



SPRÅKBANKENTEXT



Change Is Key!



UNIVERSITÉ  
DE GENÈVE



# Revealing semantic variation in Swedish using computational models of semantic proximity: results from lexicographical experiments

Emma Sköldberg, Shafqat Mumtaz Virk, Pauline Sander,  
Simon Hengchen & Dominik Schlechtweg

Euralex, October 2024

# Outline of the presentation

- ‘The Contemporary Dictionary of the Swedish Academy’ (SO) – in short
- The meaning descriptions in the SO – traditions
- The DUREl Tool
- Two experiments:
  - tool and corpora
  - results
- Summary and discussion

# 'The Contemporary Dictionary of the Swedish Academy' (SO)

65 000 headwords  
100 000 different senses



A screenshot of the Swedish Academy's online dictionary website. At the top, it says 'Svenska Akademiens ORDBÖCKER' and 'HEM OM HJÄLP GRAMMATIK KONTAKT'. Below that is a search bar with the text 'Sök i tre ordböcker på en gång' and a search button. Three dictionary entries are displayed: SAOL (published 2015), SO (published 2015), and SAOB (published 1992). The SO entry is circled in red. The SO entry includes the definition of 'ordbok' and a list of related terms like 'ordboks', 'ordboken', and 'ordbokers'. It also mentions 'Sammansättningsord' and 'Exempel'.

<https://svenska.se>

# The meaning descriptions in SO: traditions

SO is a subset of a extensive lexical database which has been developed since the 1970s. The SO-lexicographers follow principles described in e.g. Ralph et al. (1977) and Järborg (1989), e.g.:

- the senses in SO are ordered hierarchically with main senses and subsenses (e.g. figurative uses)
- the distinction into different senses is relatively “fine-grained”. The SO lexicographers are rather splitters than lumpers.

## **SOME QUESTIONS WITHIN THE DICTIONARY PROJECT:**

Are the semantic descriptions of the headwords up to date?

Have the sense of the headwords developed in some way since the 2nd edition (2021)?

The SO lexicographers' observations regarding sense development is, traditionally, based on manual work. Computer-aided methods will streamline and strengthen the editorial work.

# Tool: DUREl

- The DUREl annotation tool (<https://durel.ims.uni-stuttgart.de/>)
- Automate the measurement of **semantic proximity** between word usages
- Judgments populate weighted graphs and are **clustered**
- Clusters correlate with lexicographic **word senses**



DUREl


# Sample Corpus

- |   |      |   |   |
|---|------|---|---|
| A | 1824 | and taking a knife from her pocket, she opened a vein in her little <b>arm</b> ,                    | 😊 |
| B | 1842 | And those who remained at home had been heavily taxed to pay for the <b>arms</b> , ammunition;      | ✘ |
| C | 1860 | and though he saw her within reach of his <b>arm</b> , yet the light of her eyes seemed as far off  | 😊 |
|   |      | ...   |   |
| D | 1953 | overlooking an <b>arm</b> of the sea which, at low tide, was a black and stinking mud-flat          | 🍷 |
| E | 1975 | twelve miles of coastline lies in the southwest on the Gulf of Aqaba, an <b>arm</b> of the Red Sea. | 🍷 |
| F | 1985 | when the disembodied <b>arm</b> of the Statue of Liberty jets spectacularly out of the              | 😊 |

# Usage pair

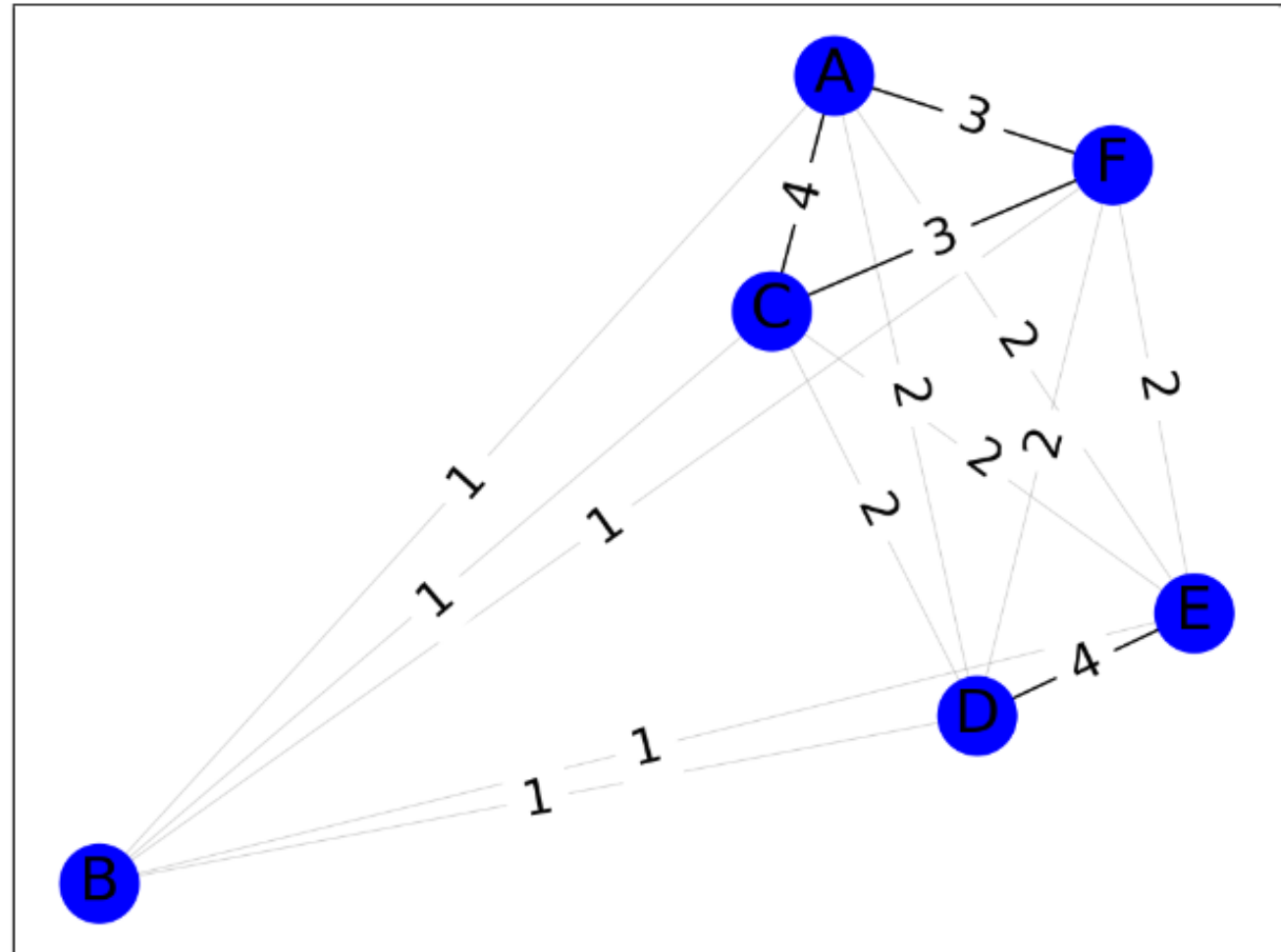
- (A) [...] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her. 😊
- (D) It stood behind a high brick wall, its back windows overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat [...]. 🙄

