



A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment

Sina Ahmadi (sina.ahmadi@insight-centre.org)

Under the supervision of Dr. John P. McCrae

January 20, 2020



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Outline

- Context
- Objectives
- Methodology
- Evaluation
- Conclusion and future steps

A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment

Sina Ahmadi¹, John P. McCrae¹,
 Thierry Declercq^{2,12}, Sanni Nimb⁷, Thomas Troelsgård⁷, Tanja Wissik⁸, Monica Monachini³,
 Bolette S. Pedersen⁴, Irene Pisan¹, Andrea Bellandi³, Fahad Khan⁴, Simon Krek⁵,
 Veronika Lipp⁶, Tamás Váradi⁶, László Simon⁶, András Györfly⁶,
 Carole Tiberius⁹, Tanneke Schoonheim⁹, Ilan Kernerman¹⁰, Raya Abu Ahmad¹⁰,
 Dorielle Lonke¹⁰, Kira Kovalenko¹¹, Oksana Dereza¹, Theodor Fransen¹
¹Insight Centre for Data Analytics, National University of Ireland, Galway
²Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Vienna, Austria
³Istituto di Linguistica Computazionale "A. Zampolli-CNR", Pisa, Italy
⁴Università di Pisa, Italy
⁵Jozef Stefan Institute, Ljubljana, Slovenia
⁶Research Institute for Linguistics, Budapest, Hungary
⁷Society for Danish Language and Literature (DSL), Copenhagen, Denmark
⁸Centre for Language Technology, University of Copenhagen, Denmark
⁹Dutch Language Institute, Leiden, The Netherlands
¹⁰K Dictionaries, Tel Aviv, Israel
¹¹Institute for Linguistic Studies of the Russian Academy of Sciences, St. Petersburg, Russia
¹²DFKI GmbH, Multilinguality and Language Technology, Germany
¹{sina.ahmadi, john.mccrae, oksana.dereza, theodor.fransen}@insight-centre.org, ²Tanja.Wissik@oeaw.ac.at
⁶{veronika.lipp, varadi.tamas, simon.laszlo, gyorffy.andras}@nyud.hu, ⁷sn@dsl.dk, tt@dsl.dk
⁸{bspedersen, saolsen}@hum.ku.dk, ⁹{carole.tiberius, tanneke.schoonheim}@ivdnt.org ¹²declercq@dfki.de

Abstract

Aligning senses across resources and languages is a challenging task with beneficial applications in the field of natural language processing and electronic lexicography. In this paper, we describe our efforts in manually aligning monolingual dictionaries. The alignment is carried out at sense-level for various resources in over 11 languages (with 4 more in development). Moreover, senses are annotated with possible semantic relationships such as broadness, narrowness, relatedness, and equivalence. In comparison to previous datasets for this task, this dataset covers a wide range of languages and resources and focuses on the more challenging task of linking general-purpose language. We believe that our data will pave the way for further advances in alignment and evaluation of word senses by creating new solutions, particularly those notoriously requiring data such as neural networks.

Keywords: lexical semantic resources, sense alignment, lexicography, language resource

1. Introduction

Dictionaries form an important foundations of numerous natural language processing (NLP) tasks, including word sense disambiguation, machine translation, question answering and automatic summarization. However, the task of combining dictionaries from different sources is difficult, especially for the case of mapping the senses of entries, which often differ significantly in granularity and coverage. Approaches so far have mostly only been evaluated on named entities and quite specific domain language. In order to support a shared task at the GLOBALEX workshop, we have developed a new baseline that covers 11 languages and will provide a new baseline for the task of monolingual word sense alignment.

Different dictionaries and related resources such as word-nets and encyclopedias have significant differences in structure and heterogeneity in content, which makes aligning information across resources and languages a challenging task. Word sense alignment (WSA) is a more specific task of linking dictionary content at sense level which has been proved to be beneficial in various NLP tasks, such as word-sense disambiguation (Navigli and Ponzetto, 2012), seman-

tic role labeling (Palmer, 2009) and information extraction (Moro et al., 2013). Moreover, combining LSRs can enhance domain coverage in terms of the number of lexical items and types of lexical-semantic information (Shi and Mihalcea, 2005; Ponzetto and Navigli, 2010; Gurevych et al., 2012).

Given the current progress of artificial intelligence and the usage of data to train neural networks, annotated data with specific features play a crucial role to tackle data-driven challenges, particularly in NLP. In recent years, a few efforts have been made to create *gold-standard* dataset, i.e., a dataset of instances used for learning and fitting parameters, for aligning senses across monolingual resources including collaboratively-curated ones such as Wikipedia¹ and expert-made ones such as WordNet. However, the previous work is limited to a handful of languages and much of it is not on the core vocabulary of the language, but instead on named entities and specialist terminology. Moreover, despite the huge endeavour of lexicographers to compile dictionaries, proper lexicographic data are rarely openly accessible to researchers. In addition many of the resources are

¹<https://www.wikipedia.org>

Context

Lexical-Semantic Resources (LSRs)

WordNet Search - 3.1
- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options: (Select option to change)

