



# Automatic Alignment of Lexicographical Data

Sina Ahmadi (<https://sinaahmadi.github.io/>)  
National University of Ireland Galway

*Analyse et traitement informatique de la langue française (ATILF)*  
July 13, 2021



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

# Outline

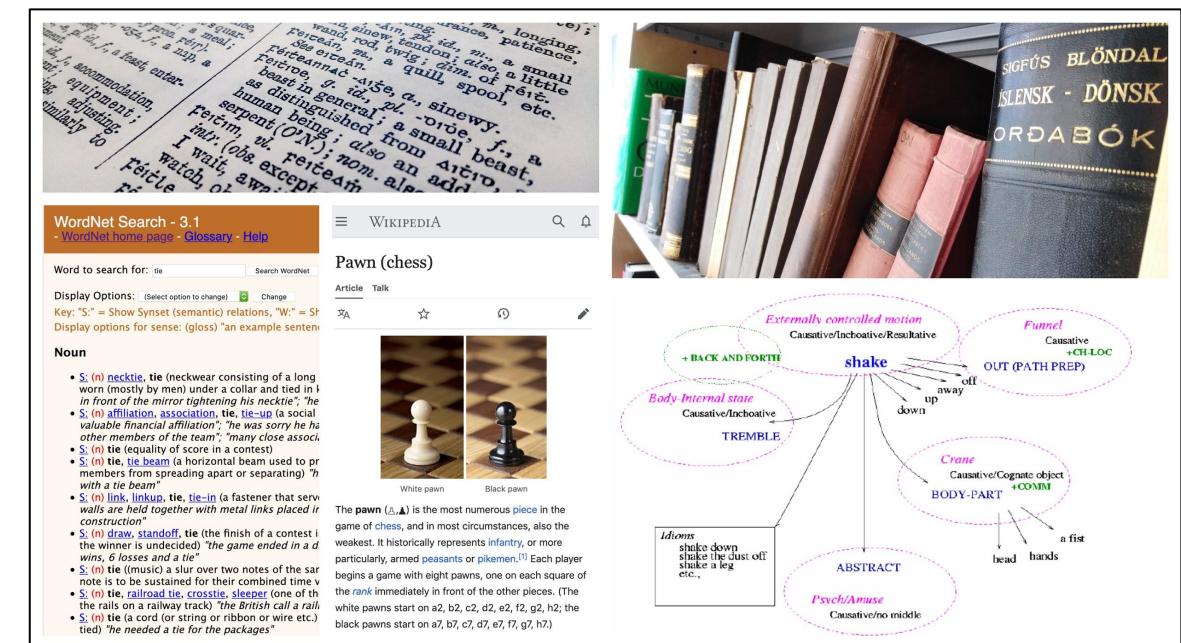
- 1. Context**
- 2. Word-sense alignment**
- 3. Bilingual lexicon induction**
- 4. Naisc**
- 5. Next steps**

# 1. Context

**"Dictionaries are treasure houses of data on the uses of words. They are also our best starting point for all questions regarding word sense distinctions, in NLP, the humanities or lexicography.**

**But to reveal the dictionary's treasures in a systematic way is no simple."**

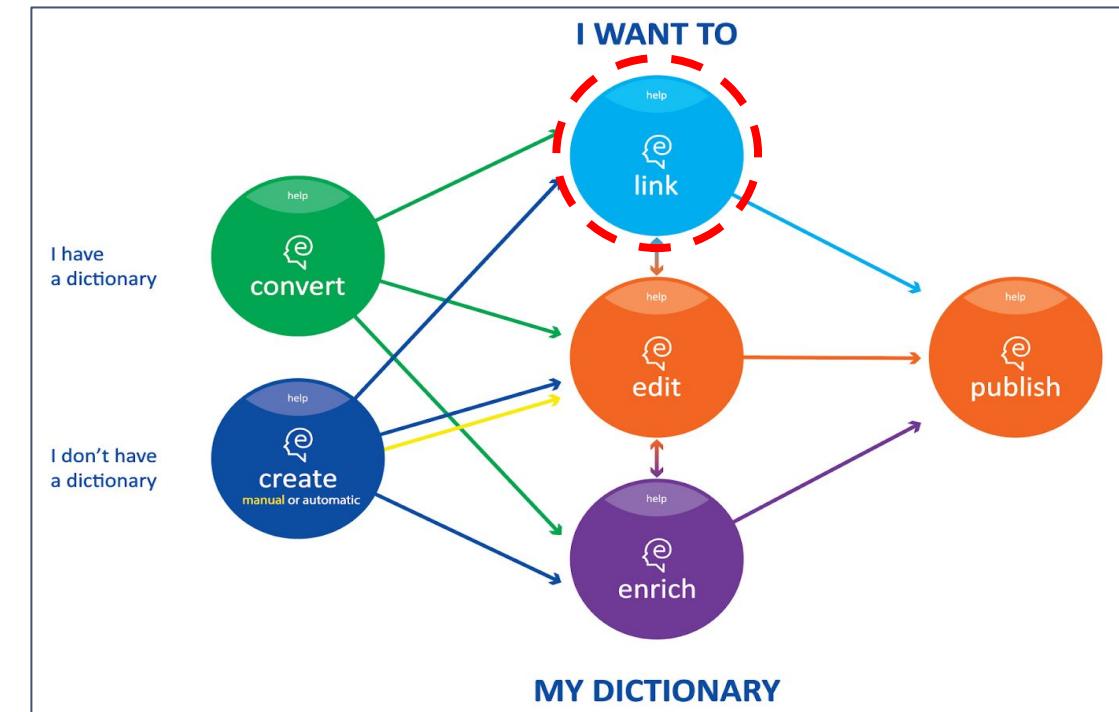
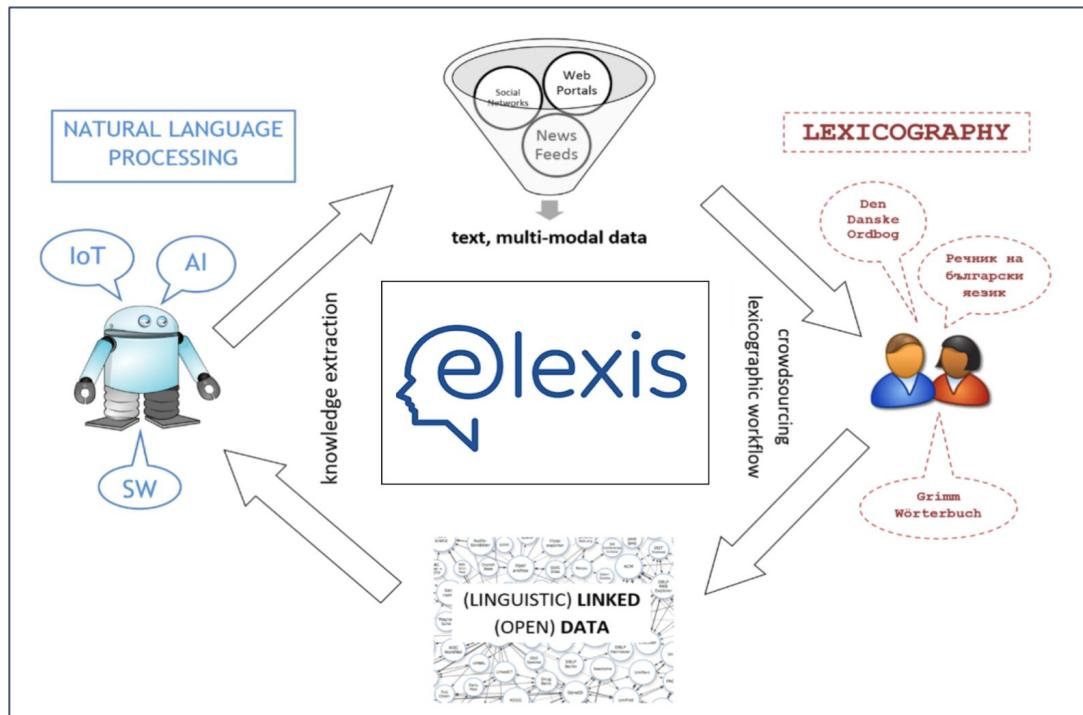
— Adam Kilgarriff (*Dictionary Word Sense Distinctions: An Enquiry into Their Nature*)



