# Towards Automatic Linking of Lexicographic Data

The case of *Den Danske Ordbog* and *Ordbog over det danske Sprog* for the Danish language

Sina Ahmadi, Sanni Nimb, John P. McCrae, Nicolai H. Sørensen

XIX EURALEX International Congress
Alexandroupolis, Greece
September 2021

DEN DANSKE ORDBOG
MODERNE DANSK SPROG

Data Science Institute
Institiúid Eolaíochta Sonraí

OÉ Gaillimh
NUI Galway

Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS

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. **Danish Lexicographical Data**
3. **Word-Sense Alignment**
4. **Experiments with ODS and DDO**
5. **Next steps**

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

# 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*)

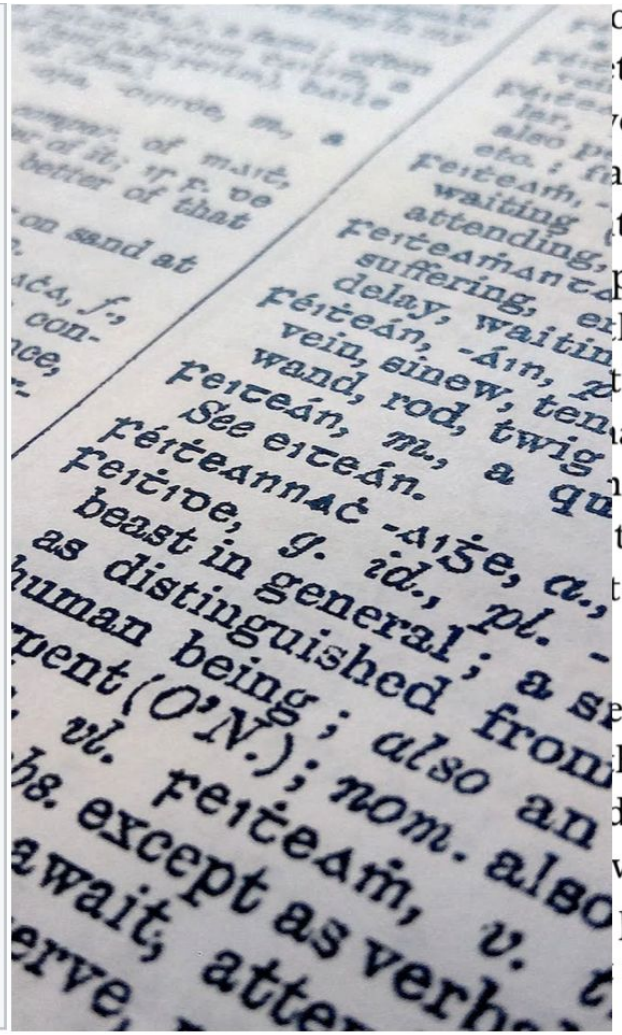

Some of the resources of Danish in the Society for Danish Language and Literature (DSL), Copenhagen

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

1. Context

# Lexical resources


Langenscheidt dictionaries



onstrued. In chapter 6, following Durkheim, we
ties, such as Australia, hold some events, places
e attacks on them as sacrilegious Durkheim posed
al an in
t the s.
plain d
he A se
the re
at cl of
n. In e,
the s t,
ts un d
es the er
he sa ns
d to e d
ward a
perp in
to erase from history knowledge of holocausts.

English

sacrilegious | ˌsakrɪ'lɪdʒəs | adjective
involving or committing sacrilege: *a sacrilegious act* | *it seems sacrilegious to say this, but it's really not that great a film.* more

English Thesaurus

sacrilegious adjective
*he condemned the book as a vicious, sacrilegious attack on their faith*: PROFANE, blasphemous, impious, sinful, irreverent, irreligious, godless, more

French - English

sacrilegious | BrE ˌsakrɪ'lɪdʒəs, AmE ˌsækrə'lɪdʒəs | adjective
Religion, figurative, humorous sacrilège

