How Web Communities Analyze Human Language

Word Senses in Wiktionary

Christian M. Meyer and Iryna Gurevych

Motivation

With the rise of the **Socio-Semantic Web**, Communities of Web users have started to create **new resources of human knowledge**, like Wikipedia. A crucial property of such resources is the **collaborative construction** process that enables **fundamentally new insights** in many areas of research unthought of before and has the potential to radically influence previously existing research paradigms.

Research Questions

Word Sense is a fundamental notion in human language analysis that has been a subject of intensive studies for centuries. Traditionally, word senses are defined by **expert** linguists or lexicographers. Emerging Web resources allow to define word senses **collaboratively** that are yet poorly investigated and have the potential to shed some light on the definition of word senses itself, which is still an open research question.

Expert defined Word Senses

Collaboratively defined Word Senses

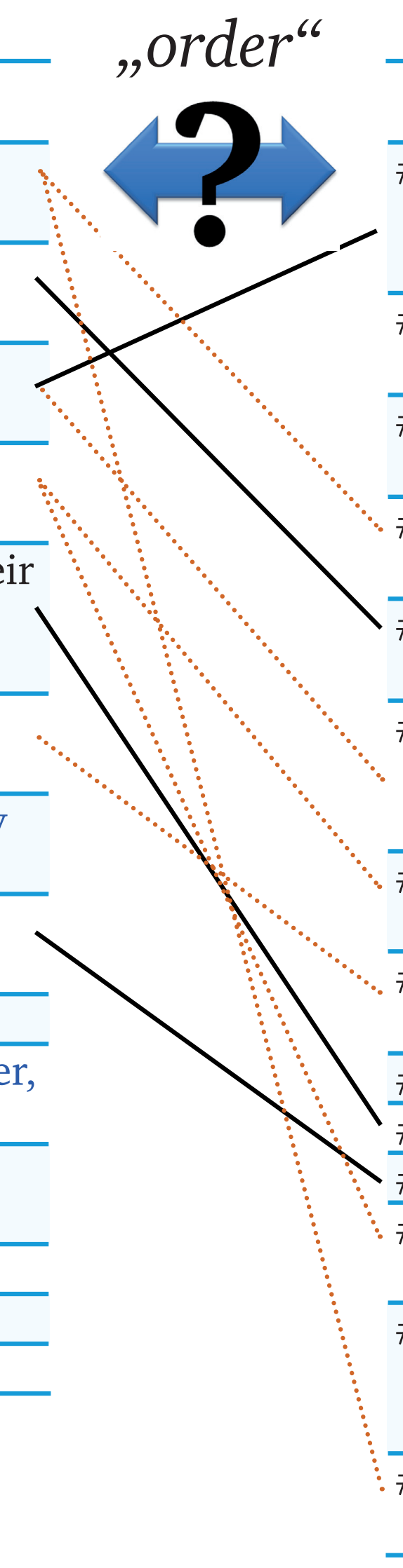
- Do they encode the same information?
- How are they defined?
- What is the coverage of word senses?
- Is the definition granularity similar?

Comparison of Manually Aligned Word Senses

Wiktionary

#1	(uncountable) Arrangement, disposition, sequence. [variant #14;#4]
#2	(uncountable) The state of being well arranged. [equals #5]
#3	(countable) A command. [equals #1]
#4	(countable) A request for some product or service [variant #12;#7]
#5	(countable) A group of religious adherents, especially monks or nuns, set apart within their religion by adherence to a particular rule or set of principles; as, the Jesuit Order. [equals #10]
#6	(countable) A society of knights; as, the Order of the Garter, the Order of the Bath. [variant #8]
#7	(countable) A decoration, awarded by a government, a dynastic house, or a religious body to an individual, usually for distinguished service to a nation or to humanity. [new]
#8	(countable/biology/taxonomy) A rank in the classification of organisms, below class and above family; a taxon at that rank [equals #11]
#9	(cricket) The sequence in which a side's batsmen bat; the batting order. [new]
#10	(electronics) a power of polynomial function in an electronic circuit's block, such as a filter, an amplifier, etc. [new]
#11	(chemistry) The overall power of the rate law of a chemical reaction, expressed as a polynomial function of concentrations of reactants and products. [new]
#12	(mathematics) The cardinality, or number of elements in a set or related structure. [new]
#13	(graph theory) The number of vertices in a graph [new]
#14	(order theory) A partially ordered set. [new]