## 1. Context

# ELEXIS – European Lexicographic Infrastructure

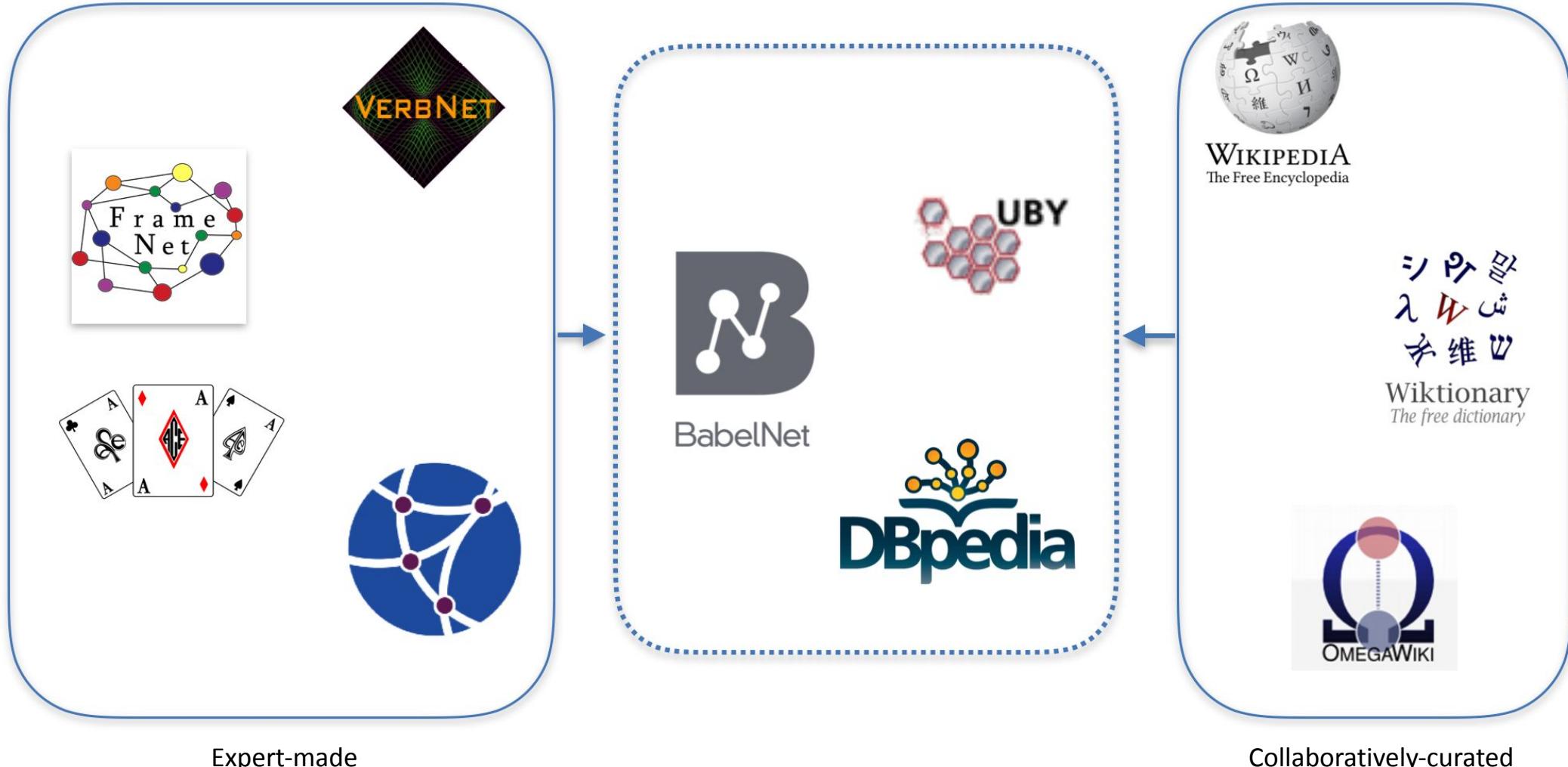
17 partners from 15 countries, 73 institutions from 37 countries  
 (February 2018 - July 2022)



\* Find out more about ELEXIS at <https://elex.is/>

## 1. Context

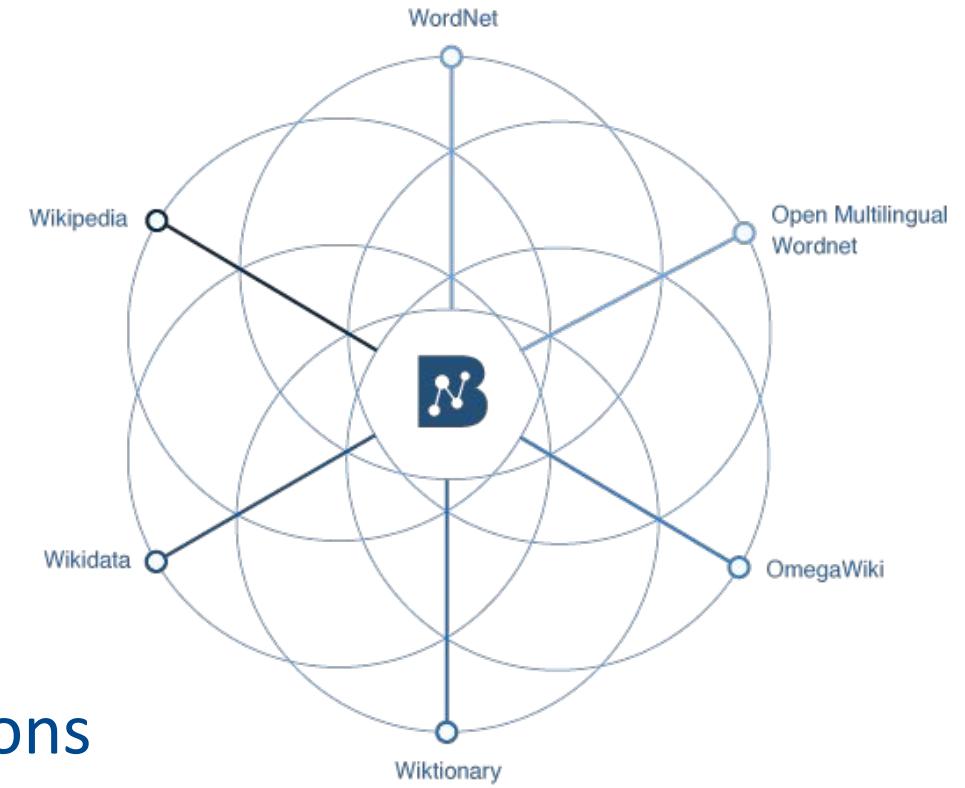
# Resource alignment



## 1. 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 languages
- To combine expert-made semantic relations
  - e.g., Hypernymy, meronymy, etc.

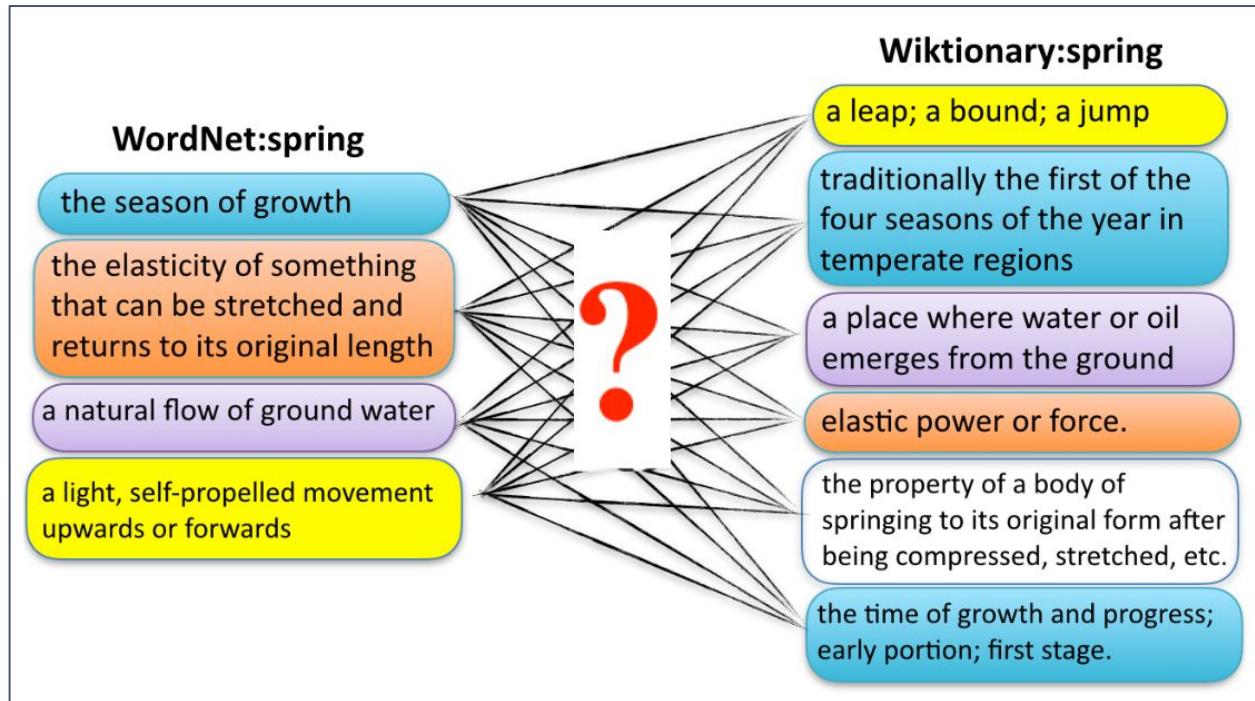


BabelNet (<https://babelnet.org/>)

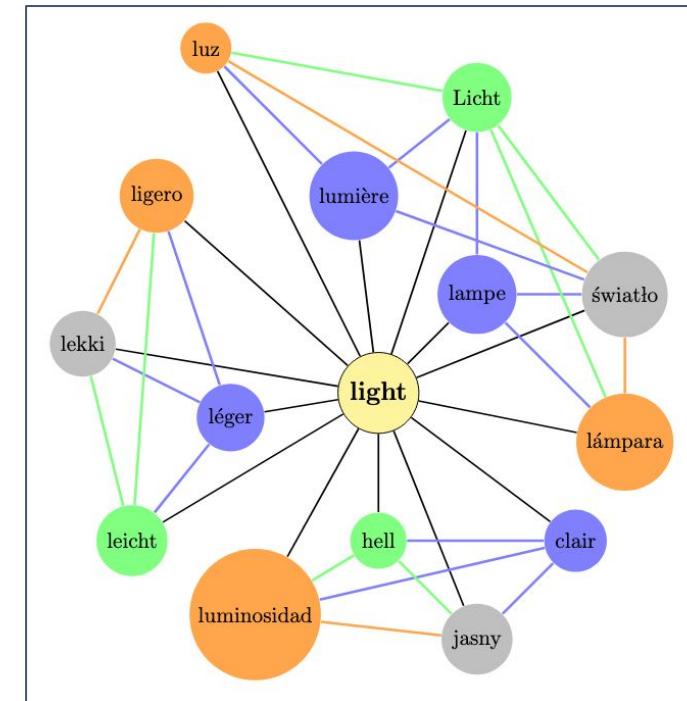
## 1. Context

# Let's set it up.

### At sense level



### At entry level



Of course, many more fields to explore, e.g. semantic relationship alignment [5], concept alignment [6] 7, and ontology alignment [7].

