AKADEMIE DER WISSENSCHAFTEN IN HAMBURG

DGS-KORPUS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Collocations in Sign Language Lexicography: Towards Semantic Abstractions for Word Sense Discrimination

Gabriele Langer, Marc Schulder
University of Hamburg, Institute for German Sign Language and Communication of the Deaf, Germany

Gold Standard in Lexicography

for well-researched written languages

- **Corpus-based:** large ***written*** corpora
  - Written form allows easy computational processing
- NLP Tools **pre-structure data** for lexicographers (e.g. lexical profiles)
  - Build on solid foundation of syntactic/semantic theory
- **Resulting dictionary** entries include:
  - Examples
  - Collocations
  - Multiword expressions, e.g. idiomatic phrases

Spoken Language Text	Signed Language
Sentence Splitting	🤔?
Token Splitting	Segmentation (manual)
Part of Speech Tagging	❌?
Lemmatisation	Glossing (manual)
Syntax Parsing	❌?

NLP

Lexicographic Analysis

Motivation

Aim: Creation of corpus-based sign language dictionary (DW-DGS)

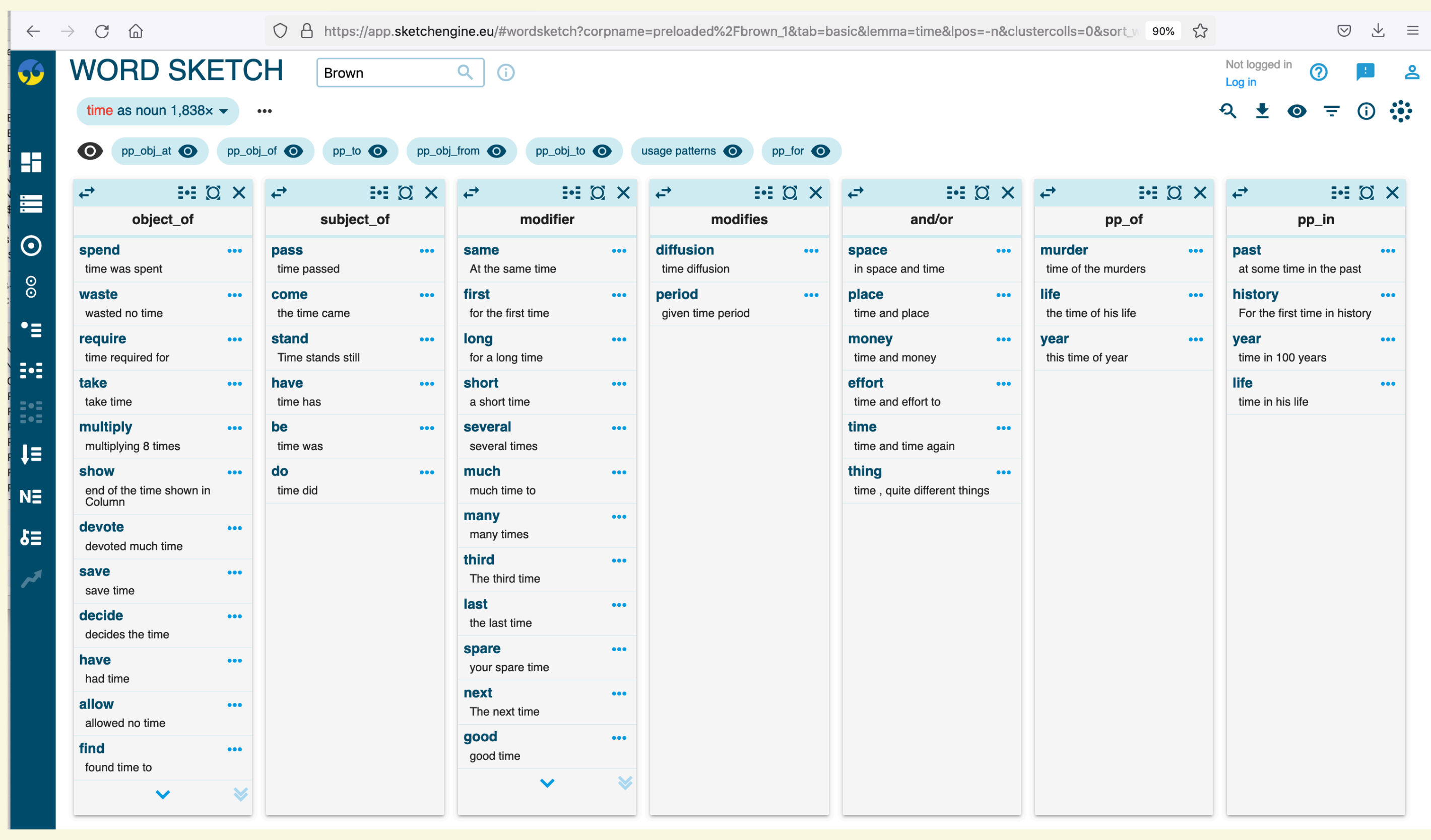
Task: Analysis of **collocations**

- Identification of typical sign combinations
  - Collocations
  - (Loan) compounds
  - Idiomatic phrases
  - Semantic preference patterns

Lexicographic uses:

- Support word/sign sense discrimination (WSD)
- Information to be included in entry
- **Challenge:** Lexicography tools/techniques available for spoken languages (in written form) do not work for signed languages