Key: "S:" = Show Synset (semantic) relations, "W:" = Show WordNet gloss
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) necktie, tie** (neckwear consisting of a long worn (mostly by men) under a collar and tied in front of the mirror tightening his necktie"; "he in front of the mirror tightening his necktie"; "he
- S: (n) affiliation, association, tie, tie-up** (a social valuable financial affiliation"; "he was sorry he had other members of the team"; "many close associ
- S: (n) tie** (equality of score in a contest)
- S: (n) tie, tie beam** (a horizontal beam used to pr members from spreading apart or separating) "h with a tie beam"
- S: (n) link, linkup, tie, tie-in** (a fastener that serv walls are held together with metal links placed in construction"
- S: (n) draw, standoff, tie** (the finish of a contest i the winner is undecided) "the game ended in a d wins, 6 losses and a tie"
- S: (n) tie** ((music) a slur over two notes of the sar note is to be sustained for their combined time v
- S: (n) tie, railroad tie, crosstie, sleeper** (one of th the rails on a railway track) "the British call a rail
- S: (n) tie** (a cord (or string or ribbon or wire etc.) tied) "he needed a tie for the packages"

Pawn (chess)

Article Talk

White pawn Black pawn

The **pawn** (♙, ♜) is the most numerous *piece* in the game of **chess**, and in most circumstances, also the weakest. It historically represents *infantry*, or more particularly, armed *peasants* or *pikemen*.^[1] Each player begins a game with eight pawns, one on each square of the *rank* immediately in front of the other pieces. (The white pawns start on a2, b2, c2, d2, e2, f2, g2, h2; the black pawns start on a7, b7, c7, d7, e7, f7, g7, h7.)

shake

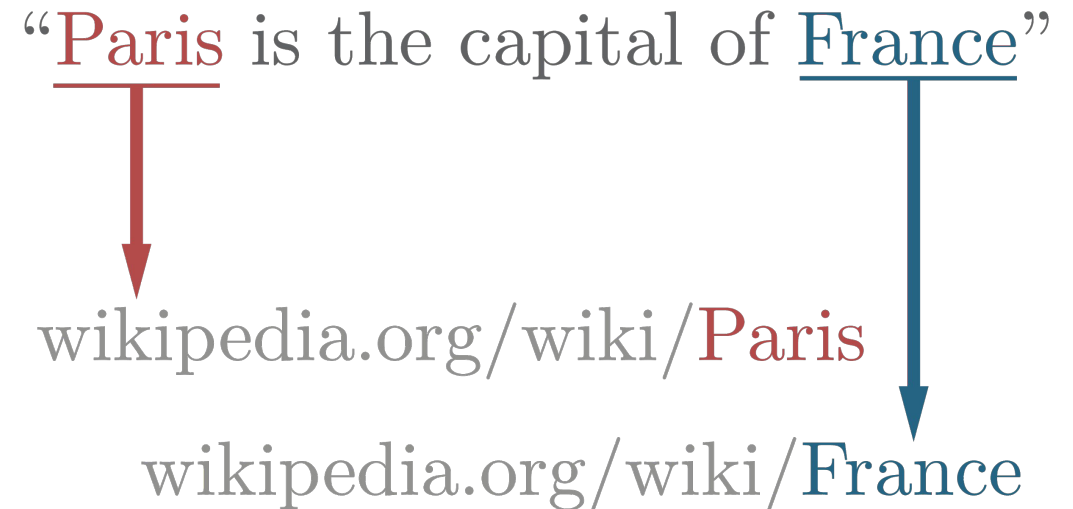
- Externally controlled motion** (Causative/Inchoative/Resultative)
- Funnel** (Causative +CH-LOC)
- Body-Internal state** (Causative/Inchoative)
- Crane** (Causative/Cognate object +COMM)
- ABSTRACT**
- Psych/Amuse** (Causative/no middle)
- Idioms**
 - shake down
 - shake the dust off
 - shake a leg etc.,
- OUT (PATH PREP)**
- BODY-PART**
 - head
 - hands