## 2. Word sense alignment (WSA)

→ linking lexical content at sense level, including glosses

**lead<sup>2</sup>** ●●○ **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: lead <sup>1</sup>	English: lead <sup>2</sup>	American: lead <sup>1</sup>	American: lead <sup>2</sup>	Specialist English: lea ▶				
16. countable noun											
A <b>lead</b> 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.											
<i>The inquiry team is also following up possible leads after receiving 400 calls from the public.</i>											
Synonyms: clue, tip, suggestion, trace				More Synonyms of lead							
17. countable noun											
<b>The lead</b> in a play, film, or show is the most important part in it. The person who plays this part can also be called the <b>lead</b> .											
<i>Performers from the Bolshoi Ballet dance the leads.</i>											
<i>Both the leads in the play are impressive.</i>											
Synonyms: leading role, principal, protagonist, title role				More Synonyms of lead							
18. countable noun											
A dog's <b>lead</b> 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]											
<i>An older man came out with a little dog on a lead.</i>											
REGIONAL NOTE: in AM, use <b>leash</b>				Collins DICTIONARY							



## 2. Word sense alignment

# Sense alignment is challenging

- Differences in structure
  - sense vs. sub-sense vs. sub-sub-sense etc.
  - formalizations, such as WordNet [8], FrameNet [9] and generative lexicon [10]
- Differences in content
  - Lexical choice:
    - **Alcohol:** vandklar vædske (water-clear liquid) vs. farveløs (colorless) in Danish
  - Definition paradigms [11], as in:
    - **footnote:** A footnote is a note of text placed at the bottom of a page (analytical)
    - **méchant :** *qui est dangereux, nuisible, néfaste* (synonymous)
    - **good:** the opposite of bad (relational)



# An example: *afstand* ['aw̥sdan?] (distance in Danish)\*

1) **fjernhed**; længden af **mellerummet** (mat.: af en ret linie) mellem to punkter. udfinde Solens Afstand fra Jorden. *Heitm. Physik.*67. (jf. Steners. *CritBet.*29 og Marg. *Klopstock.Breve.*(1760).40). Søfarende . . have ofte stor Færdighed i at bedømme Afstandene. *Heib.Pros.*II.369. \*Seer jeg .. en Hatfuld sydet Damp | Tidens Maal, Rummets Afstande flytte. *Ploug.*VV.II.13. Afstanden fra Kærsholm til Bøstrup Præstegaard var fem-seks Kilometer. *Pont.LP.*VII.65. (sj.:) han vandt Afstand (dvs.: kom længere og længere bort) fra (de angribende vilde heste). *Rist.FT.*28. || ✕ spec. om regelmæssige mellerum mellem (afdelinger af) soldater, som staar bag ved hinanden. *Sal.*IX.531. jf.: Der er Gæssene . . med Retning og Afstand som en Trup Soldater. *Bogan.*I.127. det er daarlig ridning; her er ikke spor af afstand (dvs.: der er ulige stor afstand mellem de enkelte rytttere) || | | efter præp. i. \*Alt, hvad Naturen . . | I maalløs Afstand fra hinanden spredes. *Bagges.*L.I.156. Munken . . holder sig i en ærbødig Afstand. *Oehl.*IV.161. Medens vi talte, saa jeg i lang Afstand en Dame komme. *Goldschm.*VI.274. i **afstand** (ell. † i en afstand. *Gylb.*III.215. IV.280.331. VIII.223), (sj.) **ikke tæt ved** ell. paa nært hold; (temmelig) langt borte. (nu oftere paa afstand). \*Jeg vendte om, og som en ydmyg Slave | I Afstand troe jeg fulgte Deres Vei. *Heib. Poet.*VII.272. Vandet saae klart ud nær ved, men seet i Afstand, sort som Blæk. *HCAnd.*VI.300. Jeg elsker Franskmændene – i Afstand. *Goldschm.*J.354. De dræbte ham i Afstand med Pile. *smst.*III. 197. || efter præp. paa. \*Ei en af Tillys Mænd er under Vaaben | Paa mange Miles Afstand. *Hauch.*Æ.70. Der skulde skydes (dvs.: ved en duel) paa 15 Skridts Afstand. *JakKnu.A.*211. **paa afstand**, d. s. s. i afstand. Hun havde ogsaa noget Godt i sine Øjne, naar hun var lidt paa Afstand. *Schand.*BS.118. Aftenklokken . . lød saa smukt paa Afstand. *Pont.LP.*VII.89. Enhver Voksen der har den mindste Katar bør holde sig paa Afstand fra Børnene. *Sundhedstid.*1916.266.

\* <https://ordnet.dk/ddo/ordbog?query=afstand>

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

```
<Semdel SemID="29014340"><Semem BetNo="1" SememID="29401382" odsID="Afstand_1"><dotPlus id="2666"/><SemIndhold><dotBetNo>1</dotBetNo> <i><dotSpaced>fjernhed;</dotSpaced> længden af <dotSpaced>mellerummet</dotSpaced><dotLn col="0293" lin="25" orig="mellerum-met"/>(mat.: af en ret linie) mellem to punkter.</i> <dotLn col="0293" lin="26"/>udfinde Solens Afstand fra Jorden. <i>Heitm. <dotLn col="0293" lin="27"/>Physik.67. (jf. Steners. CritBet.29 og Marg. <dotLn col="0293" lin="28"/>Klopstock.Breve.(1760).40).</i> Søfarende . . <dotLn col="0293" lin="29"/>have ofte stor Færdighed i at bedømme <dotLn col="0293" lin="30"/>Afstandene. <i>Heib.Pros. II.369.</i> <dotRaised>*</dotRaised>Seer jeg <dotLn col="0293" lin="31"/>.. . en Hatfuld sydet Damp | Tidens Maal, <dotLn col="0293" lin="32"/>Rummets Afstande flytte. <i>Ploug. VV.II.13.</i> <dotLn col="0293" lin="33"/>Afstanden fra Kærsholm til Bøstrup Præstegaard<dotLn col="0293" lin="34" orig="Præste-gaard"/>var fem-seks Kilometer. <i>Pont.LP. <dotLn col="0293" lin="35"/>VII.65. (sj.:)</i> han vandt Afstand <i>(dvs.: kom <dotLn col="0293" lin="36"/>længere og længere bort)</i> fra <i>(de angribende <dotLn col="0293" lin="37"/>vilde heste). Rist.FT.28.</i> <Planke PlankeID="29900969"><dotEdBreak/> <piktogram src="swords" title="sværd"/> <i>spec. om regelmæssige<dotLn col="0293" lin="38" orig="regel-mæssige"/>mellerum mellem (afdelinger af) <dotLn col="0293" lin="39"/>soldater, som staar bag ved hinanden. Sal. <dotLn col="0293" lin="40"/>IX.531. jf.:</i> Der er Gæssene . . med Retning<dotLn col="0293" lin="41" orig="Ret-ning"/>og Afstand som en Trup Soldater. <dotLn col="0293" lin="42"/><i>Bogan. I.127.</i> det er daarlig ridning; her er ikke spor af afstand <i>(dvs.: der er ulige stor <dotLn col="0293" lin="44"/>afstand mellem de enkelte rytttere)</i> <dotEdBreak/> <piktogram src="dotpipe" title="dobbeltbrudt streg"/> </Planke><Planke PlankeID="29900970"><dotPlus id="2667"/><dotEdBreak/> <i>efter <dotLn col="0293" lin="45"/>præp.</i> i. <dotRaised>*</dotRaised>Alt, hvad Naturen . . | I maalløs <dotLn col="0293" lin="46"/>Afstand fra hinanden spredes. <i>Bagges. L. <dotLn col="0293" lin="47"/>I.156.</i> Munken . . holder sig i en ærbødig <dotLn col="0293" lin="48"/>Afstand. <i>Oehl. IV.161.</i> Medens vi talte, saa <dotLn col="0293" lin="49"/>jeg i lang Afstand en Dame komme. <i>Goldschm. VI.274.</i><dotLn col="0293" lin="50" orig="Gold-schm.VI.274."/><b>i afstand</b> <i>(ell.</i> <piktogram src="cross" title="kors"/> i en afstand. <dotLn col="0293" lin="51"/><i>Gylb. III.215. IV.280.331. VIII.223), (sj.) <dotLn col="0293" lin="52"/><dotPlus col="293" id="2668" lin="50"/><dotSpaced>ikke tæt ved</dotSpaced> ell. paa nært hold; (temmelig) <dotLn col="0293" lin="53"/>langt borte. (nu oftere</i> paa afstand<i>)</i> <dotRaised>*</dotRaised>Jeg <dotLn col="0293" lin="54"/> vendte om, og som en ydmyg Slave | I <dotLn col="0293" lin="55"/>Poet. VII.272.</i> Vandet saa i Afstand, sort som Blæk. <dotLn col="0293" lin="58"/><dotLn col="0293" lin="59" orig="Franskmæn-dene"/> i Afstand. <i>ham i Afstand med Pile. <i>smst. III. <dotLn col="0293" lin="60"/><dotPlus id="2669"/><dotPlus id="2670" kontrol="ingen"><i>efter præp.</i> paa. <dotRaised>*</dotRaised>Ei ei Vaaben | Paa mange <dotLn col="0293" lin="63"/>Miles A lin="64"/>skydes <i>(dvs.: ved en duel)</i> paa 15 Skridts Afstand. <i>JakKnu.A.211.</i> <b>paa afstand,</b> <i>d. s. s.</i> noget Godt <dotLn col="0294" lin="01"/>i sine Øjne, naa Schand. BS.118.</i> Aftenklokken . . lød saa <dotLn col="0294" lin="04" orig="En-hver"/><i>Enhver<dotLn col="0294" lin="04"/>bør holde sig paa Afstand fra Børnene. <dotLn col="0294" lin="07"/></i></Planke></SemIndhold></Semem></Semdel>
```