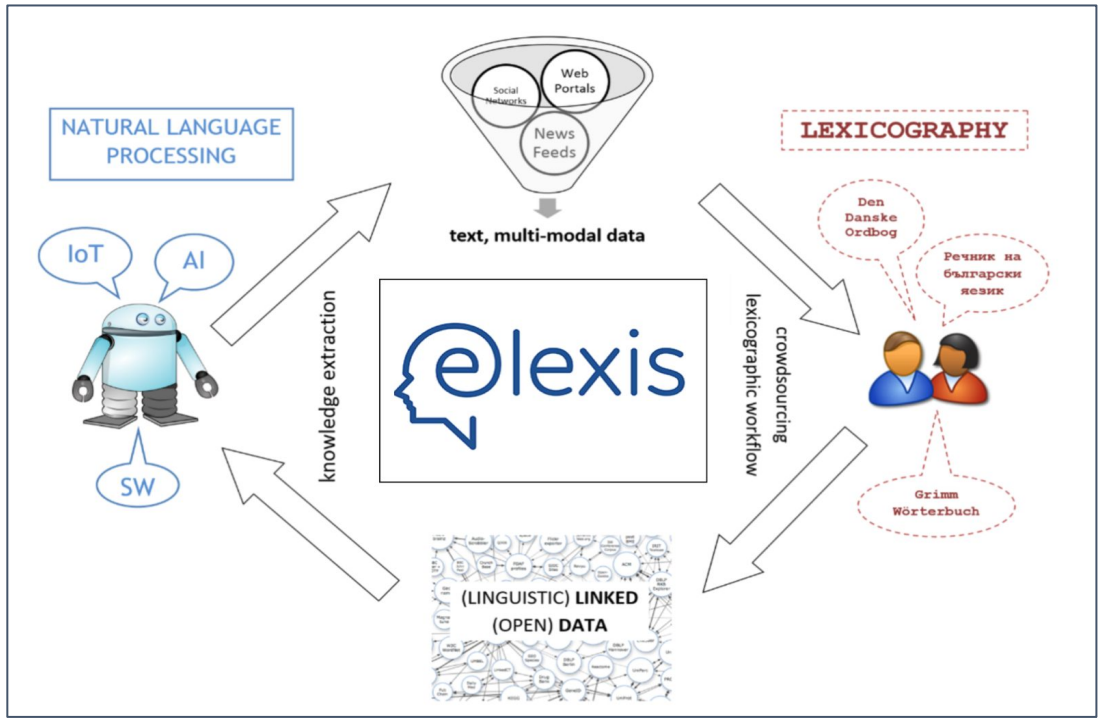
German - English

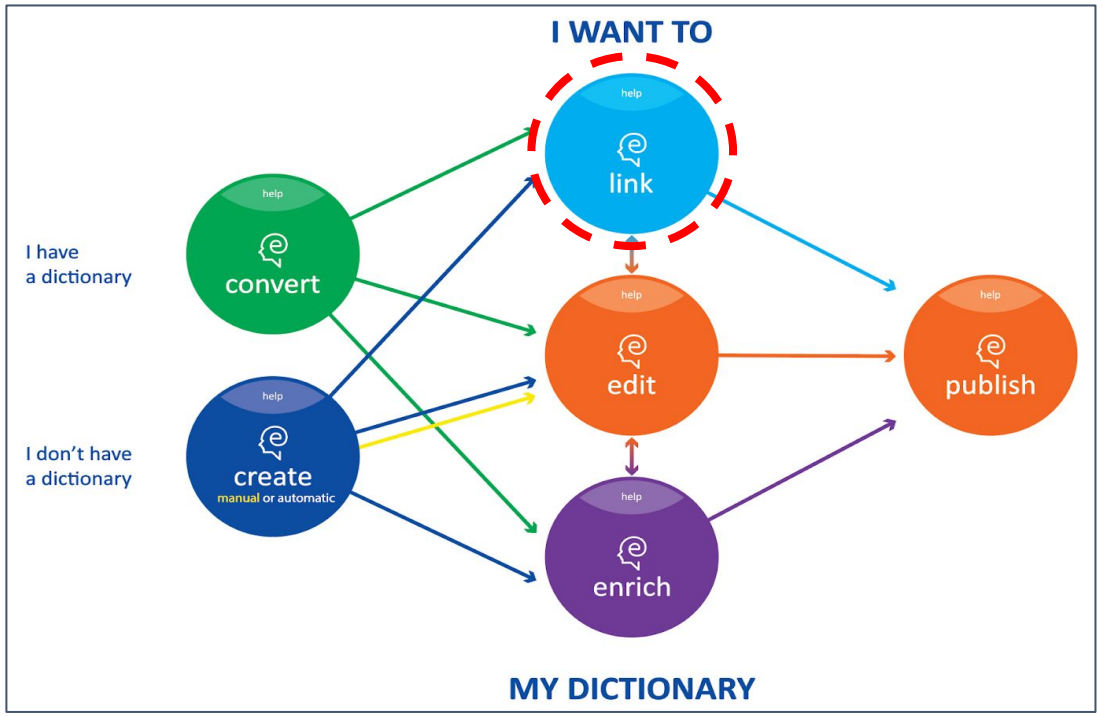
Dictionary Siri Knowledge

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

# 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/

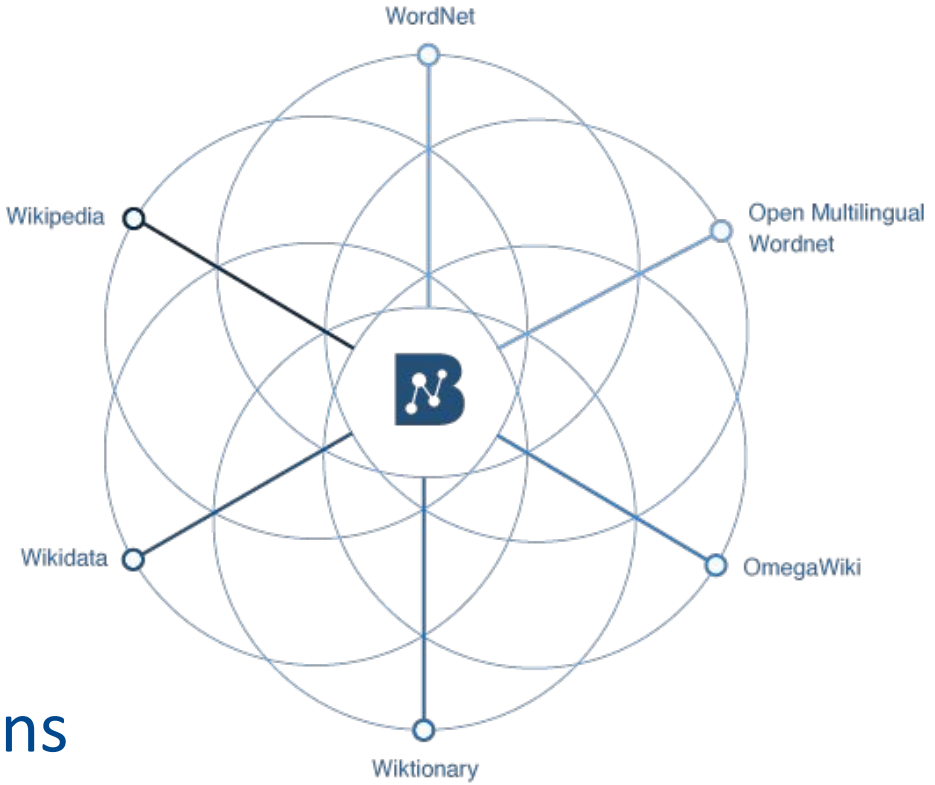


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

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.



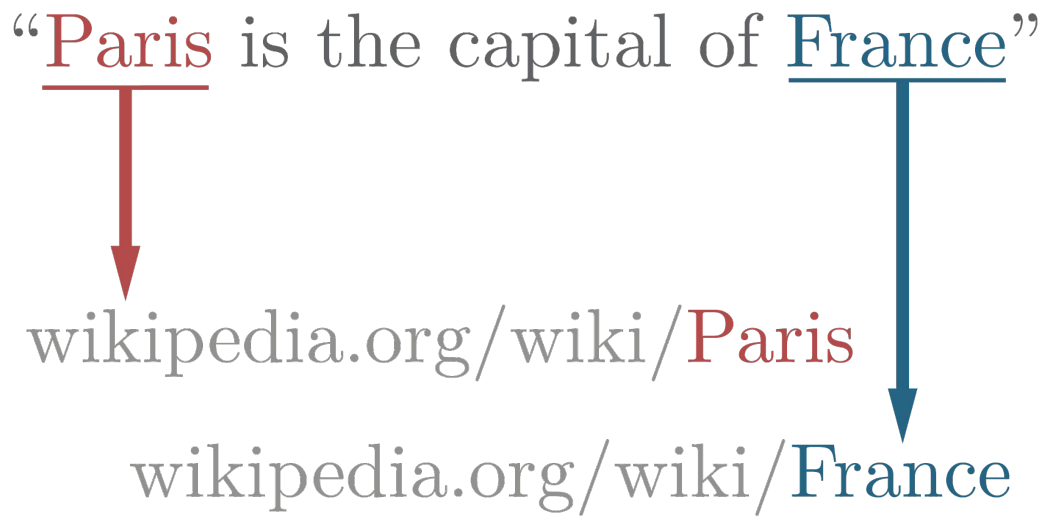

BabelNet (https://babelnet.org/)

