50&usesubcorp=COVID_lower_M&gdexconf=&attr_tooltip=nott&cattr=lemma_lc&cfromw=-3&ctow=3&cminfreq=5&cminbgr=3&cmaxitems=50&cbgrfns=t&cbgrfns=m&cbgrfns=d&csortfn=d
REF F	https://www.clarin.si/noske/run.cgi/first?corpname=parlamint21_gb&reload=&iquery=&queryselector=lemmarow&lemma=woman&phrase=&word=&char=&cql=&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&usesubcorp=REF_last15mths_lower_F&fsca_speech.from=&fsca_speech.to=&fsca_speech.speaker_name=	https://www.clarin.si/noske/run.cgi/collx?q=aword%2C%5Blemma%3D%22woman%22%5D&corpname=parlamint21_gb&viewmode=kwic&attrs=word%2Clemma%2Cdep&ctxattrs=word&structs=speech&refs=%3Dspeech.from%2C%3Dspeech.speaker_name&lemma=woman&pagesize=50&usesubcorp=REF_last15mths_lower_F&gdexconf=&attr_tooltip=nott&cattr=lemma_lc&cfromw=-3&ctow=3&cminfreq=5&cminbgr=3&cmaxitems=50&cbgrfns=t&cbgrfns=m&cbgrfns=d&csortfn=d
REF M	https://www.clarin.si/noske/run.cgi/first?corpname=parlamint21_gb&reload=&iquery=&queryselector=lemmarow&lemma=woman&phrase=&word=&char=&cql=&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&usesubcorp=REF_last15mths_lower_M&fsca_speech.from=&fsca_speech.to=&fsca_speech.speaker_name=	https://www.clarin.si/noske/run.cgi/collx?q=aword%2C%5Blemma%3D%22woman%22%5D&corpname=parlamint21_gb&viewmode=kwic&attrs=word%2Clemma%2Cdep&ctxattrs=word&structs=speech&refs=%3Dspeech.from%2C%3Dspeech.speaker_name&lemma=woman&pagesize=50&usesubcorp=REF_last15mths_lower_M&gdexconf=&attr_tooltip=nott&cattr=lemma_lc&cfromw=-3&ctow=3&cminfreq=5&cminbgr=3&cmaxitems=50&cbgrfns=t&cbgrfns=m&cbgrfns=d&csortfn=d

	CONCORDANCE of *encourage* from collocation candidates
F	https://www.clarin.si/noske/run.cgi/sortx?q=aword%2C%5Blemma%3D%22woman%22%5D&q=P-3+3+1+%5Blemma_lc%3D%22encourage%22%5D&q=sword%2Fi+-1%3C0%7E-3%3C0:corpname=parlamint21_gb&viewmode=kwic&attrs=word%2Clemma%2Cdep&ctxattrs=word&structs=speech&refs=%3Dspeech.from%2C%3Dspeech.speaker_name&lemma=woman&pagesize=50&usesubcorp=COVID_lower_F&gdexconf=&attr_tooltip=nott:skey=rc:sicase=i
M	https://www.clarin.si/noske/run.cgi/sortx?q=aword%2C%5Blemma%3D%22woman%22%5D&q=P-3+3+1+%5Blemma_lc%3D%22encourage%22%5D:corpname=parlamint21_gb&viewmode=kwic&attrs=word%2Clemma%2Cdep&ctxattrs=word&structs=speech&refs=%3Dspeech.from%2C%3Dspeech.speaker_name&lemma=woman&pagesize=50&usesubcorp=COVID_lower_M&gdexconf=&attr_tooltip=nott:skey=rc:sicase=i

TASK 3.2: Characterization of women and men

parlamint21_gb	parlamint21_gb
1297 hits for WOMAN etc.	623 hits for MAN etc.
underrepresented	breadwinner
pregnant	island
likely	labourer
twice	vicar
frightened	imam
unaware	likely
reluctant	merchant
polish	twice
concentrated	physically
unhappy	dead
desperate	mother
unable	continent
dead	entire
capable	soldier
fearful	innocent
less	perpetrator
absent	graduate
active	victim
angry	stronger
disadvantaged	father

- Analysis of top 50 collocations for
  - woman|girl|mother|female .*
  - man|boy|father|male .*
- in the subject position followed by a copula verb
  - CQL: [lemma="woman|girl|mother|female .*" & dep="nsubj"] [dep="cop"]
  - (the entire ParlaMint-GB corpus used)

	WOMAN (%)	MAN (%)
positive	25.93	20.83
negative	**66.67**	25.00
profession	7.41	**50.00**

TASK 3.2: UD search + collocations

Entire ParlaMint-GB corpus used!