## 2. Word sense alignment

# What has been done?

- A significant body of research in aligning English resources including linking the Princeton WordNet with
  - Wikipedia [12]
  - Wiktionary [13]
  - the Oxford Dictionary of English [14]
  - Wikidata [15]
- A fewer number of manually aligned monolingual resources in other languages including linking:
  - the GermaNet—the German Wordnet with
    - the German Wikipedia [16]
    - the German Wiktionary [17]
- Not too many studies on other languages!



## 2. Word sense alignment

# Monolingual WSA datasets

To address some of the current main limitations in WSA:

- Multilingualism
- Monolingual resources
- Gold-standard datasets
- Semantic relationship annotation

→ 17 manually-annotated monolingual resources for the task of WSA covering 15 languages

### A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment

Sina Ahmadi<sup>1\*</sup>, John P. McCrae<sup>2</sup>,  
 Sami Nimi<sup>1</sup>, Fahad Khan<sup>3</sup>, Monica Monachini<sup>3</sup>, Bolette S. Pedersen<sup>3</sup>, Thierry Declerck<sup>2,12</sup>, Tanja Wissis<sup>2</sup>,  
 Andrés Belland<sup>3</sup>, Irene Pisani<sup>4</sup>, Thomas Troelsgård<sup>1</sup>, Sussi Olsen<sup>5</sup>, Simon Krek<sup>6</sup>, Veronika Lipp<sup>7</sup>,  
 Tamás Vízárdi<sup>8</sup>, László Simon<sup>9</sup>, András Györfy<sup>10</sup>, Carola Tiberius<sup>10</sup>, Tanneke Schooneheim<sup>9</sup>, Yifat Ben Moshe<sup>10</sup>,  
 Maya Rudich<sup>10</sup>, Raya Abu Ahmad<sup>10</sup>, Doriele Louke<sup>10</sup>, Kira Kovaleenko<sup>11</sup>, Margit Langenmetz<sup>11</sup>, Jelena Kallas<sup>13</sup>,  
 Oksana Dereza<sup>13</sup>, Theodora Franssen<sup>13</sup>, David Cilliesen<sup>13</sup>, David Lindemann<sup>14</sup>, Mikel Alonso<sup>14</sup>, Ana Salgado<sup>14</sup>,  
 José Luis Sancho<sup>15</sup>, Rafael J. Ureña-Ruiz<sup>16</sup>, Jordi Porta Zamorano<sup>16</sup>, Kiril Simov<sup>17</sup>, Petya Osenova<sup>17</sup>,  
 Zara Kancheva<sup>17</sup>, Ivaylo Rusev<sup>17</sup>, Radka Stankova<sup>18</sup>, Andrej Perdič<sup>19</sup>, Dejan Gabrović<sup>19</sup>

<sup>1</sup>Insight Centre for Data Analytics, National University of Ireland, Galway  
 (\*sina.ahmadi.john.mccrae)@insight-centre.org  
 (other affiliations in Appendix A)

<sup>2</sup>Arora

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 15 languages. 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. Our resources are publicly available at <https://github.com/elexis-eu/MWSA>.

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

Mihalea, 2005; Punzetto 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 Wikipedia<sup>1</sup>, and expert-made ones such as WordNet. However, the previous work limited to a handful of resources and most 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 quite small and the extent to which the mapping is reliable is unclear.

In this paper, we present a set of datasets for the task of WSA containing manually-annotated monolingual resources in 15 languages. The annotation is carried out at sense level where four semantic relationships, namely, relatedness, equivalence, broadness, and narrowness, are selected. A graph of sensed in 15 languages is built by native lexicographers. Given the lexicographic context of this study, we have tried to provide lexicographic data from expert-made dictionaries. We believe that our datasets will pave the way for further developments in exploring statistical and neural methods, as well as for evaluation purposes. The rest of the paper is organized as follows: we first describe the previous work in Section 2. After having de-

<sup>1</sup><https://www.wikipedia.org>

[2]

\*Contact Authors  
<https://globalex2020.globalex.link>



## 2. Word sense alignment

# Monolingual WSA datasets

<b>elocutionary (adjective)</b>	(used of style of speaking) overly embellished of or relating to elocution	NONE exact	▼ ▼	0 -pertaining to elocution. 0 -pertaining to elc
<b>montgolfier (noun)</b>	French inventor who (with his brother Josef Michel	NONE	▼	0 -a balloon which
	French inventor who (with his brother Jacques Etie	NONE	▼	0 -a balloon which
<b>dice (verb)</b>	cut into cubes	exact	▼	1. -to cut into small cubes; .
	play dice	NONE	▼	2. -to ornament with squares, diamonds, or
<b>ebb (verb)</b>	flow back or recede	exact	▼	0 -to cause to flow
	fall away or decline	NONE	▼	0 -to cause to flow back.
<b>educated (adjective)</b>	characterized by full comprehension of the problem		▼	0 -formed or developed by education; .
	possessing an education (especially having more t	exact	▼	
<b>quaver (verb)</b>	give off unsteady sounds, alternating in amplitude	narrower	▼	0 -to utter with quavers.
	sing or play with trills, alternating with the half note	broader	▼	
<b>rangy (adjective)</b>	adapted to wandering or roaming	related	▼	0 -inclined or able to range, or rove about, fo
	allowing ample room for ranging	NONE	▼	
	tall and thin and having long slender limbs		▼	
<b>smiler (noun)</b>	a person who smiles		▼	0 -one who smiles.
	the human face ('kisser' and 'smiler' and 'mug' are		▼	
<b>chair (noun)</b>				

```
{
  "lemma": "splenetic",
  "POS_tag": "adjective",
  "gender": "",
  "meta_ID": "",
  "resource_1_senses": [
    {
      "#text": "of or relating to the spleen",
      "external_ID": "splenic.a.01"
    },
    {
      "#text": "very irritable",
      "external_ID": "bristly.s.01"
    }
  ],
  "resource_2_senses": [
    {
      "#text": "affected with spleen; malicious; → spiteful; peevish; fretful.",
      "external_ID": ""
    }
  ],
  "alignment": [
    {
      "sense_source": "very irritable",
      "sense_target": "affected with spleen; → malicious; spiteful; peevish; → fretful.",
      "semantic_relationship": "exact"
    }
  ]
}
```