Lexical Profile

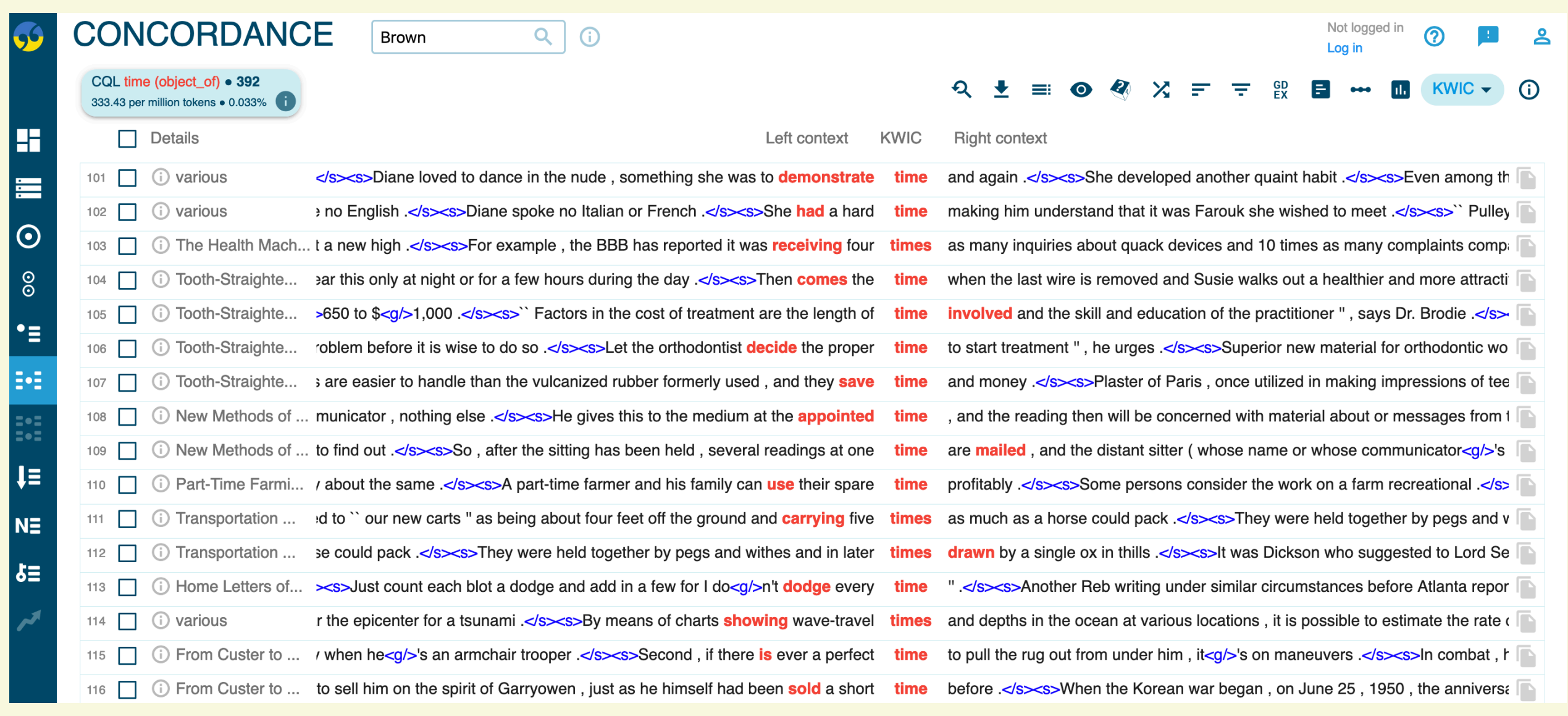


WORD SKETCH

object_of	subject_of	modifier	modifies	and/or	pp_of	pp_in
spend	pass	same	diffusion	space	murder	past
waste	come	first	period	place	life	history
require	stand	long		money	year	year
take	have	short		effort		life
multiply	be	several		time		
show	do	much		thing		
devote		many				
save		third				
decide		last				