VERBNET

Context

LSRs: a few applications

- Entity linking
- Semantic role labeling
- Word-sense disambiguation

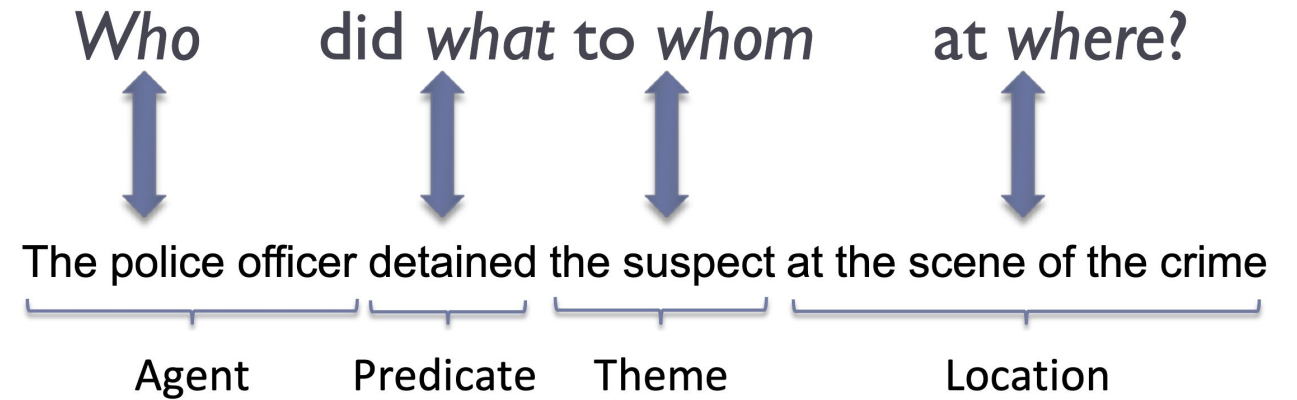


Source: https://en.wikipedia.org/wiki/Entity_linking

Context

LSRs: a few applications

- Entity linking
- **Semantic role labeling**
- Word-sense disambiguation



Source: https://web.stanford.edu/~jurafsky/slp3/slides/22_SRL.pdf

Context

LSRs: a few applications

- Entity linking
- Semantic role labeling
- **Word-sense disambiguation**

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

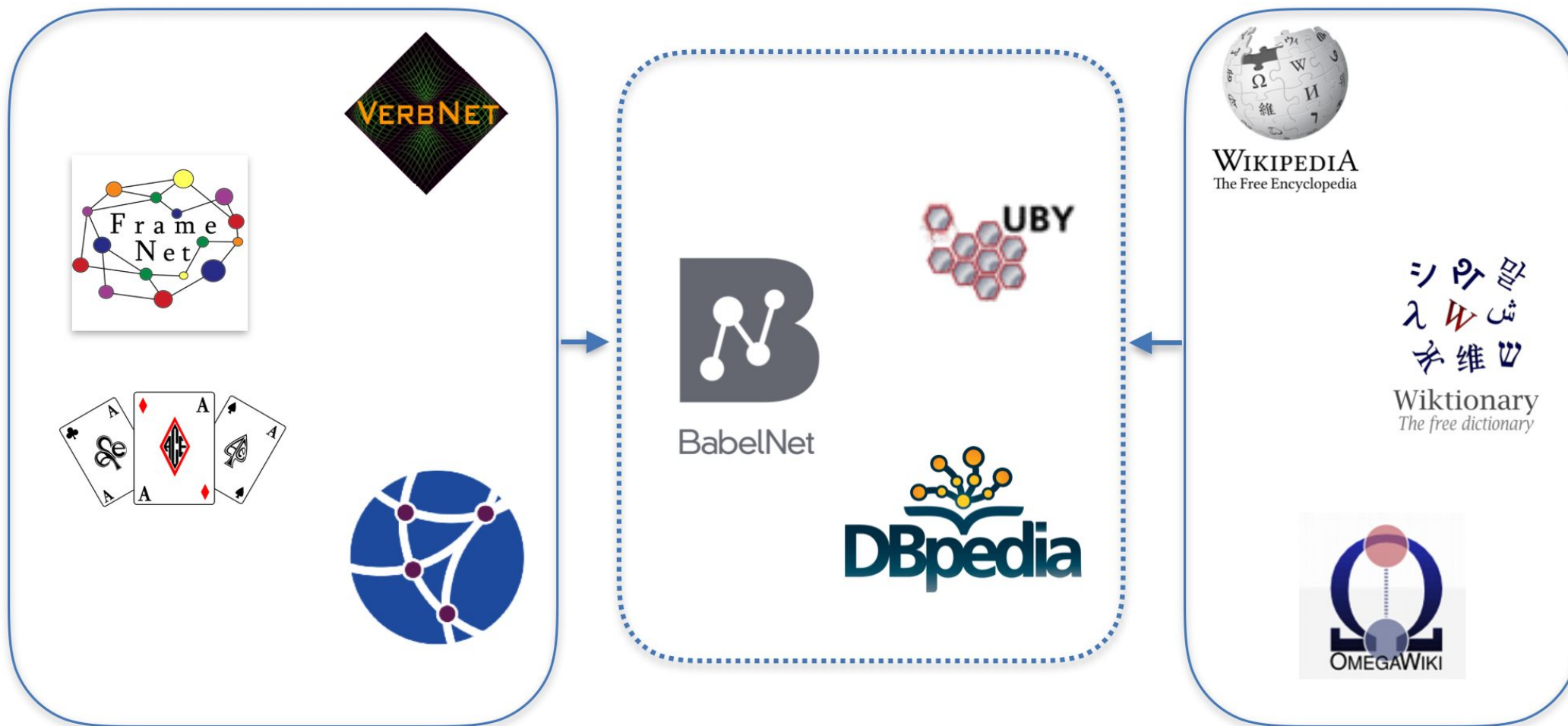
Display options for sense: (gloss) "an example sentence"

Noun

- [S:](#) (n) [necktie](#), **tie** (neckwear consisting of a long narrow piece of material worn (mostly by men) under a collar and tied in knot at the front) *"he stood in front of the mirror tightening his necktie"; "he wore a vest and tie"*
- [S:](#) (n) [affiliation](#), [association](#), **tie**, [tie-up](#) (a social or business relationship) *"a valuable financial affiliation"; "he was sorry he had to sever his ties with other members of the team"; "many close associations with England"*
- [S:](#) (n) **tie** (equality of score in a contest)
- [S:](#) (n) **tie**, [tie beam](#) (a horizontal beam used to prevent two other structural members from spreading apart or separating) *"he nailed the rafters together with a tie beam"*
- [S:](#) (n) [link](#), [linkup](#), **tie**, [tie-in](#) (a fastener that serves to join or connect) *"the walls are held together with metal links placed in the wet mortar during construction"*
- [S:](#) (n) [draw](#), [standoff](#), **tie** (the finish of a contest in which the score is tied and the winner is undecided) *"the game ended in a draw"; "their record was 3 wins, 6 losses and a tie"*
- [S:](#) (n) **tie** ((music) a slur over two notes of the same pitch; indicates that the