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

1. Context

# A few applications

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

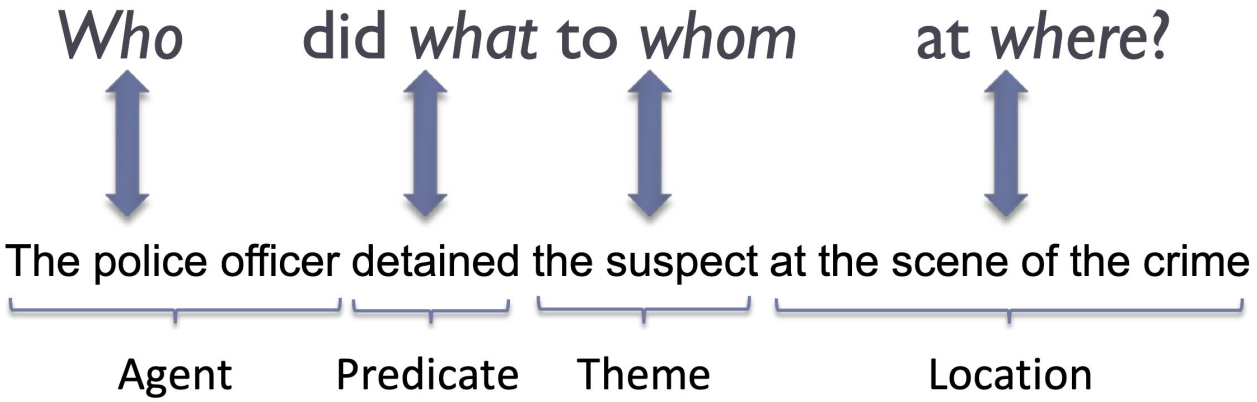



Source: https://en.wikipedia.org/wiki/Entity_linking

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

# A few applications

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



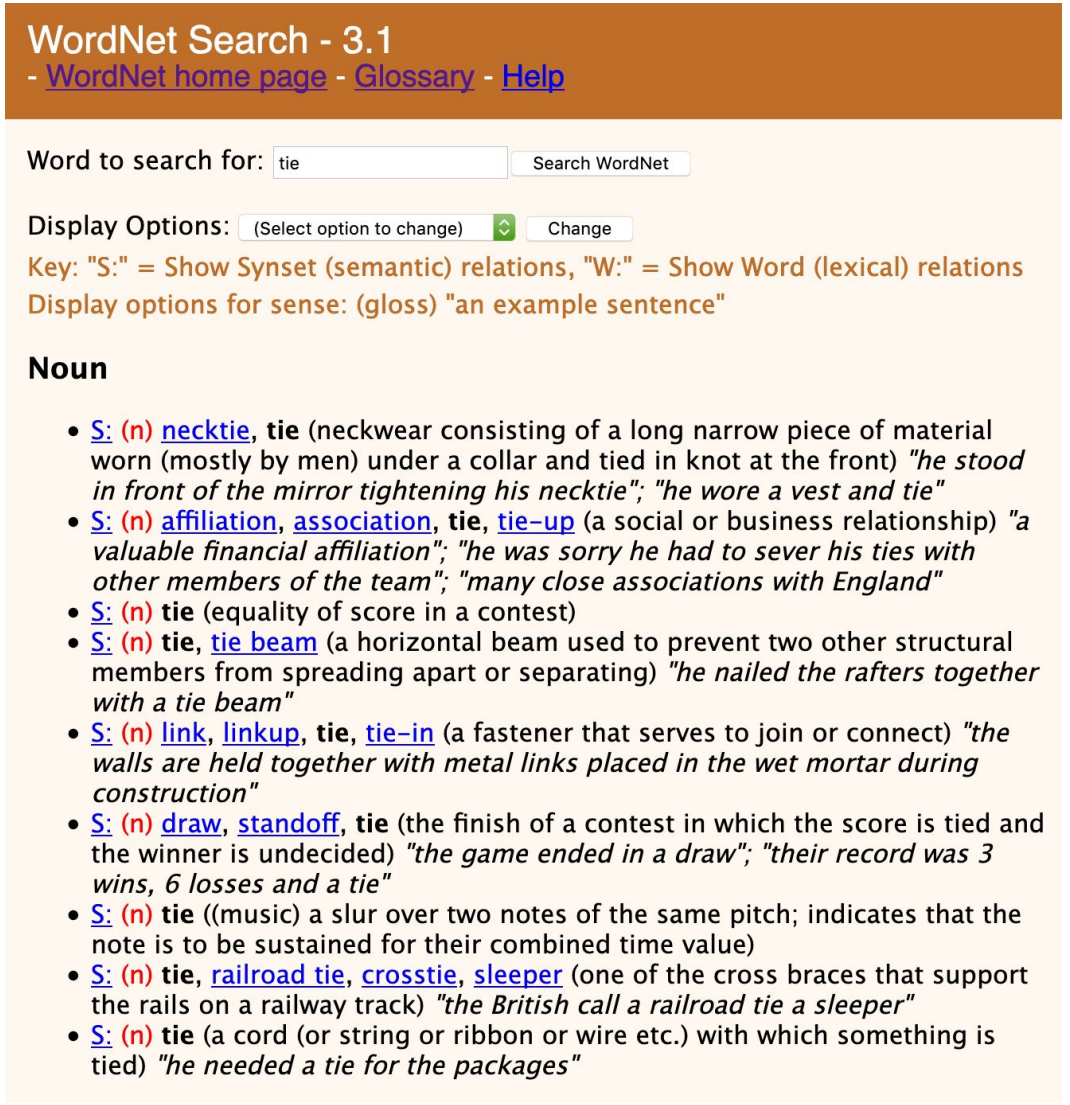

Source: https://web.stanford.edu/~jurafsky/slp3/slides/22_SRL.pdf

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

1. Context

# A few applications

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



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: tie Search WordNet

Display Options: (Select option to change) Change

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 note is to be sustained for their combined time value)
- S: (n) **tie**, railroad tie, crosstie, sleeper (one of the cross braces that support the rails on a railway track) *"the British call a railroad tie a sleeper"*
- S: (n) **tie** (a cord (or string or ribbon or wire etc.) with which something is tied) *"he needed a tie for the packages"*