## 2. Word sense alignment

# Semantic relationships (based on SKOS\*)

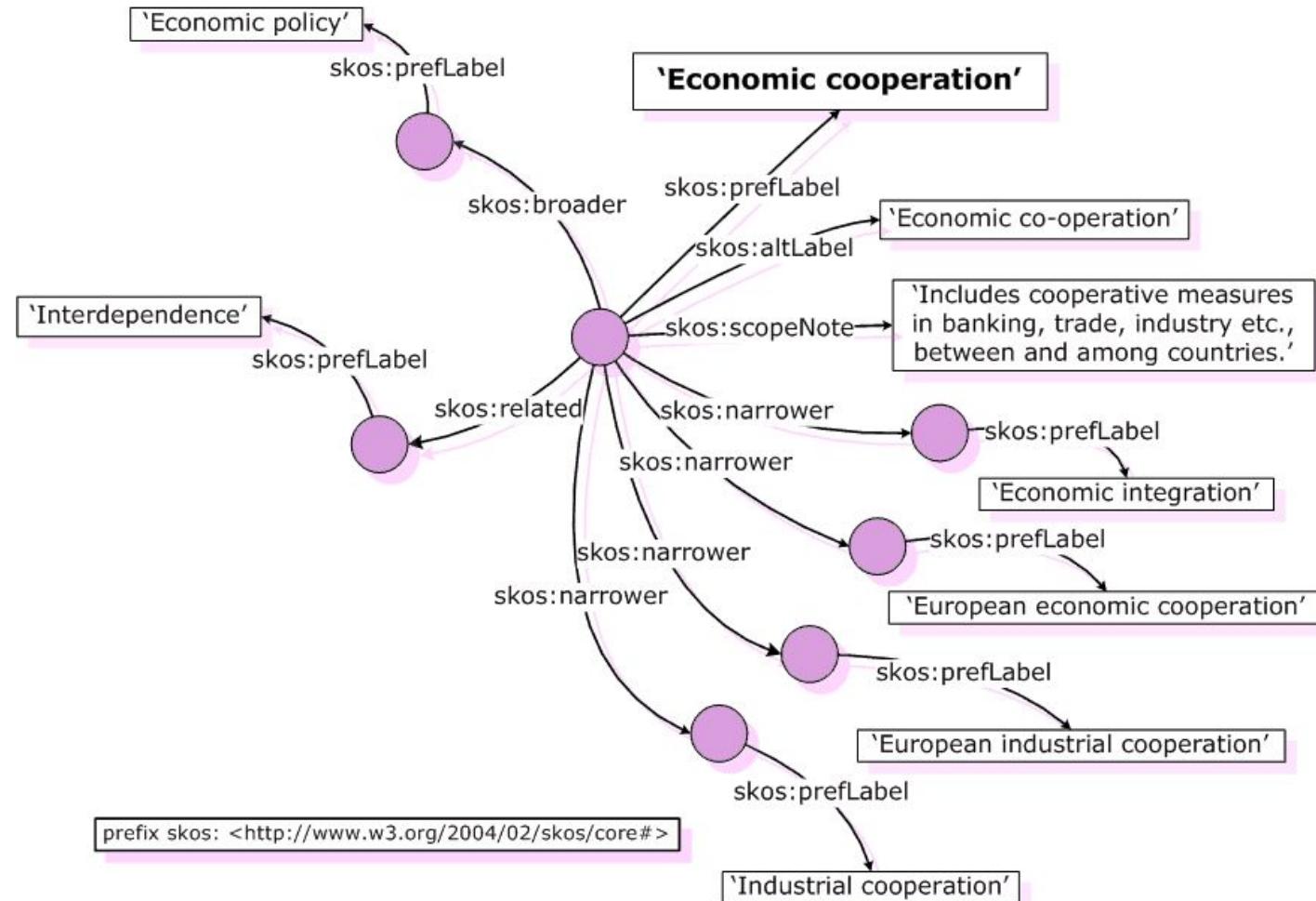
- **exact**: The sense are the same, for example the definitions are simply paraphrases
- **broader**: The sense in the first dictionary completely covers the meaning of the sense in the second dictionary and is applicable to further meanings
- **narrower**: The sense in the first dictionary is entirely 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/>



## 2. Word sense alignment

# Semantic relationships: an example



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



Language	Resource	Nouns	Verbs	Adjectives	Adverbs	Other	All
Basque	Basque Wordnet	929 (6836)	0 (0)	0 (0)	0 (0)	0 (0)	929 (6836)
	<i>Euskal Hiztegia</i>	971 (7754)	0 (0)	0 (0)	0 (0)	0 (0)	971 (7754)
Bulgarian	BTB-WN	1394 (15649)	175 (1698)	305 (3187)	50 (338)	0 (0)	1924 (20872)
	Bulgarian Wiktionary	1273 (12883)	164 (1107)	194 (1418)	39 (306)	0 (0)	1670 (15714)
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>	1459 (28979)	405 (5185)	527 (7878)	106 (2662)	0 (0)	2497 (44704)
	<i>Algemeen Nederlands Woordenboek</i>	497 (8443)	140 (1542)	109 (1393)	13 (172)	0 (0)	759 (11550)
English (KD)	Global	92 (532)	107 (617)	80 (457)	57 (257)	61 (283)	397 (2146)
	Password	66 (536)	72 (417)	62 (324)	33 (177)	46 (188)	279 (1642)
English (NUIG)	Webster	1131 (11606)	741 (4622)	373 (2585)	45 (269)	0 (0)	2290 (19082)
	<i>Princeton WordNet</i>	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 Wiktionary	2026 (15160)	0 (0)	0 (0)	0 (0)	0 (0)	2026 (15160)
	German OmegaWiki	1266 (14354)	0 (0)	0 (0)	0 (0)	0 (0)	1266 (14354)
Hungarian	Comprehensive						1355 (14654)
	Explanatory						1038 (10934)
Irish	<i>An Foclóir Beag</i>	891 (8053)	11 (95)	55 (267)	10 (56)	36 (171)	1003 (8642)
	Irish Wiktionary	1209 (6696)	8 (45)	61 (181)	10 (41)	36 (109)	1324 (7072)
Italian	ItalWordNet	408 (3128)	352 (2411)	0 (0)	0 (0)	0 (0)	760 (5539)
	SIMPLE	290 (1990)	218 (1240)	0 (0)	0 (0)	0 (0)	508 (3230)
Serbian	Serbian WordNet	691 (5864)	985 (6522)	92 (713)	0 (0)	0 (0)	1768 (13099)
	Dictionary of Serbo-Croatian Literary Language	289 (2360)	281 (1527)	29 (215)	0 (0)	0 (0)	599 (4102)
Slovenian (JSI)	Slovene WordNet	409 (1106)	303 (901)	237 (733)	44 (133)	0 (0)	993 (2873)
	Slovene Lexical Database	284 (2237)	191 (1047)	220 (1486)	29 (102)	0 (0)	724 (4872)
Slovenian (ISJFR)	Standard Slovenian Dictionary (eSSKJ)	229 (2060)	109 (911)	76 (620)	0 (0)	60 (588)	474 (4179)
	<i>Kostelski slovar</i>	151 (1050)	61 (308)	45 (257)	0 (0)	38 (263)	295 (1878)
Spanish	<i>Diccionario de la lengua española</i>	617 (7986)	225 (2426)	305 (3269)	26 (161)	24 (250)	1197 (14092)
	Spanish Wiktionary	602 (6421)	227 (2045)	294 (2825)	25 (129)	22 (123)	1170 (11543)
Portuguese	<i>Dicionário da Língua Portuguesa Contemporânea</i>	285 (4060)	58 (686)	110 (1287)	9 (143)	1 (9)	463 (6185)
	<i>Dicionário Aberto</i>	199 (1521)	53 (203)	67 (372)	3 (15)	1 (5)	323 (2116)
Russian	<i>Ozhegov-Shvedova</i>	258 (2038)	109 (615)	101 (533)	15 (77)	44 (368)	527 (3631)
	Dictionary of the Russian Language (MAS)	310 (2811)	173 (1338)	190 (1219)	20 (114)	71 (1010)	764 (6492)

## 2. Word sense alignment

# MWSA Datasets Statistics

## Number of senses and number of tokens (in parentheses)

 Data openly available at:  
<https://github.com/elexis-eu/mwsa>



## 2. Word sense alignment

# Approaches and state-of-the-art

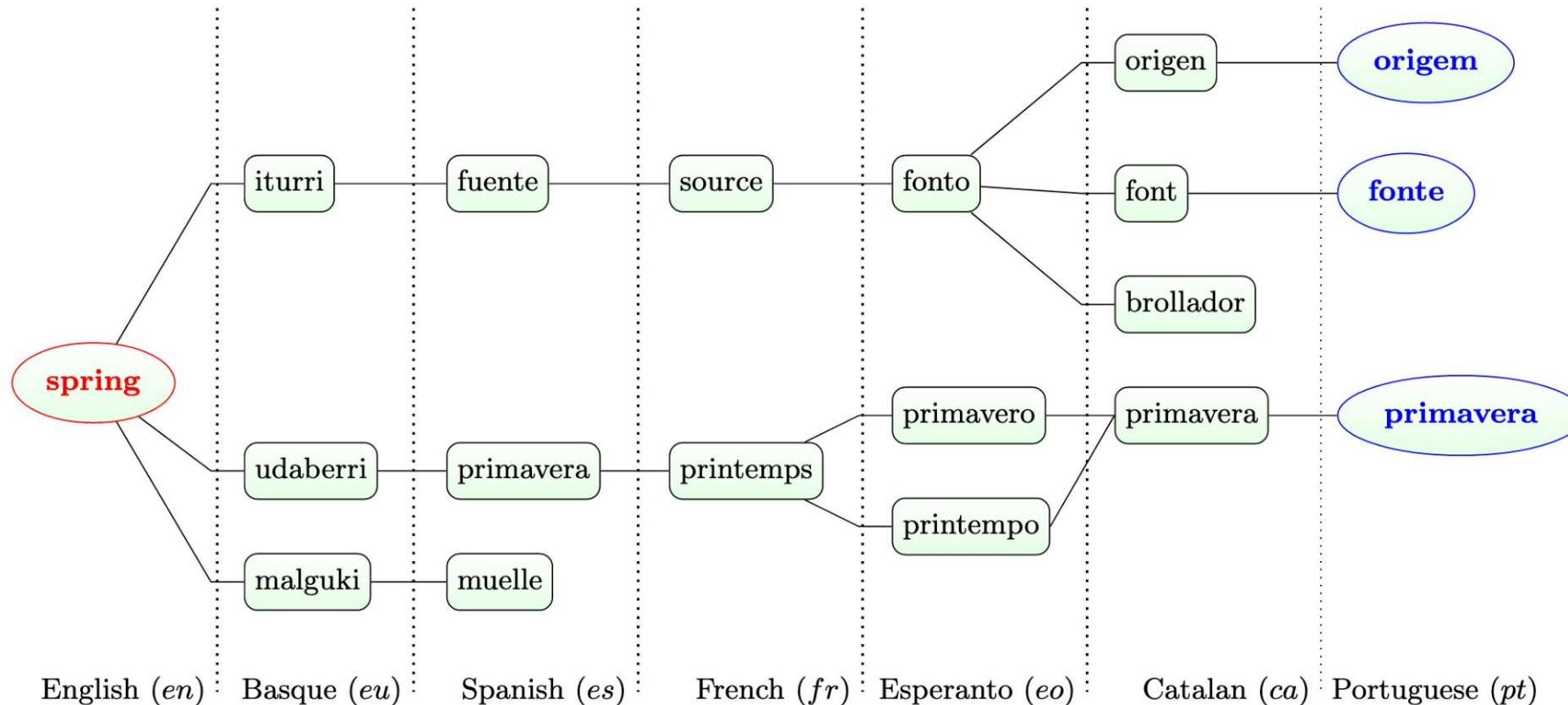
Results of the English monolingual WSA in the context of the Monolingual Word Sense Alignment Shared Task (LREC 2020): <https://competitions.codalab.org/competitions/22163>