[equals] equivalent word sense in other resource
[new] Not represented in other resource
[variant] variant of a word sense in other resource; more specific or generic



WordNet

#1	(often plural) a command given by a superior (e.g., a military or law enforcement officer) that must be obeyed; *„the British ships dropped anchor and waited for orders from London“* [equals #3]
#2	a degree in a continuum of size or quantity; *„it was on the order of a mile“; „an explosion of a low order of magnitude“* [new]
#3	established customary state (especially of society); *„order ruled in the streets“; „law and order“* [new]
#4	logical or comprehensible arrangement of separate elements; *„we shall consider these questions in the inverse order of their presentation“* [variant #1]
#5	a condition of regular or proper arrangement; *„he put his desk in order“; „the machine is now in working order“* [equals #2]
#6	a legally binding command or decision entered on the court record (as if issued by a court or judge); *„a friend in New Mexico said that the order caused no trouble out there“* [variant #3]
#7	a commercial document used to request someone to supply something in return for payment and providing specifications and quantities [variant #4]
#8	a formal association of people with similar interests; *„men from the fraternal order will staff the soup kitchen today“* [variant #6]
#9	a body of rules followed by an assembly [new]
#10	a group of person living under a religious rule [equals #5]
#11	(biology) taxonomic group containing one or more families [equals #8]
#12	a request for something to be made, supplied, or served; *„I gave the waiter my order“* [variant #4]
#13	(architecture) one of original three styles of Greek architecture distinguished by the type of column and entablature used or a style developed from the original three by the Romans [new]
#14	the act of putting things in a sequential arrangement; *„there were mistakes in the ordering of items on the list“* [variant #1]

Resource Coverage

	Wiktionary	WordNet	Overlap
Number of Lexemes:	323,264	156,584	75,750
...only Nouns:	200,217	119,034	48,681
...only Verbs:	55,483	11,531	8,967
...only Adjectives:	46,636	21,538	14,484
...only Adverbs:	9,660	4,481	3,618
...other POS:	11,268	0	0
Inflected Forms:	102,476	–	–
Latin Terms:	–	7,082	–
Abbreviations:	7,051	1,014	624
Proper Names:	13,494	14,236	3,110
Neologisms (1,192):	156	21	18

Sense Distribution

Senses	Nouns		Verbs		Adjectives	
	WKT	WN	WKT	WN	WKT	WN
1	86%	87%	87%	55%	82%	77%
2	9%	8%	7%	22%	13%	15%
3	3%	2%	3%	10%	3%	5%
4	1%	1%	1%	5%	1%	2%
≥ 5	1%	2%	2%	8%	1%	1%
avg	1.26	1.23	1.26	2.17	1.27	1.39
max	57	33	58	59	22	27

Polysemic Difference: difference in the number of encoded senses; 60% have $\Delta = 0$; 95% have $\Delta \leq 2$.

Sense Comparison

Dimension	Shared	New		Variant	
Word Frequency		WKT	WN	WKT	WN
Seldomly used	4	6	0	1	1
Medium usage	9	2	3	3	0
Commonly used	16	7	5	16	9
Polysemic Difference					
Low ($\Delta = 0$)	6	7	4	3	6
High ($\Delta \geq 30$)	8	32	1	5	0
Part of Speech					
Nouns	30	46	7	14	7
Verbs	9	6	4	8	7
Adjectives	4	2	2	6	5

Conclusions

Resource Coverage

- Overlap of the resources at term level is surprisingly low.
- The missing terms induce also many missing word senses.

Word Sense Distribution

- Word sense distribution is mostly similar.
- On average, more word senses for verbs in WordNet.
- Higher maximum number of word senses in Wiktionary.
- 60% of the shared lexemes encode the same number of word senses; 95% have a polysemic difference of less than 3.

Word Sense Comparison

- Wiktionary encodes word senses for seldomly used terms.
- Better coverage for slang-related and domain-specific word senses.
- WordNet shows a better coverage of senses from social sciences and humanities, while Wiktionary has a better coverage of senses from natural sciences, sports, and military.
- Good agreement of senses for words with a medium language frequency.
- Many Wiktionary word senses for commonly used words are missing from WordNet.

We argue that collaborative word sense inventories have a great potential and aim to combine expert and collaborative resources in the future.