CQL queries:

- [lemma="woman|girl|mother|female .*" & dep="nsubj"] [dep="cop"]
- [lemma="man|boy|father|male .*" & dep="nsubj"] [dep="cop"]

Collocation candidates: lemma (lc); 0/+3

	CONCORDANCES	COLLOCATIONS
WOMAN	https://www.clarin.si/noske/run.cgi/first?corpname=parlamint21_gb&reload=&iquery=&queryselector=cqlrow&lemma=&phrase=&word=&char=&cql=%5Blemma%3D%22woman%7Cgirl%7Cmother%7Cfemale+.*%22+%26+dep%3D%22nsubj%22%5D+%5Bdep%3D%22cop%22%5D&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&usesubcorp=&fsca_speech.from=&fsca_speech.to=&fsca_speech.speaker_name=	https://www.clarin.si/noske/run.cgi/collx?q=aword%2C%5Blemma%3D%22woman%7Cgirl%7Cmother%7Cfemale+.*%22+%26+dep%3D%22nsubj%22%5D+%5Bdep%3D%22cop%22%5D&corpname=parlamint21_gb&viewmode=kwic&attrs=word%2Clemma%2Cdep&ctxattrs=word&structs=speech&refs=%3Dspeech.from%2C%3Dspeech.speaker_name&pagesize=50&gdexconf=&attr_tooltip=nott&cattr=lemma_lc&cfromw=0&ctow=3&cminfreq=5&cminbgr=3&cmaxitems=50&cbgrfns=t&cbgrfns=m&cbgrfns=d&csortfn=d
MAN	https://www.clarin.si/noske/run.cgi/first?corpname=parlamint21_gb&reload=&iquery=&queryselector=cqlrow&lemma=&phrase=&word=&char=&cql=%5Blemma%3D%22man%7Cboy%7Cfather%7Cmale+.*%22+%26+dep%3D%22nsubj%22%5D+%5Bdep%3D%22cop%22%5D&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&usesubcorp=&fsca_speech.from=&fsca_speech.to=&fsca_speech.speaker_name=	https://www.clarin.si/noske/run.cgi/collx?q=aword%2C%5Blemma%3D%22man%7Cboy%7Cfather%7Cmale+.*%22+%26+dep%3D%22nsubj%22%5D+%5Bdep%3D%22cop%22%5D&corpname=parlamint21_gb&viewmode=kwic&attrs=word%2Clemma%2Cdep&ctxattrs=word&structs=speech&refs=%3Dspeech.from%2C%3Dspeech.speaker_name&pagesize=50&gdexconf=&attr_tooltip=nott&cattr=lemma_lc&cfromw=0&ctow=5&cminfreq=5&cminbgr=3&cmaxitems=50&cbgrfns=t&cbgrfns=m&cbgrfns=d&csortfn=d

Creating a tutorial

Image by Alexas Fotos

Identifying the aim

- Transferring CorpLing techniques to **other SSH disciplines** (other than linguistics)
- Making it relevant for **trans-national research** (comparable, multilingual corpora)
- Showcasing the potential of **special data type** (parliamentary records)
- Making it available to **non-coders** (corpora containing large amounts of data)
- Making it **accessible** (by using non-commercial resources and tools)
- Making it a **self-study manual** divided into several modules (theory supporting the tasks, mentioning potential errors in interpretation)

Choosing resources and tools

- Parliamentary corpora in CLARIN
  - available for several languages, comparable (ParlaMint)
  - richly annotated (with metadata)
  - well-documented
  - free and open-access
  - available through free online concordancers

- Several paths tested:
  - Scenario 1: Use several parliamentary corpora for comparative analysis (but suboptimal before ParlaMint; uneven time spans, metadata, annotations, different concordancers)
  - Scenario 2: Use English corpora for internationally understandable examples, but the concordancer did not offer the functionalities needed for the selected research problem
  - Scenario 3: siParl corpus (comprehensive in size and annotations) in NoSketchEngine (offering all major CorpLing analytical techniques)

Creating the tutorial

- **Adapting RQ** to the target audience (SSH fields)
- Embedding the analysis into **their theoretical and methodological framework** (manual analysis of topics based on the ministries, in a later study on Comparative Agendas Project schema)
- Making it possible to **skip and return** to individual parts in theory and hands-on (by analytical design and formatting)
- Leveraging the **affordances of the online format** (hyperlinks, cross-references, screencasts)
- Exposing it to reviews and tests

Pitfalls to avoid

- Using commercial tools and resources
- Not checking the maturity of the resource or scheduled updates to the tools
- Making it too specific to hinder adaptability to other data resources
- Not adapting the RQ to the interests of the target audience
- Making it too long or not modular

Why use/adapt the Voices Of The Parliament tutorial?

- **Teach powerful corpus linguistic techniques**
  - No programming skills needed
  - No previous experience with concordancers needed
- **Show elaborated hands-on examples supported by theory**
  - (Screencasts available)
  - Rich in cross-references to dig deeper
- **Help your students assimilate the methods to the point where they can think of possible uses on other data**
  - Similar parliamentary corpora
  - Other discourse types
- **Wide target audience**
  - Researchers outside of linguistics
  - Students in linguistics
  - Anyone interested in multidisciplinary research

Try it out, use it in the classroom, create spin-offs and let us know what you think.

Thank you for your attention!