Context

Resource alignment



Expert-made

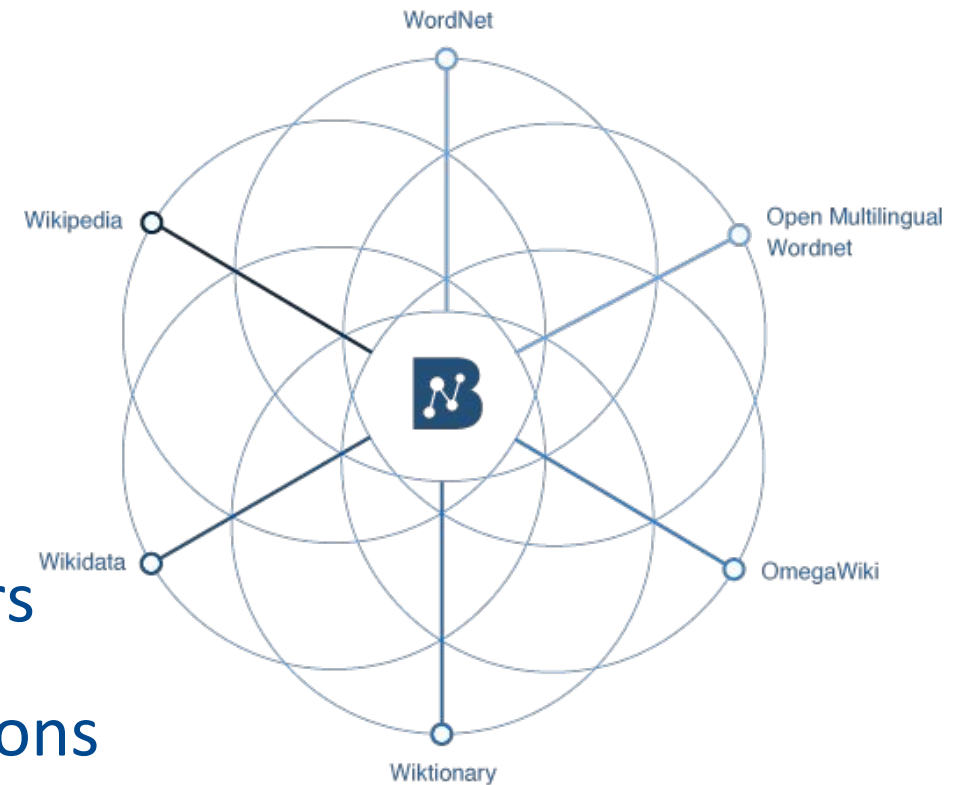
Collaboratively-curated



Context

Why linking resources?

- To improve word and concept coverage
 - e.g., named entities, new senses
- To improve domain coverage
- To improve multilingualism
- Creating resources for new language pairs
- To combine expert-made semantic relations
 - e.g., Hypernymy, meronymy, etc.



BabelNet (<https://babelnet.org/>)

Word sense alignment (WSA)

→ linking lexical content at sense level

lead² ●●○ **S3** **W2** noun 🔊 🔊

1 → the lead

2 **[singular]** the amount or distance by which one competitor is ahead of another

🔊 The Chicago Bulls **had a narrow lead** (=were winning by a small number of points).

lead over

🔊 The Socialists now have a commanding lead over their opponents.

3 **[singular]** if someone follows someone else's lead, they do the same as the other person has done

🔊 Other countries are likely to **follow** the U.S.'s **lead**.

🔊 The Government should **give** industry a **lead** in tackling racism (=show what other people should do).

🔊 The black population in the 1960s **looked to** Ali **for a lead** (=looked to him to show them what they should do).

4 → take the lead (in doing something)

5 **[countable]** a piece of information that may help you to solve a crime or mystery **SYN** clue

🔊 The police have checked out dozens of leads, but have yet to find the killer.

6 **[countable]** the main acting part in a play, film etc, or the main actor
play the lead/the lead role

🔊 He will play the lead role in 'Hamlet'.

🔊 Powers was **cast in the lead role** (=he was chosen to play it).

the male/female lead

🔊 They were having trouble casting the female lead.

🔊 the film's **romantic lead**

7 → lead singer/guitarist etc



lead lead Video English: lead¹ English: lead² American: lead¹ American: lead² Specialist English: lea ▶

16. countable noun

A **lead** is a piece of information or an idea which may help people to discover the facts in a situation where many facts are not known, for example in the investigation of a crime or in a scientific experiment.

The inquiry team is also following up possible leads after receiving 400 calls from the public.

Synonyms: clue, tip, suggestion, trace [More Synonyms of lead](#)

17. countable noun

The lead in a play, film, or show is the most important part in it. The person who plays this part can also be called the **lead**.

Performers from the Bolshoi Ballet dance the leads.

Both the leads in the play are impressive.

Synonyms: leading role, principal, protagonist, title role [More Synonyms of lead](#)

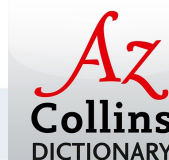
18. countable noun

A dog's **lead** is a long, thin chain or piece of leather which you attach to the dog's collar so that you can control the dog.

[mainly British]

An older man came out with a little dog on a lead.