have		spare				
allow		next				
find		good				

Collocates grouped by *part of speech* (POS), syntactic roles, etc.

Concordance View (KWIC)



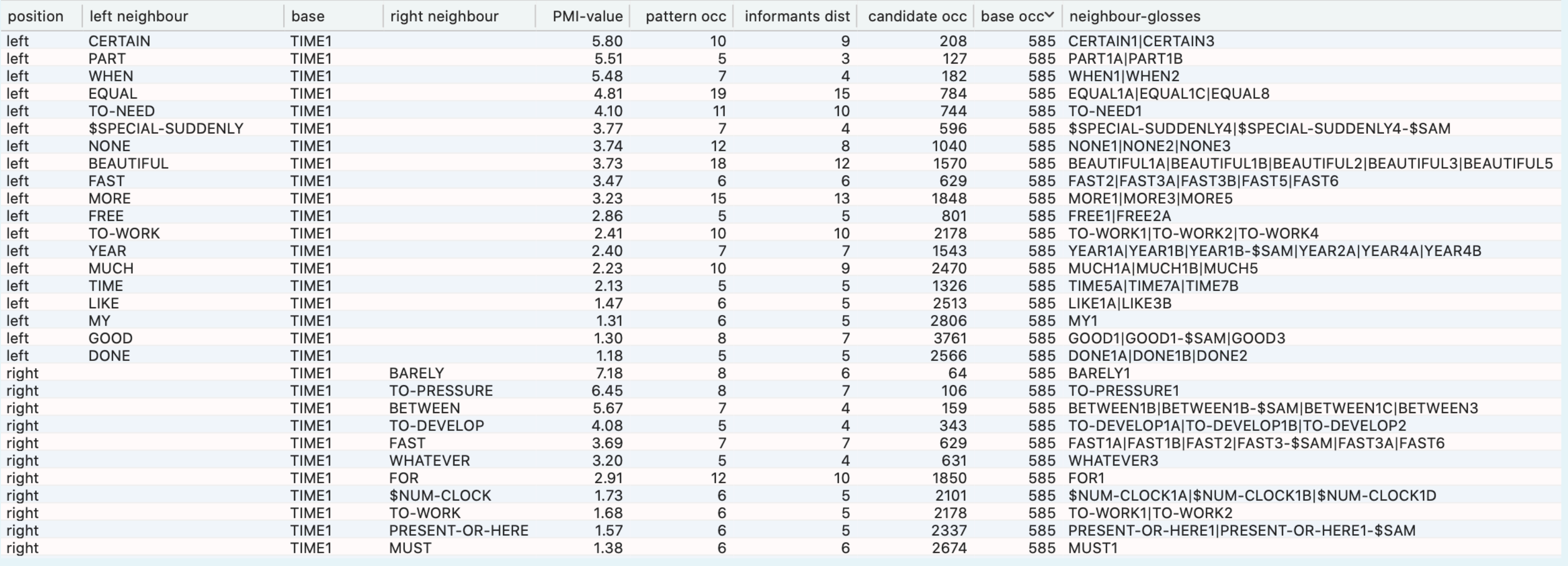
CONCORDANCE

Pre-sorted overview display of individual tokens in context

Profile and KWIC view from SketchEngine Website (BROWN corpus) see https://www.sketchengine.eu/ (2022-06-17)