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

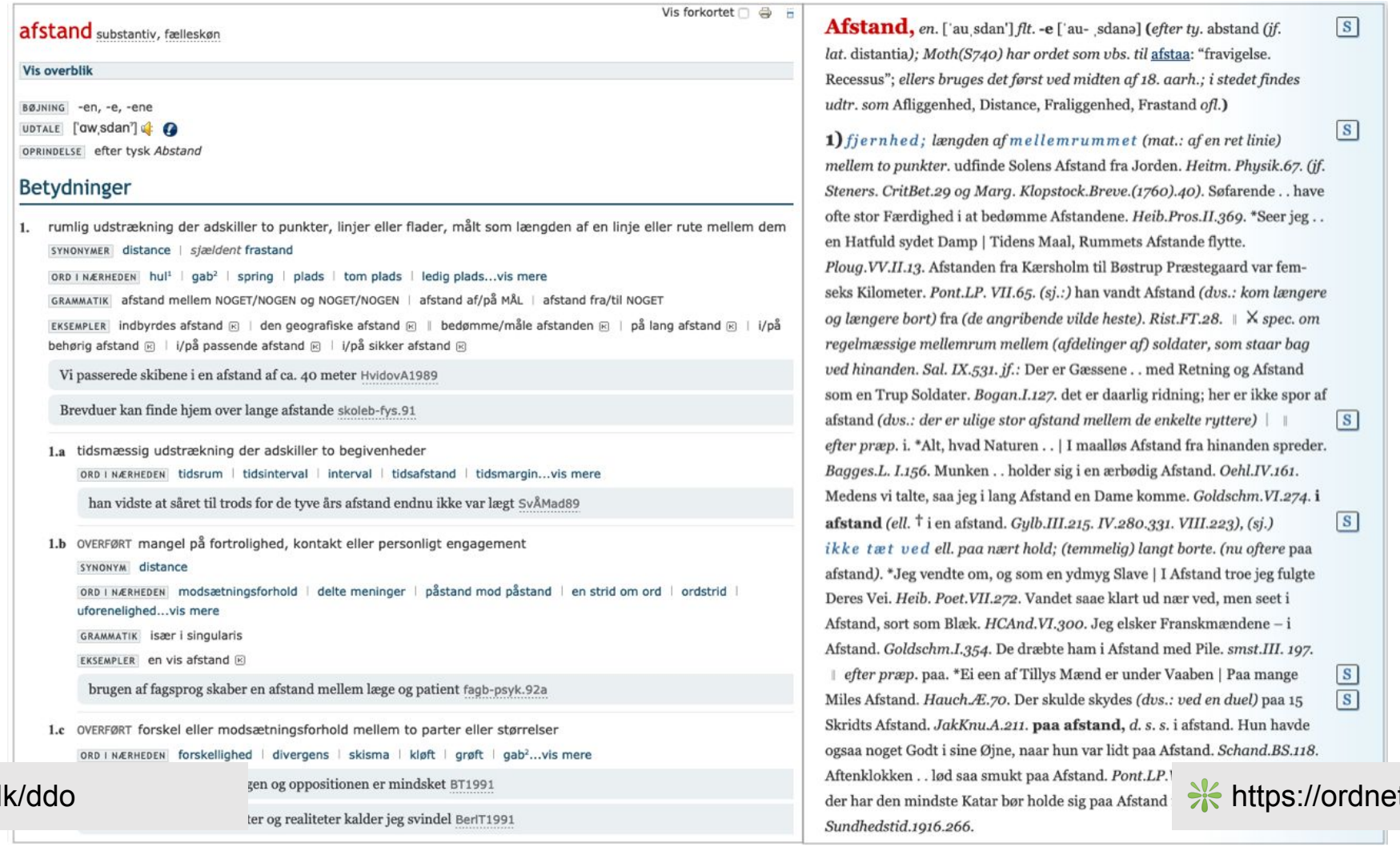
# 2. Danish lexicographical resources

- Lexicographic resources at DSL
  - DanNet - the Danish WordNet
  - Ordbog over det danske Sprog (ODS) (open access: https://ordnet.dk/ods)
  - Den Danske Ordbog (DDO) (open access: https://ordnet.dk/ddo)
  - The Danish Thesaurus
  - Danish FrameNet (aligned with DDO)
- Our focus is on on linking ODS and DDO at **sense level**

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

## 2. Danish lexicographical resources

# ODS vs. DDO



**afstand** substantiv, fælleskøn

Vis overblik

BØJNING -en, -e, -ene
UDTALE ['ɑw,sdan']
OPRINDELSE efter tysk *Abstand*

### Betydninger

1. rumlig udstrækning der adskiller to punkter, linjer eller flader, målt som længden af en linje eller rute mellem dem
   SYNONYMER distance | *sjældent* frastand
   ORD I NÆRHEDEN hul[1] | gab[2] | spring | plads | tom plads | ledig plads...vis mere
   GRAMMATIK afstand mellem NOGET/NOGEN og NOGET/NOGEN | afstand af/på MÅL | afstand fra/til NOGET
   EKSEMPLER indbyrdes afstand | den geografiske afstand | bedømme/måle afstanden | på lang afstand | i/på behørig afstand | i/på passende afstand | i/på sikker afstand
   Vi passerede skibene i en afstand af ca. 40 meter HvidovA1989
   Brevduer kan finde hjem over lange afstande skoleb-fys.91

   1.a tidsmæssig udstrækning der adskiller to begivenheder
   ORD I NÆRHEDEN tidsrum | tidsinterval | interval | tidsafstand | tidsmargin...vis mere
   han vidste at såret til trods for de tyve års afstand endnu ikke var lægt SvÅMad89

   1.b OVERFØRT mangel på fortrolighed, kontakt eller personligt engagement
   SYNONYM distance
   ORD I NÆRHEDEN modsætningsforhold | delte meninger | påstand mod påstand | en strid om ord | ordstrid | uforenelighed...vis mere
   GRAMMATIK især i singularis
   EKSEMPLER en vis afstand
   brugen af fagsprog skaber en afstand mellem læge og patient fagb-psyk.92a

   1.c OVERFØRT forskel eller modsætningsforhold mellem to parter eller størrelser
   ORD I NÆRHEDEN forskellighed | divergens | skisma | kløft | grøft | gab[2]...vis mere
   gen og oppositionen er mindsket BT1991
   ter og realiteter kalder jeg svindel BerlT1991

**Afstand,** *en.* ['au,sdan'] *flt.* **-e** ['au- ,sdanə] **(***efter ty.* abstand *(jf. lat.* distantia*); Moth(S740) har ordet som vbs. til* afstaa: "fravigelse. Recessus"; *ellers bruges det først ved midten af 18. aarh.; i stedet findes udtr. som* Afliggenhed, Distance, Fraliggenhed, Frastand *ofl.)*

**1)** *fjernhed; længden af mellemrummet (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 mellemrum 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 ryttere)* | ‖ *efter præp.* i. *Alt, hvad Naturen . . | I maalløs Afstand fra hinanden spreder. *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.I.354.* De dræbte ham i Afstand med Pile. *smst.III. 197.* ‖ *efter præp.* paa. *Ei een 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.* der har den mindste Katar bør holde sig paa Afstand *Sundhedstid.1916.266.*

https://ordnet.dk/ddo

https://ordnet.dk/ods

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

2. Danish lexicographical resources

# ODS vs. DDO

## ODS

- Historical Danish dictionary (1700-1955)
- Richly-described entries
- Retrodigitized ⟶ noisy senses (both, in printed and XML versions)
- Senses not linked to any external resource
- Obsolete senses


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

# 2. Danish lexicographical resources

# *afstand* [ˈɑwˌsdanˀ] (distance in Danish)*

**1)** *fjernhed; længden af mellemrummet (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 mellemrum 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 ryttere)* ‖ ‖ *efter præp.* i. *Alt, hvad Naturen . . | I maalløs Afstand fra hinanden spreder. *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.I.354.* De dræbte ham i Afstand med Pile. *smst.III. 197.* ‖ *efter præp.* paa. *Ei een 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.*