REGIONAL NOTE:
 in AM, use **leash**



Context

Sense alignment is challenging

- Differences in structure
 - Where a sense is a “sense” and where a “subsense”?
- Differences in content
 - Lexical choice:
 - **Alcohol:** *vandklar vædske* (water-clear liquid) vs. *farveløs* (colorless) in Danish
 - Orthographic variations: *kjøn* vs. *køn*, *paa* vs. *på*
 - Description methods

Context

LSRs: the current state

- A significant body of research in aligning English resources including linking the Princeton WordNet with
 - Wikipedia (McCrae, 2018)
 - Wiktionary (Meyer and Gurevych, 2011)
 - the Oxford Dictionary of English (Navigli, 2006)
- A fewer number of manually aligned monolingual resources in other languages including linking:
 - the GermaNet–the German Wordnet with
 - the German Wikipedia (Henrich et al., 2012)
 - the German Wiktionary (Henrich et al., 2011)
- Various solutions have been proposed to semi-automatically link and merge existing LSRs

Objectives

- To address some of the current main limitations in WSA:
 - Multilingualism
 - Monolingual resources
 - Gold-standard datasets
 - Semantic relationship annotation

→ **manually-annotated monolingual resources for the task of WSA for 11 languages***

*Ultimately, 15 languages

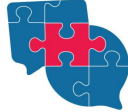

Context

Data to nourish

Annotated data play a crucial role to
tackle data-driven challenges
particularly given the current progress
of artificial intelligence



Methodology

- Resource selection → higher priority given to dictionaries
- Preprocessing including spelling variation normalization
- Organizing resources in one of the following formats:
 - a. OntoLex- Lemon (Cimiano et al., 2016)
 - b. TEI-Lex0 (Romary and Tasovac, 2018)
 - c. or a simple TSV (tab-separated values) format
- Conversion of data into dynamic spreadsheets 
- Manual annotation of corresponding resources and languages
- Conversion to the final structure of the dataset 

Data Selection

The selection of the initial set of lemmas and senses to be aligned is guided by the following criteria:

- lemmas should represent all open class words
- lemmas should represent different degrees of polysemy
- lemmas in the two resources have the same part-of-speech tags

→ Spelling variations are normalized to an identical variation.

Data selection

The Princeton WordNet

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Verb

- **S: (v) clog**, [choke off](#), [clog up](#), [back up](#), [congest](#), [choke](#), [foul](#) (become or cause to become obstructed) *"The leaves clog our drains in the Fall"; "The water pipe is backed up"*
- **S: (v) clog** (dance a clog dance)
- **S: (v) clog** (impede the motion of, as with a chain or a burden) *"horses were clogged until they were tamed"*
- **S: (v) clog**, [constipate](#) (impede with a clog or as if with a clog) *"The market is being clogged by these operations"; "My mind is constipated today"*
- **S: (v) clog**, [clot](#) (coalesce or unite in a mass) *"Blood clots"*
- **S: (v) clog**, [overload](#) (fill to excess so that function is impaired) *"Fear clogged her mind"; "The story was clogged with too many details"*

Webster's dictionary of 1913

Clog

Clog, v. t. [*imp. & p. p. Clogged* (?); *p. pr. & vb. n. Clogging*.]

1. To encumber or load, especially with something that impedes motion; to hamper.

The winds of birds were *clogged* with ace and snow.
Dryden.

2. To obstruct so as to hinder motion in or through; to choke up; as, to *clog* a tube or a channel.

3. To burden; to trammel; to embarrass; to perplex.

The commodities are *clogged* with impositions.
Addison.

You 'll rue the time
That *clogs* me with this answer.
Shak.

Syn. – Impede; hinder; obstruct; embarrass; burden; restrain; restrict.

Methodology

Conversion to dynamic spreadsheets

Headword (pos)	Sense IDs	WordNet sense	Semantic relation	Sense match	Webster sense	Sense IDs	
allow (verb)							
	allow.v.04	give or assign a resource to a particular person or cause	exact	3.-to sanction	0 -to admit; to concede; to make allowance or abatement.		
	allow.v.06	allow or plan for a certain possibility; concede the truth or validity of something	exact	5.-to own or ;	1.-to praise; to approve of; hence, to sanction.		
	leave.v.06	make a possibility or provide opportunity for; permit to be attainable or cause to remain	broader	7.-to grant lic	2.-to like; to be suited or pleased with.		
	allow.v.03	let have	broader	4.-to grant, g	3.-to sanction; to invest; to intrust.		
	admit.v.05	afford possibility	broader	4.-to grant, g	4.-to grant, give, admit, accord, afford, or yield; to let one have;		
	let.v.01	make it possible through a specific action or lack of action for something to happen	broader	7.-to grant lic	5.-to own or acknowledge; to accept as true; to concede; to accede to an opinion;		
	give_up.v.11	allow the other (baseball) team to score	broader	4.-to grant, g	6.-to grant (something) as a deduction or an addition; esp. to abate or deduct; .		
	allow.v.09	grant as a discount or in exchange	exact	6.-to grant (s	7.-to grant license to; to permit; to consent to; .		
	allow.v.10	allow the presence of or allow (an activity) without opposing or prohibiting	broader	7.-to grant lic			
	permit.v.01	consent to, give permission	broader	7.-to grant lic			