Approach	Accuracy	Precision	Recall	F1-measure
Jaccard similarity with the Hungarian Algorithm [19]	0.752	0	0	0
similarity methods as well as similarities coming from ELMo and BERT [20]	0.763	0.619	0.782	0.691
word-sense disambiguation, multiple features from BERT combined using a random forest [21]	0.798	0.746	0.353	0.48
BERT & Siamese LSTM [22]	0.759	0.586	0.692	0.634
Features from ConceptNet & restricted Boltzmann Machine [23]	<b>89</b>	<b>82.35</b>	<b>82.87</b>	<b>82.61</b>



### 3. Bilingual Lexicon Induction (BLI)

Given two dictionaries in two different languages, generate new translation pairs → linking dictionaries at entry level



Based on the in the Apertium translation data: <http://tiad2021.unizar.es/>

### 3. Bilingual Lexicon Induction (BLI)

## Translation Inference across Dictionaries (TIAD)

- automatic **generation** of new bilingual (and multilingual) dictionaries from existing ones
- no translations can be obtained directly among source and target languages based on the available RDF data
- no external resource with alignments between the source and target languages is allowed to be used → **unsupervised**

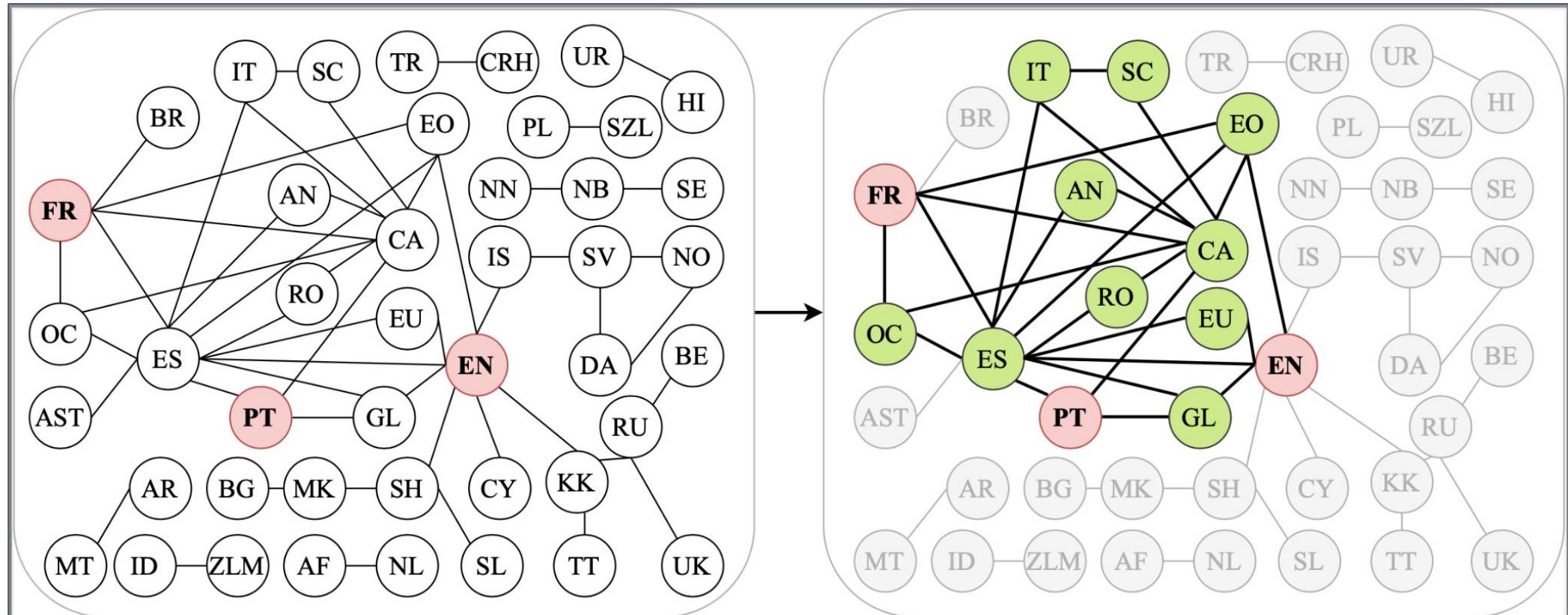
 See more at <https://tiad2021.unizar.es/>

Year	Target dictionaries	Paper	Approach	External resources
TIAD 2017	German-Portuguese	[1]	graph analysis	-
	Danish-Spanish	[16]	graph analysis and collocation-based models	Europarl corpus
	Dutch-French	[7]	Support Vector Machine using features based on the translation graph and string similarity	-
TIAD 2019	English	[2]	multi-way neural machine translation	corpora of languages from the same family and Wiktionary
	French	[21]	graph analysis and neural machine translation	Directorate General for Translation corpus [18]
	Portuguese	[9]	pivot-based and cross-lingual word embeddings	monolingual corpora
		[6]	multi-lingual word embedding	pretrained embedding model
		[14]	unsupervised document embedding using Orthonormal Explicit Topic Analysis	Wikipedia corpora
TIAD 2020	English	[15]	unsupervised multi-way neural machine translation and unsupervised document embedding	Directorate General for Translation corpus [18]
	French	[4]	propagation of concepts over a graph of interconnected dictionaries using WordNet synsets and lexical entries as concepts	WordNet
	Portuguese	[13]	graph analysis and cross-lingual word embeddings	monolingual corpora of Common Crawl and Wikipedia
		[8]	graph analysis relying on paths, synonyms, similarities and cardinality in the translation graph	-

References available at [1]

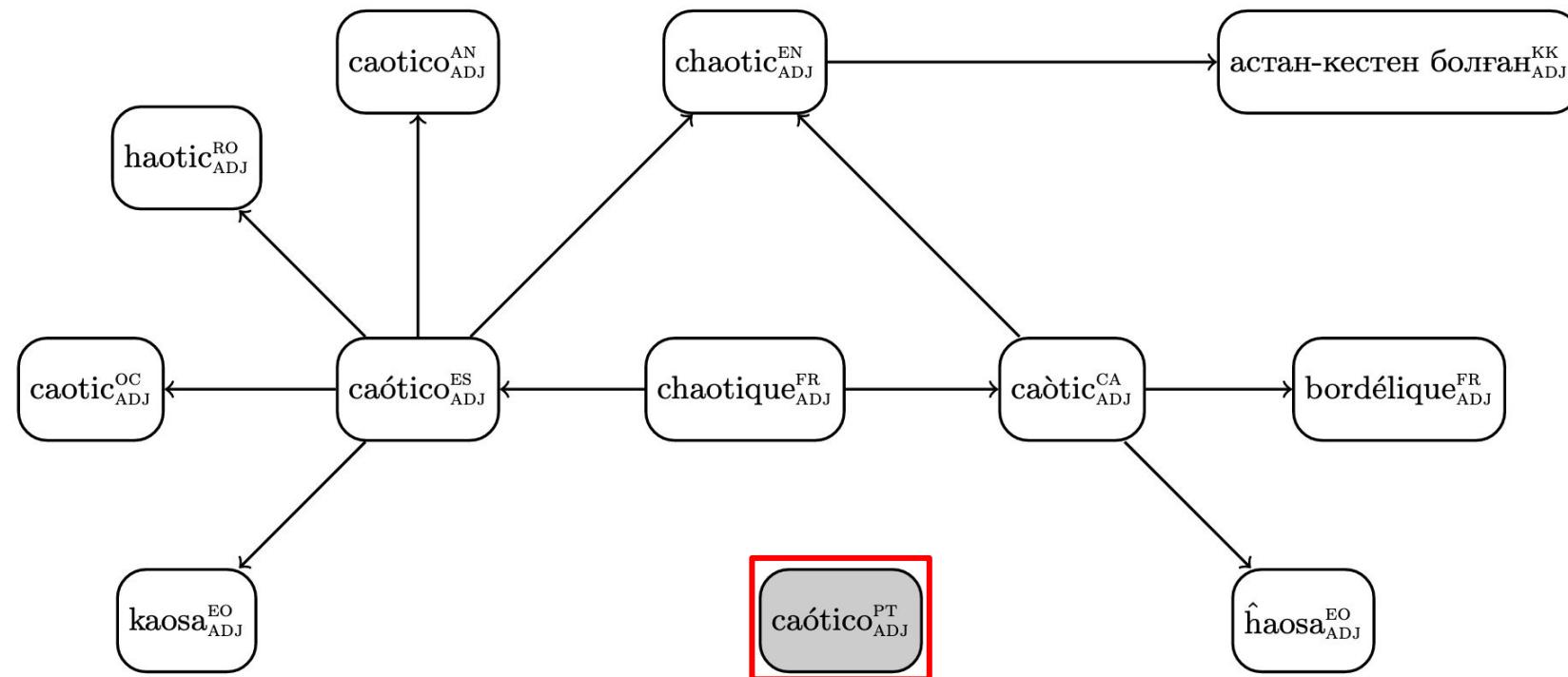
### 3. Bilingual Lexicon Induction (BLI)