our work

Gloss Name Collocations



position	left neighbour	base	right neighbour	PMI-value	pattern occ	informants dist	candidate occ	base occ	neighbour-glosses
left	CERTAIN	TIME1		5.80	10	9	208	585	CERTAIN1|CERTAIN3
left	PART	TIME1		5.51	5	3	127	585	PART1A|PART1B
left	WHEN	TIME1		5.48	7	4	182	585	WHEN1|WHEN2
left	EQUAL	TIME1		4.81	19	15	784	585	EQUAL1A|EQUAL1C|EQUAL8
left	TO-NEED	TIME1		4.10	11	10	744	585	TO-NEED1
left	$SPECIAL-SUDDENLY	TIME1		3.77	7	4	596	585	$SPECIAL-SUDDENLY4\|$SPECIAL-SUDDENLY4-$SAM
left	NONE	TIME1		3.74	12	8	1040	585	NONE1|NONE2|NONE3
left	BEAUTIFUL	TIME1		3.73	18	12	1570	585	BEAUTIFUL1A|BEAUTIFUL1B|BEAUTIFUL2|BEAUTIFUL3|BEAUTIFUL5
left	FAST	TIME1		3.47	6	6	629	585	FAST2|FAST3A|FAST3B|FAST5|FAST6
left	MORE	TIME1		3.23	15	13	1848	585	MORE1|MORE3|MORE5
left	FREE	TIME1		2.86	5	5	801	585	FREE1|FREE2A
left	TO-WORK	TIME1		2.41	10	10	2178	585	TO-WORK1|TO-WORK2|TO-WORK4
left	YEAR	TIME1		2.40	7	7	1543	585	YEAR1A|YEAR1B|YEAR1B-$SAM|YEAR2A|YEAR4A|YEAR4B
left	MUCH	TIME1		2.23	10	9	2470	585	MUCH1A|MUCH1B|MUCH5
left	TIME	TIME1		2.13	5	5	1326	585	TIME5A|TIME7A|TIME7B
left	LIKE	TIME1		1.47	6	5	2513	585	LIKE1A|LIKE3B
left	MY	TIME1		1.31	6	5	2806	585	MY1
left	GOOD	TIME1		1.30	8	7	3761	585	GOOD1|GOOD1-$SAM|GOOD3
left	DONE	TIME1		1.18	5	5	2566	585	DONE1A|DONE1B|DONE2
right		TIME1	BARELY	7.18	8	6	64	585	BARELY1
right		TIME1	TO-PRESSURE	6.45	8	7	106	585	TO-PRESSURE1
right		TIME1	BETWEEN	5.67	7	4	159	585	BETWEEN1B|BETWEEN1B-$SAM|BETWEEN1C|BETWEEN3
right		TIME1	TO-DEVELOP	4.08	5	4	343	585	TO-DEVELOP1A|TO-DEVELOP1B|TO-DEVELOP2
right		TIME1	FAST	3.69	7	7	629	585	FAST1A|FAST1B|FAST2|FAST3-$SAM|FAST3A|FAST6
right		TIME1	WHATEVER	3.20	5	4	631	585	WHATEVER3
right		TIME1	FOR	2.91	12	10	1850	585	FOR1
right		TIME1	$NUM-CLOCK	1.73	6	5	2101	585	$NUM-CLOCK1A\|$NUM-CLOCK1B|$NUM-CLOCK1D
right		TIME1	TO-WORK	1.68	6	5	2178	585	TO-WORK1|TO-WORK2
right		TIME1	PRESENT-OR-HERE	1.57	6	5	2337	585	PRESENT-OR-HERE1|PRESENT-OR-HERE1-$SAM
right		TIME1	MUST	1.38	6	6	2674	585	MUST1