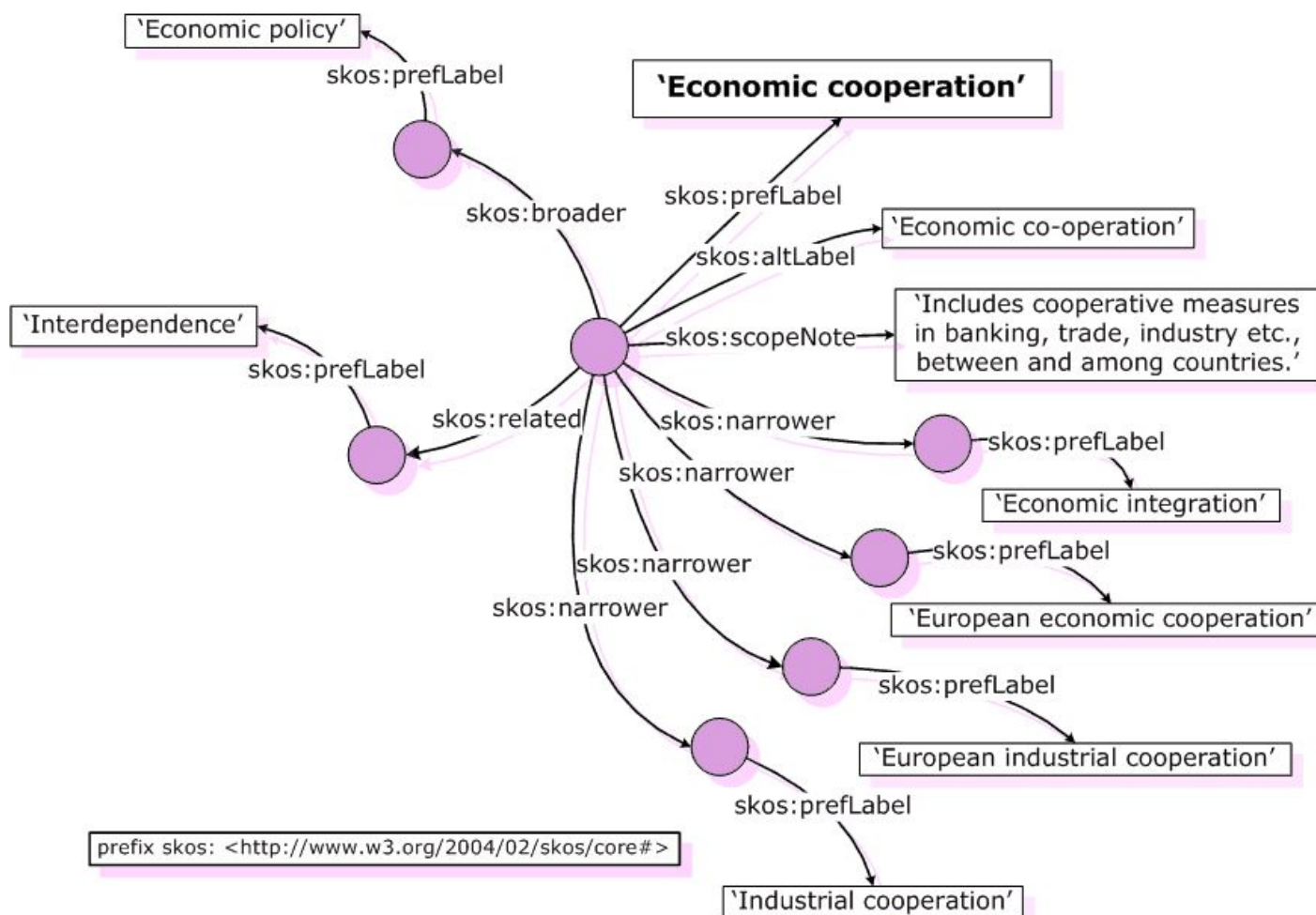
Semantic relationships (based on SKOS*)

- **exact**: The sense are the same, for example the definitions are simply paraphrases
- **broader**: The sense in the first dictionary completely covers the meaning of the sense in the second dictionary and is applicable to further meanings
- **narrower**: The sense in the first dictionary is entirely covered by the sense of the second dictionary, which is applicable to further meanings
- **related**: There are cases when the senses may be equal but the definitions in both dictionaries differ in key aspects
- **none**: There is no match for this sense

* <https://www.w3.org/2004/02/skos/>

Methodology

Semantic relationships: an example



Source: <https://www.w3.org/2004/02/skos/core/guide/2005-10-06/>

Evaluation

1. Statistics of the datasets
2. Frequency of the number of senses
3. Sense granularity
4. Sense alignments
5. Inter-annotator agreement