<Semdel SemID="29014340"><Semem BetNo="1" SememID="29401382" odsID="Afstand_1"><dotPlus id="2666"/><SemIndhold><dotBetNo>1)</dotBetNo> <dotSpaced>fjernhed;</dotSpaced> længden af <dotSpaced>mellemrummet</dotSpaced><dotLn col="0293" lin="25" orig="mellemrum-met"/>(mat.: af en ret linie) mellem to punkter. <dotLn col="0293" lin="26"/>udfinde Solens Afstand fra Jorden. Heitm. <dotLn col="0293" lin="27"/>Physik.67. (jf. Steners. CritBet.29 og Marg. <dotLn col="0293" lin="28"/>Klopstock.Breve.(1760).40). Søfarende . . <dotLn col="0293" lin="29"/>have ofte stor Færdighed i at bedømme <dotLn col="0293" lin="30"/>Afstandene. Heib.Pros.II.369. <dotRaised>*</dotRaised>Seer jeg <dotLn col="0293" lin="31"/>. . en Hatfuld sydet Damp | Tidens Maal, <dotLn col="0293" lin="32"/>Rummets Afstande flytte. Ploug.VV.II.13. <dotLn col="0293" lin="33"/>Afstanden fra Kærsholm til Bøstrup Præstegaard<dotLn col="0293" lin="34" orig="Præste-gaard"/>var fem-seks Kilometer. Pont.LP. <dotLn col="0293" lin="35"/>VII.65. (sj.:) han vandt Afstand (dvs.: kom <dotLn col="0293" lin="36"/>længere og længere bort) fra (de angribende <dotLn col="0293" lin="37"/>vilde heste). Rist.FT.28. <Planke PlankeID="29900969"><dotEdBreak/> <piktogram src="swords" title="sværd"/> spec. om regelmæssige<dotLn col="0293" lin="38" orig="regel-mæssige"/>mellemrum mellem (afdelinger af) <dotLn col="0293" lin="39"/>soldater, som staar bag ved hinanden. Sal. <dotLn col="0293" lin="40"/>IX.531. jf.: Der er Gæssene . . med Retning<dotLn col="0293" lin="41" orig="Ret-ning"/>og Afstand som en Trup Soldater. <dotLn col="0293" lin="42"/>Bogan.I.127. det er daarlig ridning; her <dotLn col="0293" lin="43"/>er ikke spor af afstand (dvs.: der er ulige stor <dotLn col="0293" lin="44"/>afstand mellem de enkelte ryttere) <piktogram src="dotpipe" title="dobbeltbrudt streg"/> </Planke><Planke PlankeID="29900970"><dotPlus id="2667"/><dotEdBreak/> efter <dotLn col="0293" lin="45"/>præp. i. <dotRaised>*</dotRaised>Alt, hvad Naturen . . | I maalløs <dotLn col="0293" lin="46"/>Afstand fra hinanden spreder. Bagges.L. <dotLn col="0293" lin="47"/>I.156. Munken . . holder sig i en ærbødig <dotLn col="0293" lin="48"/>Afstand. Oehl.IV.161. Medens vi talte, saa <dotLn col="0293" lin="49"/>jeg i lang Afstand en Dame komme. Goldschm.VI.274.<dotLn col="0293" lin="50" orig="Gold-schm.VI.274."/>i afstand (ell. <piktogram src="cross" title="kors"/> i en afstand. <dotLn col="0293" lin="51"/>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 paa afstand). <dotRaised>*</dotRaised>Jeg <dotLn col="0293" lin="54"/>vendte om, og som en ydmyg Slave | I <dotLn col="0293" lin="56"/>Poet.VII.272. Vandet sa i Afstand, sort som Blæk. <dotLn col="0293" lin="58"/>0293" lin="59" orig="Franskmæn-dene"/>– i Afstand. ham i Afstand med Pile. smst.III. <dotLn col="0293" <dotPlus id="2669"/><dotPlus id="2670" kontrol="ingen efter præp. paa. <dotRaised>*</dotRaised>Ei Vaaben | Paa mange <dotLn col="0293" lin="63"/>Miles lin="64"/>skydes (dvs.: ved en duel) paa 15 Sk JakKnu.A.211. paa afstand, d. s. s.</ noget Godt <dotLn col="0294" lin="01"/>i sine Øjne, na i>Schand.BS.118. Aftenklokken . . lød saa <dotLn  Enhver<dotLn col="0294" lin="04" orig="En-hver"/> "/>bør holde sig paa Afstand fra Børnene. <dotLn col="0294" lin="07"/></Planke></SemIndhold></Semem></Semde

* 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.

2. Danish lexicographical resources

# ODS vs. DDO

DDO

- modern Danish dictionary 1955-->
- DSL's direct follower to ODS --> many lemmas and senses in common
- Well-structured sense descriptions
- ~50% linked to DanNet, ~90% to The Danish Thesaurus at sense level