Frequent left and right neighbours for TIME1

Using Collocation Patterns for WSD

- Use gloss names as **rough indication of meaning**
- Collapse phonological and lexical neighbour variants into same group
- Use statistics (PMI) with frequency threshold: ≥ 5 co-occurrences
- Frequent neighbours indicate typical semantic contexts of use

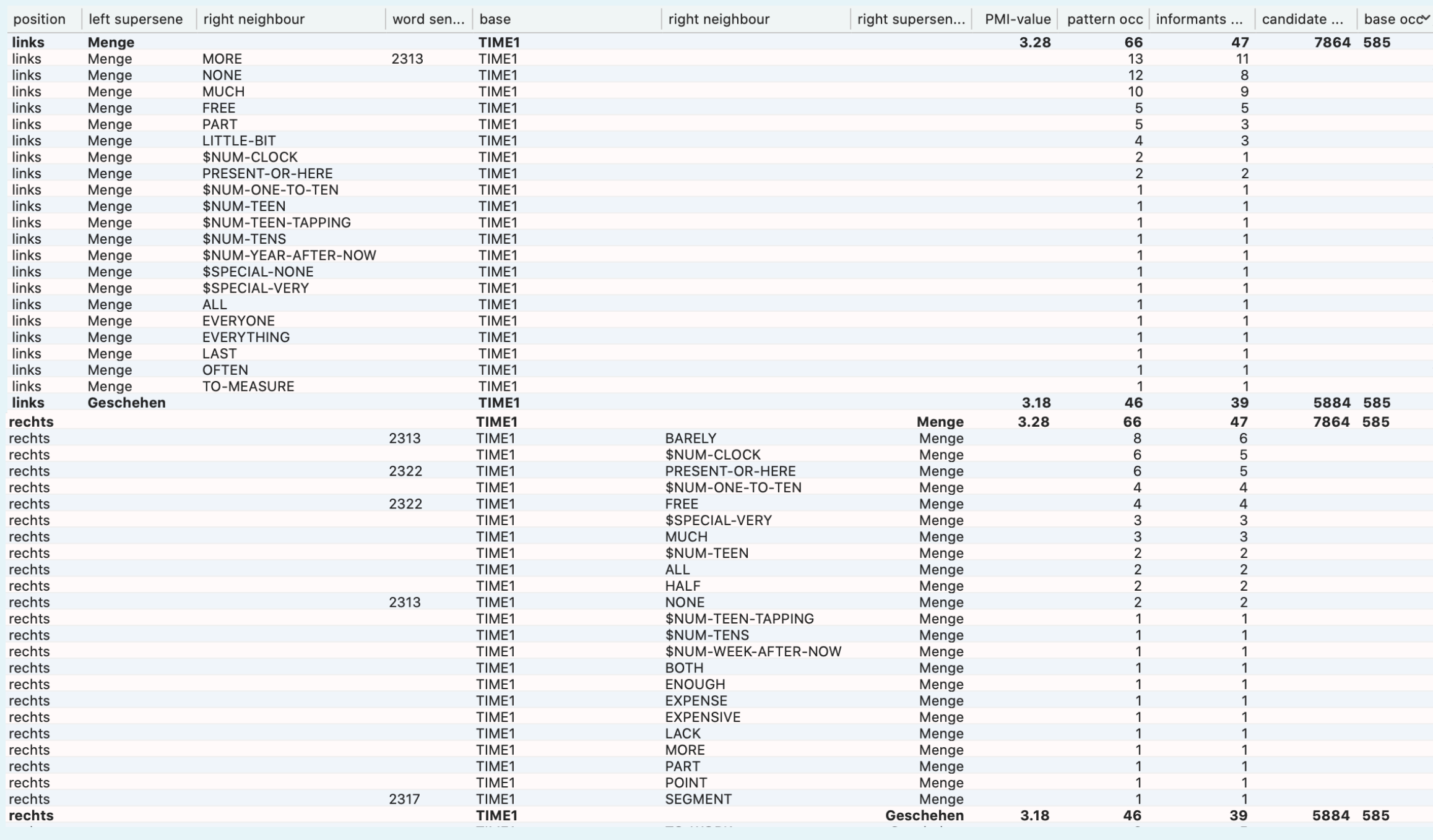
Caveats:

- Glosses are neither translations nor sense tagging ➜ not precise
- Sometimes misleading polysemous/homonymic gloss names (spoken language interference, gloss conventions)

How to cluster by roles?

Idea: No theoretical foundation for syntax parse, so let's look look at semantic groups instead: **Supersenses** (coarse semantic categories) taken from a WordNet.

Supersense Collocations

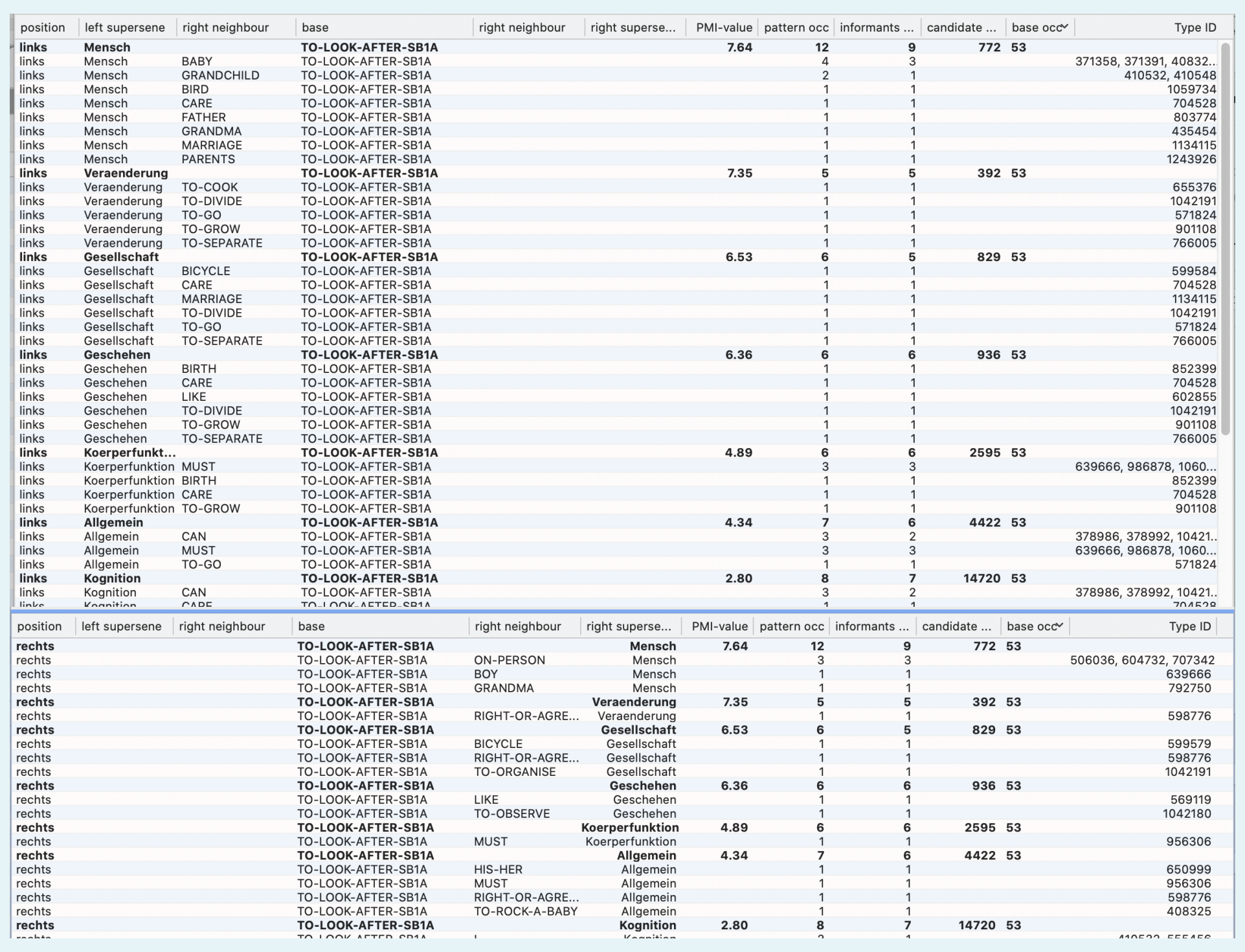


Supersense (GermaNet)	Supersense (Princeton WordNet)	Explanation (Princeton WordNet)