Evaluation

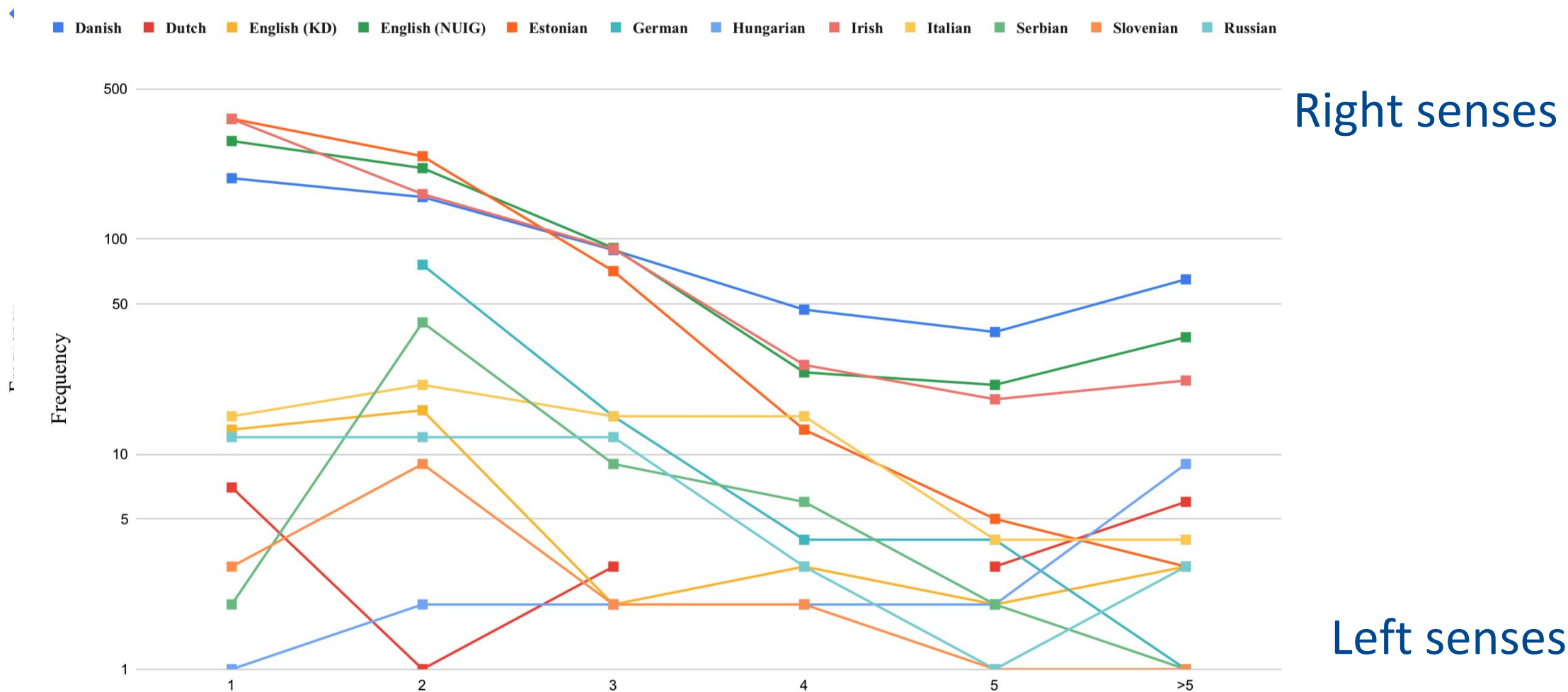
1. Statistics

Number of senses and
number of tokens (in
parentheses)

Language	Resource	Nouns	Verbs	Adjectives	Adverbs	Other	All
Danish	<i>Ordbog over det danske Sprog</i>	2176 (282040)	983 (119163)	436 (60599)	0 (0)	0 (0)	3595 (461802)
	<i>Den Danske Ordbog</i>	1036 (12326)	383 (4045)	248 (2228)	0 (0)	0 (0)	1667 (18599)
Dutch	<i>Woordenboek der Nederlandsche Taal</i>	124 (2861)	34 (507)	18 (208)	0 (0)	0 (0)	176 (3576)
	<i>Algemene Nederlandse Woordenboek</i>	47 (848)	22 (324)	12 (173)	0 (0)	0 (0)	81 (1345)
English (KD)	Global	16 (125)	14 (120)	36 (273)	34 (162)	25 (147)	125 (827)
	Password	16 (138)	9 (47)	31 (189)	22 (129)	24 (107)	102 (610)
English (NUIG)	Webster	1131 (11606)	741 (4622)	373 (2585)	45 (269)	0 (0)	2290 (19082)
	Princeton WordNet	730 (12166)	496 (6980)	249 (2892)	24 (207)	0 (0)	1499 (22245)
Estonian	Dictionary of Estonian (EKS)	543 (4012)	273 (1598)	151 (747)	98 (451)	78 (370)	1143 (7178)
	Estonian Basic Dictionary (PSV)	543 (4492)	273 (1983)	151 (1097)	98 (596)	79 (468)	1144 (8636)
German	German OmegaWiki	419 (2926)	0 (0)	0 (0)	0 (0)	0 (0)	419 (2926)
	Wiktionary	247 (3013)	0 (0)	0 (0)	0 (0)	0 (0)	247 (3013)
Hungarian	Comprehensive						132 (1379)
	Explanatory						103 (1150)
Irish	<i>An Foclóir Beag</i>	891 (8053)	11 (95)	55 (267)	10 (56)	36 (171)	1003 (8642)
	Wiktionary	1209 (6696)	8 (45)	61 (181)	10 (41)	36 (109)	1324 (7072)
Italian	ItalWordNet	128 (933)	142 (1054)	0 (0)	0 (0)	0 (0)	270 (1987)
	SIMPLE	105 (789)	103 (612)	0 (0)	0 (0)	0 (0)	208 (1401)
Serbian	Serbian WordNet	188 (1459)	183 (1208)	15 (131)	0 (0)	0 (0)	386 (2798)
	Dictionary of Serbo-Croatian Literary Language	101 (857)	48 (248)	4 (34)	0 (0)	0 (0)	153 (1139)
Slovenian	Slovene WordNet	49 (117)	33 (94)	14 (46)	4 (22)	0 (0)	100 (279)
	Slovene Lexical Database	30 (271)	7 (36)	9 (61)	2 (9)	0 (0)	48 (377)
Russian	Ozhegov-Shvedova	33 (269)	16 (90)	17 (79)	4 (14)	30 (263)	100 (715)
	Dictionary of the Russian Language (MAS)	38 (298)	21 (143)	29 (162)	4 (12)	55 (823)	147 (1438)