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

# 3. Word sense alignment (WSA)

## → linking lexical content at sense level, including glosses

**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).

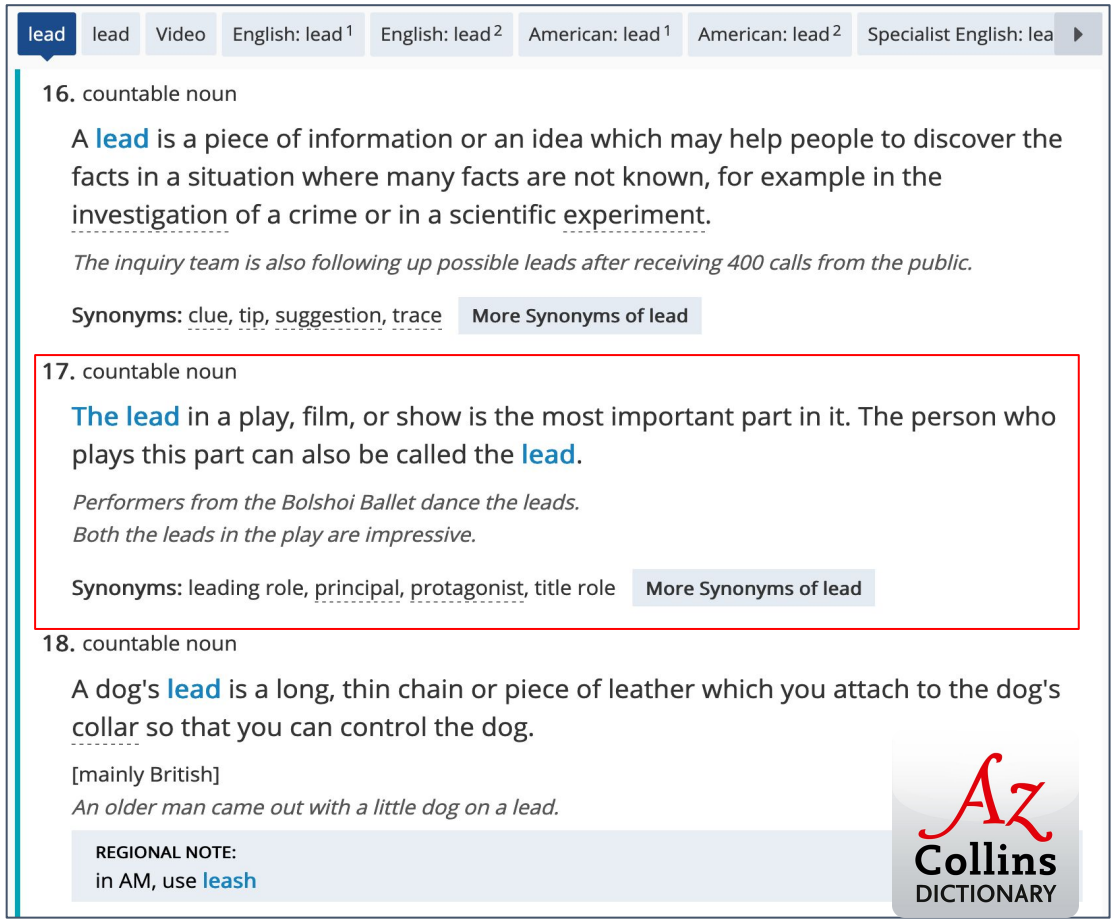
**the male/female lead**

They were having trouble casting the female lead.

the film's **romantic lead**

**7** → **lead singer/guitarist etc**

Longman



lead | lead | Video | English: lead 1 | English: lead 2 | American: lead 1 | American: lead 2 | 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

Az Collins DICTIONARY

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

3. 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)

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

3. 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

A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment

Sina Ahmadi*, John P. McCrae*,
Sanni Nimb[1], Fahad Khan[3], Monica Monachini[3], Bolette S. Pedersen[8], Thierry Declerck[2,12], Tanja Wissik[2], Andrea Bellandi[3], Irene Pisani[4], Thomas Troelsgård[1], Sussi Olsen[8], Simon Krek[5], Veronika Lipp[6], Tamás Váradi[6], László Simon[6], András Győrffy[6], Carole Tiberius[9], Tanneke Schoonheim[9], Yifat Ben Moshe[10], Maya Rudich[10], Raya Abu Ahmad[10], Dorielle Lonke[10], Kira Kovalenko[11], Margit Langemets[13], Jelena Kallas[13], Oksana Dereza[7], Theodorus Fransen[7], David Cillessen[7], David Lindemann[14], Mikel Alonso[14], Ana Salgado[15], José Luis Sancho[16], Rafael-J. Ureña-Ruiz[16], Jordi Porta Zamorano[16], Kiril Simov[17], Petya Osenova[17], Zara Kancheva[17], Ivaylo Radev[17], Ranka Stanković[18], Andrej Perdih[19], Dejan Gabrovšek[19]
*Insight Centre for Data Analytics, National University of Ireland, Galway
{sina.ahmadi,john.mccrae}@insight-centre.org
(other affiliations in Appendix A)

**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 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

**1. Introduction**

Lexical semantic resources (LSRs) are knowledge repositories that provide the vocabulary of a language in a descriptive and structured way. One of the famous examples of LSRs are dictionaries. Dictionaries form an important foundation 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[1], we have developed a new baseline that covers 15 languages and will provide a new baseline for the task of monolingual word sense alignment.

Different dictionaries and related resources such as wordnets and encyclopedia 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), semantic 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 Wikipedia[2], 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 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 for each pair of senses in the two resources 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-

* Contact Authors
[1] https://globalex2020.globalex.link/
[2] https://www.wikipedia.org

[2]

→ 17 manually-annotated monolingual resources for the task of WSA covering 15 languages, including **Danish**



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

# Monolingual WSA datasets

| | | | | |
|---|---|---|---|---|
| **elocutionary (adjective)** | | | | |
| | (used of style of speaking) overly embellished | NONE | | 0\|-pertaining to elocution. |
| | of or relating to elocution | exact | 0\|-pertaining to elc | |
| **montgolfier (noun)** | | | | |
| | French inventor who (with his brother Josef Michel | NONE | 0\|-a balloon which | 0\|-a balloon which ascends by the buoyancy |
| | French inventor who (with his brother Jacques Etie | NONE | 0\|-a balloon which | |
| **dice (verb)** | | | | |
| | cut into cubes | exact | 1.\|-to cut into smal | 1.\|-to cut into small cubes; . |
| | play dice | NONE | | 2.\|-to ornament with squares, diamonds, or |
| **ebb (verb)** | | | | |
| | flow back or recede | exact | 0\|-to cause to flow | 0\|-to cause to flow back. |
| | fall away or decline | NONE | | |
| | hem in fish with stakes and nets so as to prevent th | NONE | | |
| **educated (adjective)** | | | | |
| | characterized by full comprehension of the problem | | | 0\|-formed or developed by education; . |
| | possessing an education (especially having more th | exact | | |
| **quaver (verb)** | | narrower | | |
| | give off unsteady sounds, alternating in amplitude | broader | | 0\|-to utter with quavers. |
| | sing or play with trills, alternating with the half note | related | | |
| **rangy (adjective)** | | NONE | | |
| | adapted to wandering or roaming | | | 0\|-inclined or able to range, or rove about, fo |
| | allowing ample room for ranging | | | |
| | tall and thin and having long slender limbs | | | |
| **smiler (noun)** | | | | |
| | a person who smiles | | | 0\|-one who smiles. |
| | the human face (`kisser' and `smiler' and `mug' are | | | |
| **chair (noun)** | | | | |


```
{
    "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"}
    ]
}
```