…		
Kommunikation	communication	*nouns denoting communicative processes and contents; verbs of telling, asking, ordering, singing*
Menge	quantity	*nouns denoting quantities and units of measure*
Mensch	person	*nouns denoting people*
Ort	location	*nouns denoting spatial position*
…		

Frequent Supersense Neighbours for TIME1 – Menge (amount)

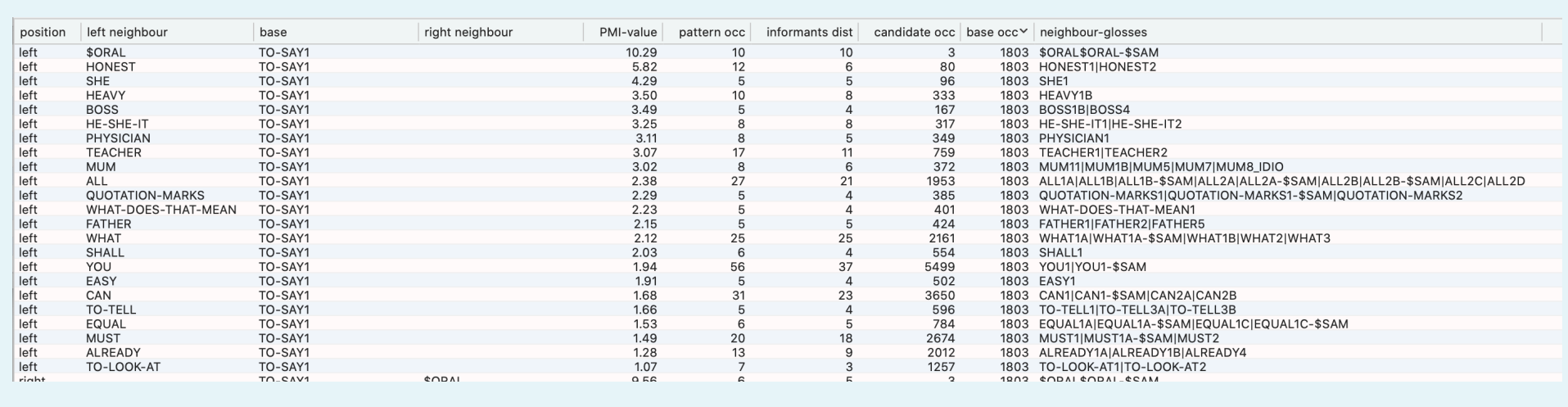
Supersense Collocations add to SL Lexicographers' Toolbox

Who is looking after whom?