## Exploring the graph structure



### 3. Bilingual Lexicon Induction (BLI)

# Limitations due to coverage

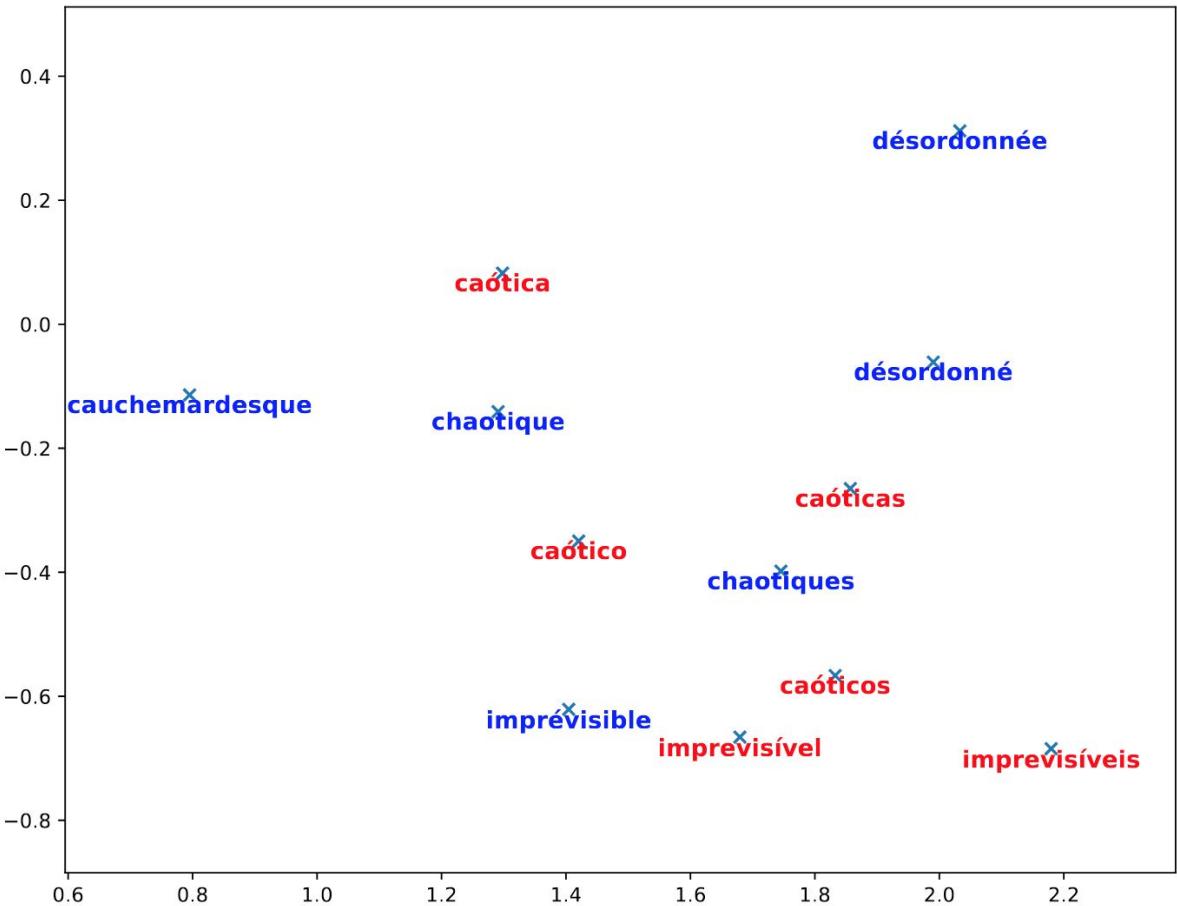


### 3. Bilingual Lexicon Induction (BLI)

## Cross-lingual word embeddings mapping

Using unsupervised cross-lingual word embedding mapping techniques, find a **mapping between the monolingual word embedding spaces** of the source and target languages.

- VecMap [24]
- MUSE [25]
- Both methods, induce a seed lexicon automatically assuming approximate isomorphism between source and target spaces

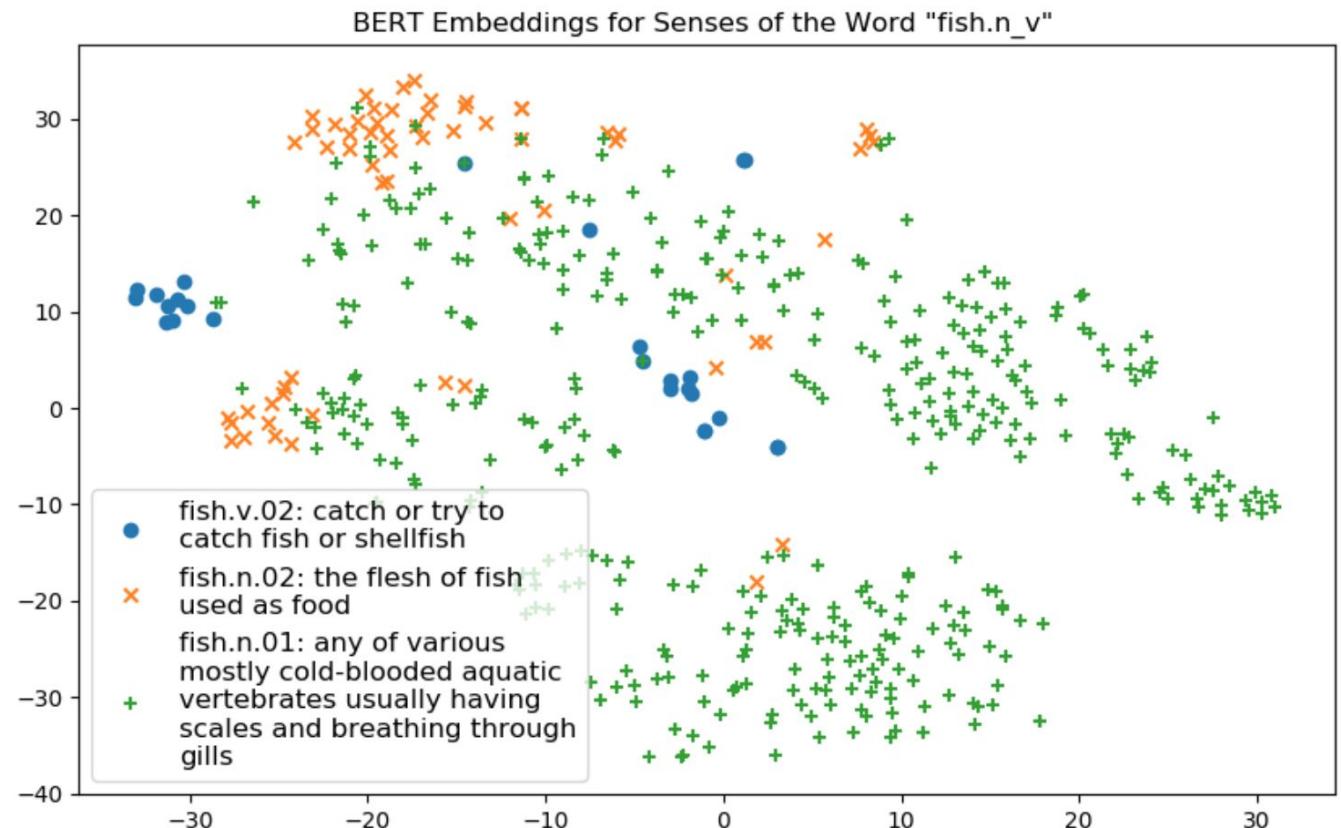


t-SNE visualization of *chaotique* (adjective in French) in the MUSE multilingual word embeddings of French and Portuguese

### 3. Bilingual Lexicon Induction (BLI)

## Contextual embeddings

- How to incorporate representation of sense distinction in the two previous tasks?
- A fundamental question: Are there a certain number of senses for a word? What determines it?