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

3. Word sense alignment

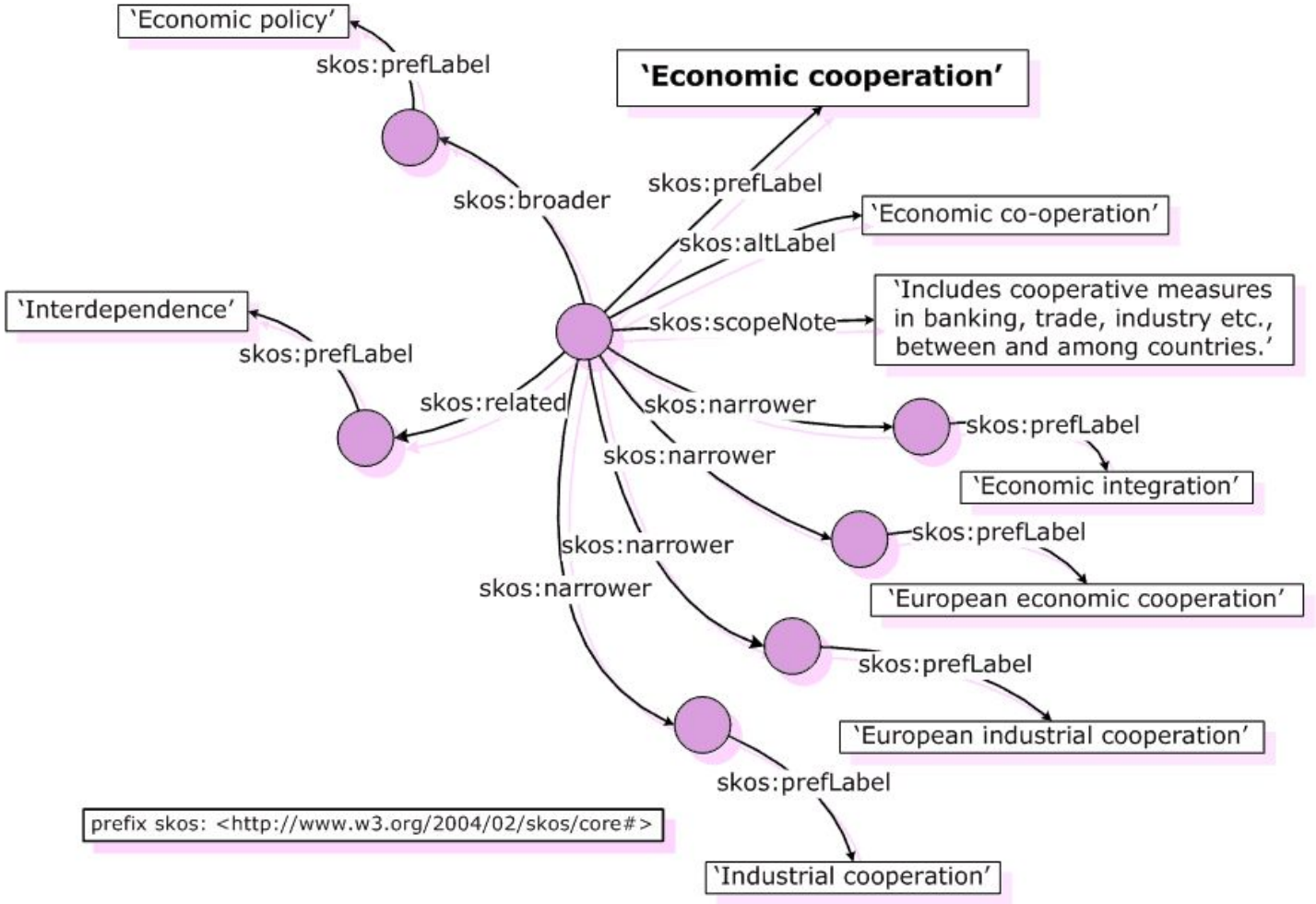
# Semantic relations (based on SKOS*)

- **`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/

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

# Semantic relations: an example




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

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

# Manual alignment of ODS-DDO senses, examples

- **`ODS exact DDO`**: noun *passager* ('passenger'): ODS 'person traveling with mail coach etc.', DDO: 'person traveling with private or public means of transportation'.
- **`ODS broader than DDO`**: noun *værge* ('guardian'): ODS 'a guardian of anything or anybody', DDO 'a guardian in legal context'.
- **`ODS narrower than DDO`**: adjective *spids* ('sharp'): ODS two narrower senses, one about a sound and another one about a smell → DDO 'pungent in an unpleasant way (about smell, taste or sound)'.
- **`ODS related to DDO`**: noun *søvn* ('sleep'): ODS 'being able to sleep', DDO 'the state of sleeping'

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

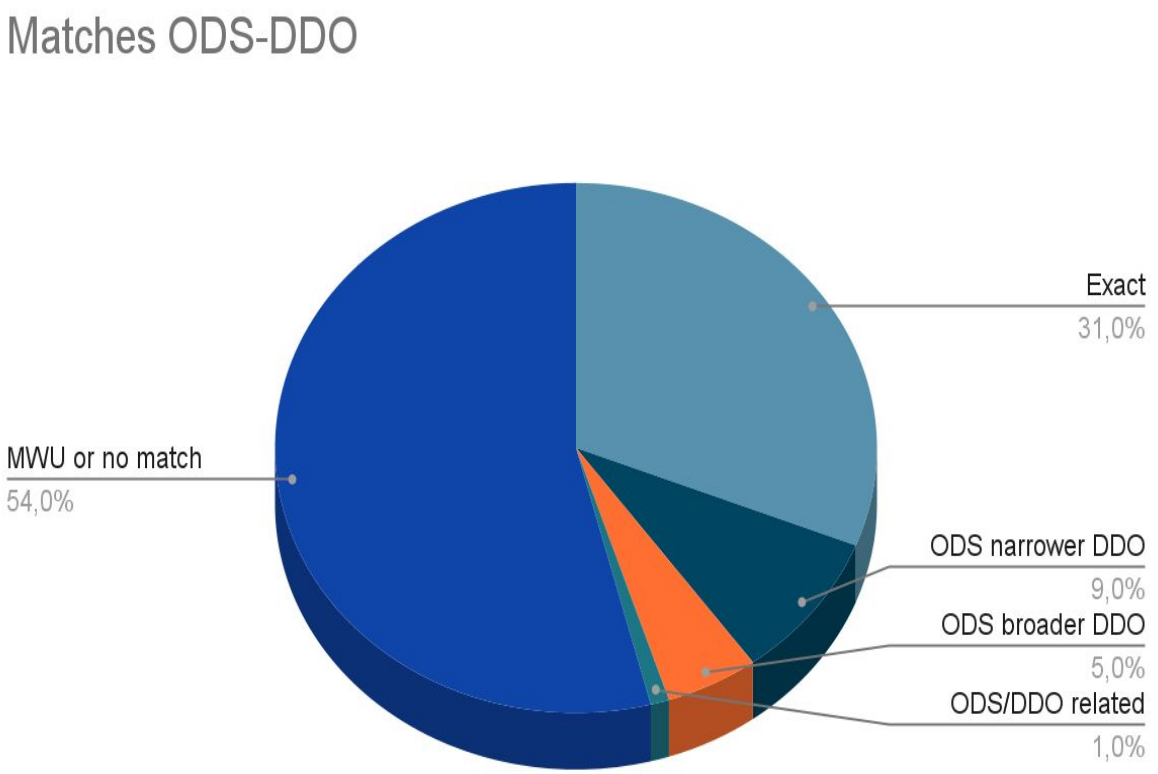
3. Word sense alignment

# Manual alignment of ODS-DDO senses

- 380 nouns, 93 verbs, 63 adjectives.
- ODS 3,595 senses, DDO 1,667 senses
  → 1,000 exact matches

- Polysemy: average DDO 3.11 senses per lemma (ODS 6.7)

- Difference in dictionary structure:
  - ODS senses include MWU
  - DDO senses: no MWU
  - ODS: heading main senses
  - DDO: no heading main senses