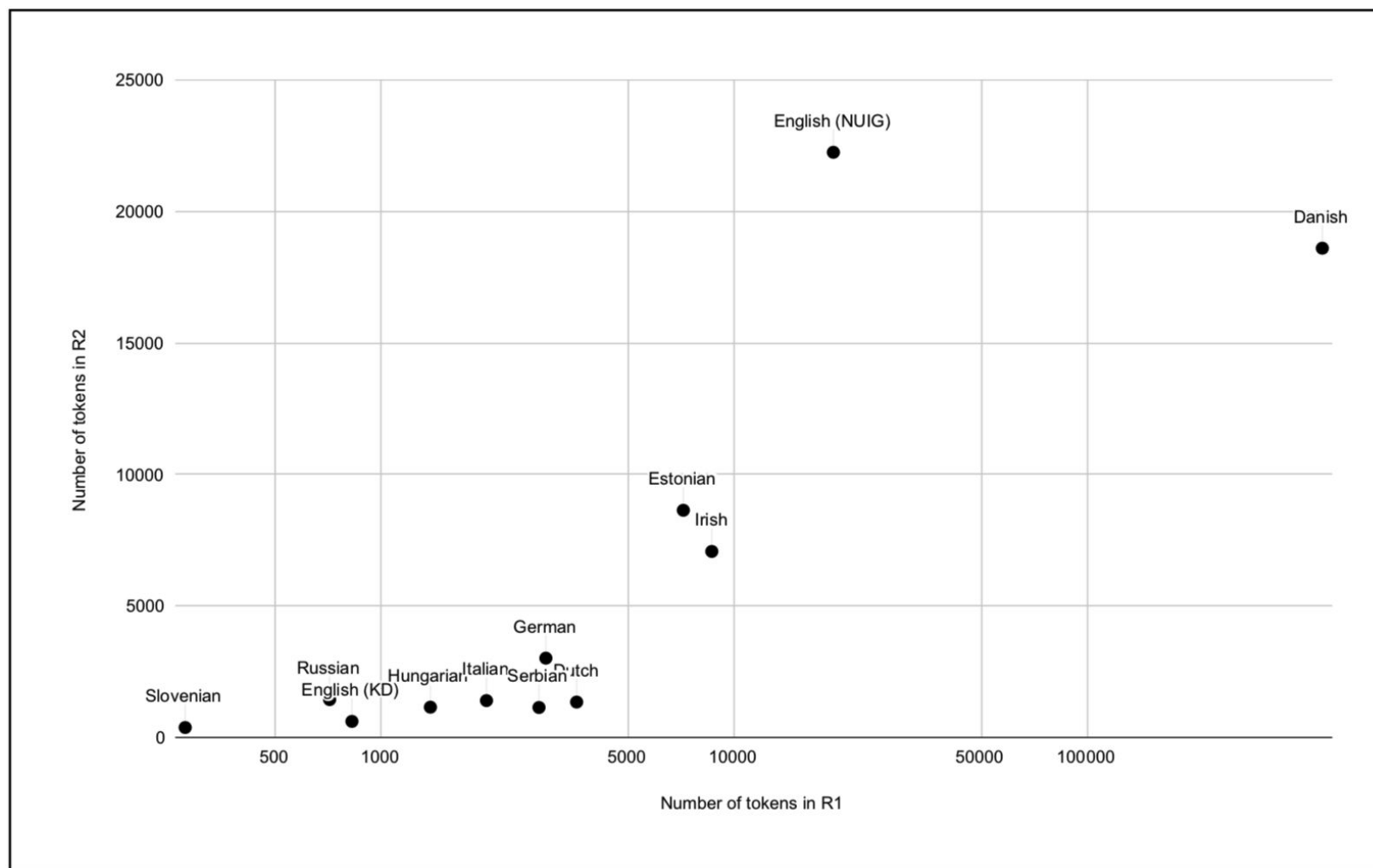
Datasets description

2. Frequency of the number of senses



Evaluation

3. Sense granularity: Correlation between number of tokens



Evaluation

4. Sense alignments

Language	Semantic relationship					k_1	k_2	k
	exact	narrower	broader	related	all			
Danish	1103	316	189	36	1644	0.46	0.99	0.62
Dutch	46	1	0	0	47	0.27	0.58	0.37
English (KD)	55	24	7	22	108	0.86	1.06	0.95
English (NUIG)	885	339	42	67	1333	0.58	0.89	0.70
Estonian	1043	61	54	4	1162	1.02	1.02	1.02
German	50	71	104	2	227	0.54	0.92	0.68
Hungarian	49	13	10	9	81	0.61	0.79	0.69
Irish	731	45	67	132	975	0.97	0.74	0.84
Italian	116	39	16	28	199	0.74	0.96	0.83
Serbian	85	8	24	28	145	0.38	0.95	0.54
Slovenian	20	23	14	6	63	0.63	1.31	0.85
Russian	63	2	40	21	126	1.26	0.86	1.02

$K \rightarrow$ the average degree of senses

Evaluation

5. Inter-annotator agreement using Fleiss' Kappa

	Agreement (5-class)	Agreement (2-class)
Irish	0.373	0.468
English (NUIG)	0.931	0.95
Danish	0.837	0.562

Conclusion and future steps

We believe that our datasets will

- pave the way for further developments in exploring statistical and neural methods in **semantic relationship detection** and **word sense alignment** in monolingual data
- be useful for evaluation purposes

References

McCrae, J. P. (2018). Mapping wordnet instances to Wikipedia. In Proceedings of the 9th Global WordNet Conference (GWC 2018), pages 62–69.

Meyer, C. M. and Gurevych, I. (2011). What psycholinguists know about chemistry: Aligning Wiktionary and WordNet for increased domain coverage. In Proceedings of 5th International Joint Conference on Natural Language Processing, pages 883–892.

Navigli, R. (2006). Meaningful clustering of senses helps boost word sense disambiguation performance. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 105–112. Association for Computational Linguistics.

Henrich, V., Hinrichs, E., and Vodolazova, T. (2011). Semi-automatic extension of GermaNet with sense definitions from Wiktionary. In Proceedings of the 5th Language and Technology Conference (LTC 2011), pages 126–130.

Henrich, V., Hinrichs, E. W., and Suttner, K. (2012). Automatically linking GermaNet to Wikipedia for harvesting

Thanks 😊
Any question?