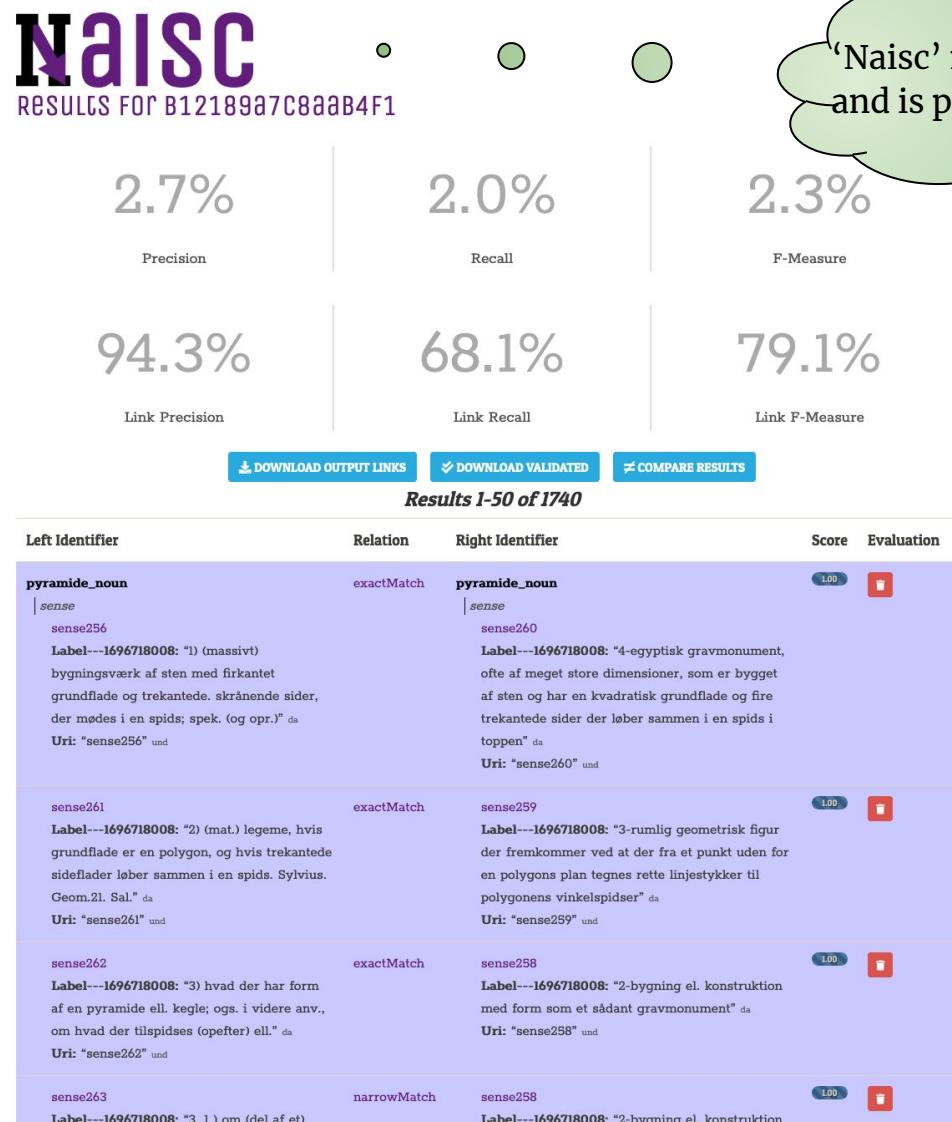
t-SNE visualization of *fish* (noun, verb) in BERT vector for each occurrence in the SemCor corpus (<http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor>)

## 4. Naisc

- Naisc is a system for aligning RDF datasets
- It takes as input 2 RDF documents
- It outputs an alignment (set of RDF triples) between these two documents



Openly available at  
<https://github.com/insight-centre/naisc>



## 5. Next steps?

- Cross-lingual word-sense alignment
- More robust techniques for semantic relationship detection across word senses and glosses
- Exploring other lexicographical information such as:
  - etymological data: **school** of fish vs. elementary **school**
  - cognates: *eau* (water, French) vs. *aw* (water, Kurdish) from \**wd-r/akʷ-* (PIE)
  - pronunciations: lead /li:d/ vs. lead /lɛd/
- Promoting interoperability among existing and future lexicographical resources, as in Linguistic Linked Open Data
- And many more...

# Merci beaucoup !



# References

- [1] Ahmadi, S., Atul Kr. Ojha, Banerjee S., McCrae, J. P. (September 2021). NUIG at TIAD 2021: Cross-lingual WordEmbeddings for Translation Inference. In Proceedings of the Translation Inference Across Dictionaries Workshop (TIAD 2021).
- [2] Ahmadi, S., McCrae, J. P., Nimb, S., Khan, F., Monachini, M., Pedersen, B. S., ... & Gabrovsek, D. (2020). A multilingual evaluation dataset for monolingual word sense alignment. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). European Language Resources Association (ELRA).
- [3] Arcan, M., Torregrosa, D., Ahmadi, S., & McCrae, J. P. (2019, May). Inferring translation candidates for multilingual dictionary generation with multi-way neural machine translation. In Proceedings of the Translation Inference Across Dictionaries Workshop (TIAD 2019).
- [4] Torregrosa, D., Arcan, M., Ahmadi, S., & McCrae, J. P. (2019). Tiad 2019 shared task: Leveraging knowledge graphs with neural machine translation for automatic multilingual dictionary generation. Translation Inference Across Dictionaries.
- [5] Miller, T., & Gurevych, I. (2014, May). WordNet—Wikipedia—Wiktionary: Construction of a Three-way Alignment. In LREC (pp. 2094-2100).
- [6] Jiang, W., Lin, Y., & Li, Y. (2018, April). Concept alignment of product taxonomies based on semantic similarity. In 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 517-521). IEEE.
- [7] Chakraborty, J., Bansal, S. K., Virgili, L., Konar, K., & Yaman, B. (2021, March). Ontoconnect: Unsupervised ontology alignment with recursive neural network. In Proceedings of the 36th Annual ACM Symposium on Applied Computing (pp. 1874-1882).
- [8] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.
- [9] Goddard, C. (2011). Semantic analysis: A practical introduction. Oxford University Press.
- [10] Pustejovsky, J. (1998). The generative lexicon. MIT press.
- [11] Westerhout, E. (2010). Definition extraction for glossary creation: a study on extracting definitions for semi-automatic glossary creation in Dutch. Netherlands Graduate School of Linguistics.
- [12] McCrae, J. P. (2018). Mapping wordnet instances to Wikipedia. In Proceedings of the 9th Global WordNet Conference (GWC 2018), pages 62–69.
- [13] 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.
- [14] 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.
- [15] McCrae, J. P., & Cillessen, D. (2021, January). Towards a Linking between WordNet and Wikidata. In Proceedings of the 11th Global Wordnet Conference (pp. 252-257).
- [16] 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.
- [17] Henrich, V., Hinrichs, E. W., & Suttn, K. (2012). Automatically Linking GermaNet to Wikipedia for Harvesting Corpus Examples for GermaNet Senses. J. Lang. Technol. Comput. Linguistics, 27(1), 1-19.
- [18] Ahmadi, S., & McCrae, J. P. (2021, January). Monolingual Word Sense Alignment as a Classification Problem. In Proceedings of the 11th Global Wordnet Conference (pp. 73-80).
- [19] Kernerian, I., Krek, S., McCrae, J. P., Gracia, J., Ahmadi, S., & Kabashi, B. (2020, May). Proceedings of the 2020 Globalex Workshop on Linked Lexicography. In Proceedings of the 2020 Globalex Workshop on Linked Lexicography.
- [20] Bajčetić, L., & Yim, S. B. (2020, May). Implementation of Supervised Training Approaches for Monolingual Word Sense Alignment: ACDH-CH System Description for the MWSA Shared Task at GlobaLex 2020. In Proceedings of the 2020 Globalex Workshop on Linked Lexicography (pp. 84-91).
- [21] Păiş, V., Tufiş, D., & Ion, R. (2020, May). MWSA Task at GlobaLex 2020: RACAI's Word Sense Alignment System using a Similarity Measurement of Dictionary Definitions. In Proceedings of the 2020 Globalex Workshop on Linked Lexicography (pp. 69-75).
- [22] Manna, R., Speranza, G., Di Buono, M. P., & Monti, J. (2020, May). UNIOR NLP at MWSA Task-GlobaLex 2020: Siamese LSTM with Attention for Word Sense Alignment. In Proceedings of the 2020 Globalex Workshop on Linked Lexicography (pp. 76-83).
- [23] Ahmadi, S., & McCrae, J. P. (2021, January). Monolingual Word Sense Alignment as a Classification Problem. In Proceedings of the 11th Global Wordnet Conference (pp. 73-80).
- [24] Artetxe, M., Labaka, G., & Agirre, E. (2017, July). Learning bilingual word embeddings with (almost) no bilingual data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 451-462).
- [25] Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. A. (2017). Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043.