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

| Language | Resource | Nouns | Verbs | Adjectives | Adverbs | Other | All |
|---|---|---|---|---|---|---|---|
| Basque | Basque Wordnet | 929 (6836) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 929 (6836) |
| | *Euskal Hiztegia* | 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 | *Ordbog over det danske Sprog* | 2176 (282040) | 983 (119163) | 436 (60599) | 0 (0) | 0 (0) | 3595 (461802) |
| | *Den Danske Ordbog* | 1036 (12326) | 383 (4045) | 248 (2228) | 0 (0) | 0 (0) | 1667 (18599) |
| Dutch | *Woordenboek der Nederlandsche Taal* | 1459 (28979) | 405 (5185) | 527 (7878) | 106 (2662) | 0 (0) | 2497 (44704) |
| | *Algemeen Nederlands Woordenboek* | 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) |
| | *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 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 | *An Foclóir Beag* | 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) |
| | *Kostelski slovar* | 151 (1050) | 61 (308) | 45 (257) | 0 (0) | 38 (263) | 295 (1878) |
| Spanish | *Diccionario de la lengua española* | 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 | *Dicionário da Língua Portuguesa Contemporânea* | 285 (4060) | 58 (686) | 110 (1287) | 9 (143) | 1 (9) | 463 (6185) |
| | *Dicionário Aberto* | 199 (1521) | 53 (203) | 67 (372) | 3 (15) | 1 (5) | 323 (2116) |
| Russian | *Ozhegov-Shvedova* | 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) |

elexis european lexicographic infrastructure

## 3. 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



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

# 4. Experiments with ODS and DDO

→ linking lexical content at sense level, including glosses

In order to predict the semantic similarity between two given senses, we trained various models based on the following similarity features:

1. String metrics
    - Longest common substring
    - Length ratio
    - Average word length ratio
    - Jaccard, Dice, and Containment
2. Word Embeddings (based on cosine similarity)
3. Automatic feature extraction

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

## 4. Experiments with ODS and DDO

# 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


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

## 4. Experiments with ODS and DDO

# Results

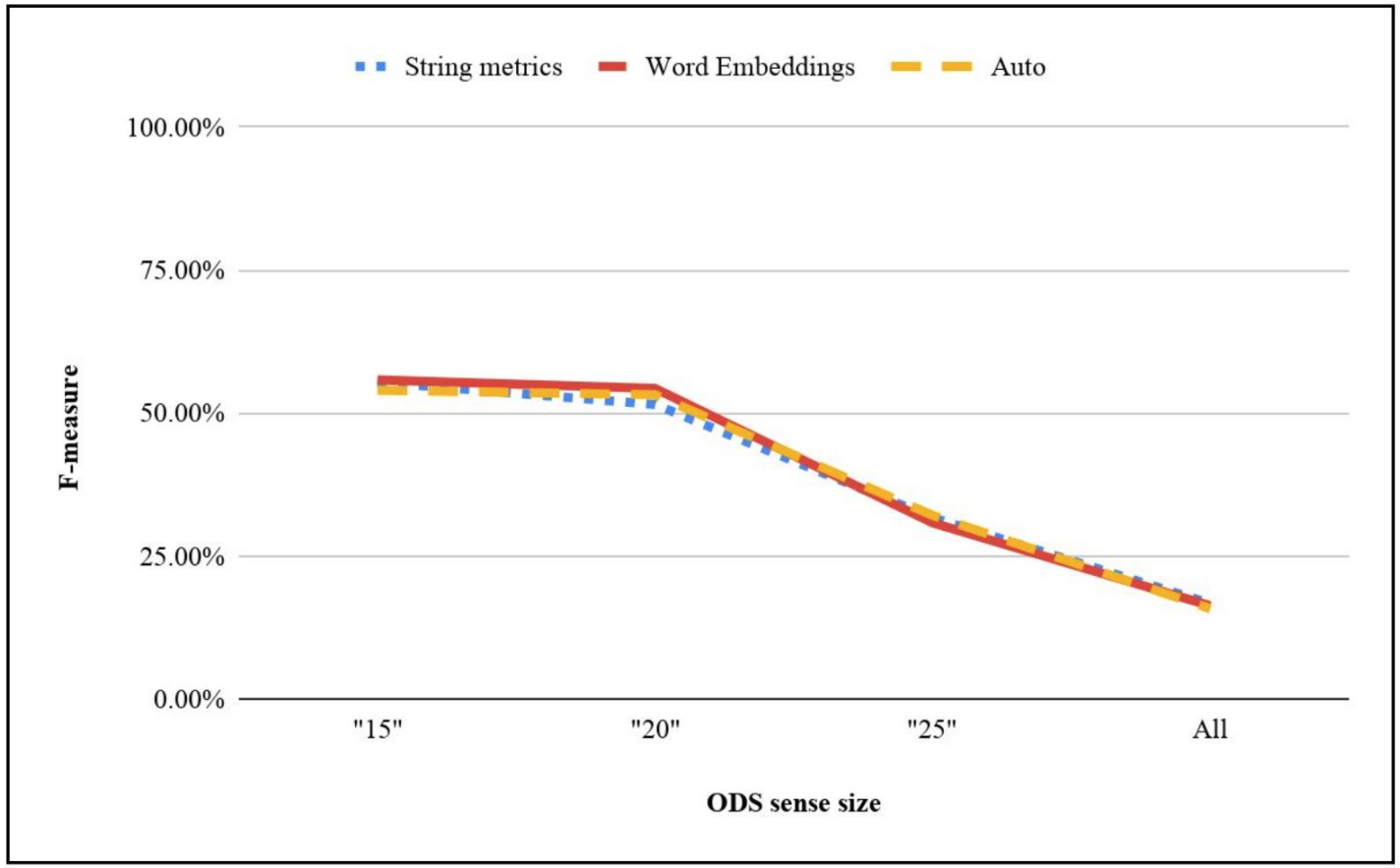
Performance of our similarity detection models for automatic alignment of DDO and ODS within a specific limit of space-separated tokens (15, 20, 25 and all tokens)

| ODS sense size | Model | Precision | Recall | F-measure |
|---|---|---|---|---|
| 15 | String metrics | 65.3% | **48.1%** | 55.4% |
| | Word Embeddings | **66.7%** | 48.0% | **55.8%** |
| | Auto | 64.0% | 46.6% | 54.0% |
| 20 | String metrics | 61.5% | 44.3% | 51.5% |
| | Word Embeddings | 64.7% | 46.7% | 54.3% |
| | Auto | 63.3% | 45.8% | 53.2% |
| 25 | String metrics | 57.5% | 21.9% | 31.7% |
| | Word Embeddings | 55.9% | 21.2% | 30.8% |
| | Auto | 58.5% | 22.2% | 32.1% |
| All | String metrics | 54.7% | 9.8% | 16.7% |
| | Word Embeddings | 50.7% | 9.7% | 16.3% |
| | Auto | 50.3% | 9.4% | 15.8% |

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

## 4. Experiments with ODS and DDO

# Results




The correlation of sense sizes in ODS with F-measure using various methods

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

# Future steps

- Text paraphrasing: converting ODS senses into more representative ones by
- Removing excessive descriptions, e.g. historical citations
- Lexical and orthographic normalization
- Paraphrasing techniques

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