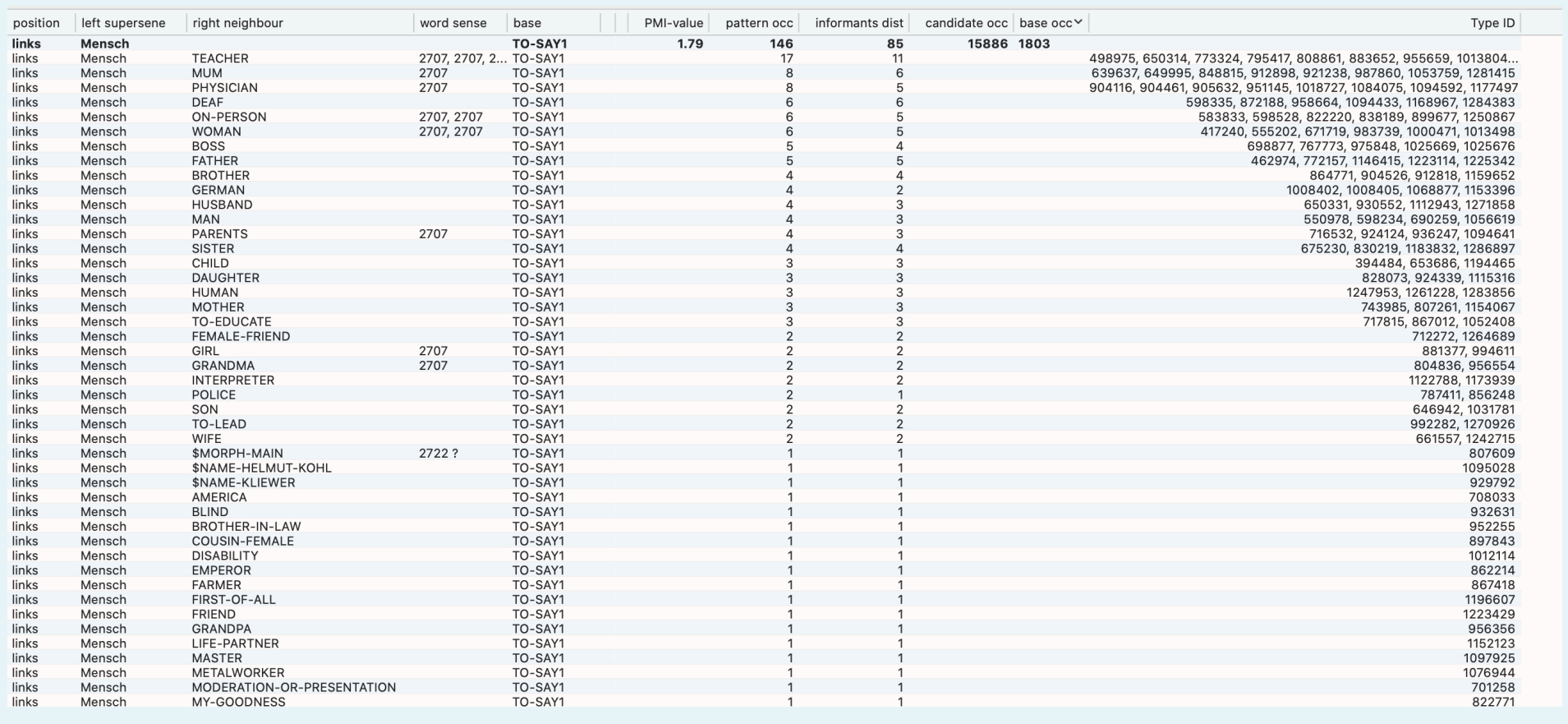


Frequent left and right supersense neighbours of TO-LOOK-AFTER-SB1A

Who says?



Frequent left neighbours of TO-SAY1



Frequent left supersense neighbours of TO-SAY1

In cases where gloss name collocation view comes up empty because of its frequency threshold, the supersense collocation view may still reveal useful patterns.

Signs denoting persons are frequent left neighbours of the sign TO-SAY1. In supersense collocations even members of a semantic group that only appear once in the data are considered in the detection of the larger semantic preference pattern.

DGS Corpus (www.dgs-korpus.de) accessed via iLex using SQL query integrated into iLex user interface (2022-06-17)

Getting from Signs to WordNet

- Gloss name ➜ German lemma ➜ word senses ➜ supersenses
- Supersense collocations cluster glosses by semantic category

Caveats:

- String-based matching to senses **very noisy**
- Reinforces spoken language word sense assumptions

Conclusion

- First step towards automatic support of SL lexicographic work
- Matching to spoken language is a crutch, output will be noisy.
- Observations must always be checked against original video data.

Poster presented at the 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources. LREC 2022, Marseille, France. June 25th, 2022. **Paper published** in Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. LREC 2020. May 16th, 2020.

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.