# Semantic Proximity/Relatedness

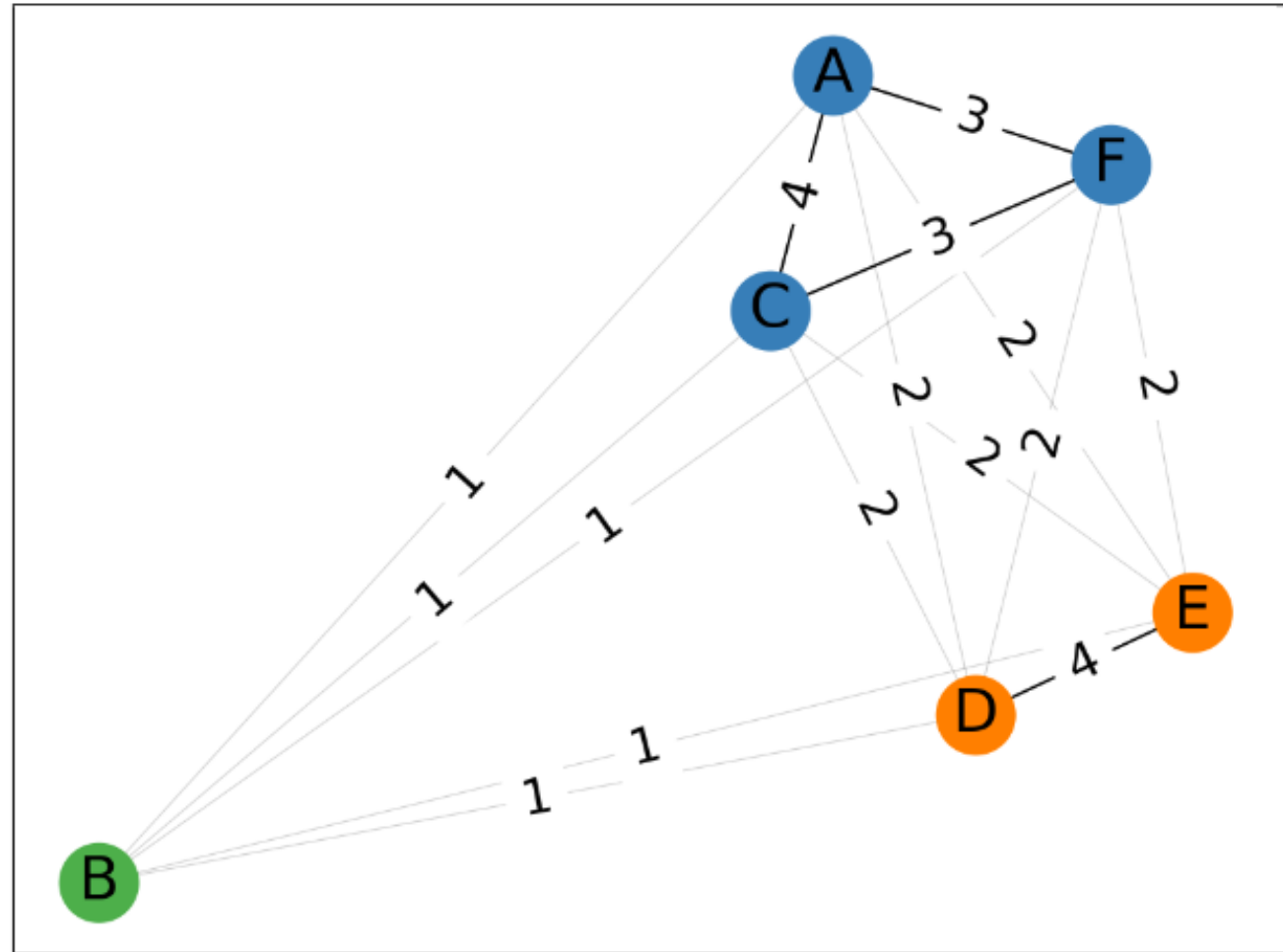
- 
- 4: Identical
  - 3: Closely Related
  - 2: Distantly Related
  - 1: Unrelated



# Graph Representation



# Clustering



# Corpora

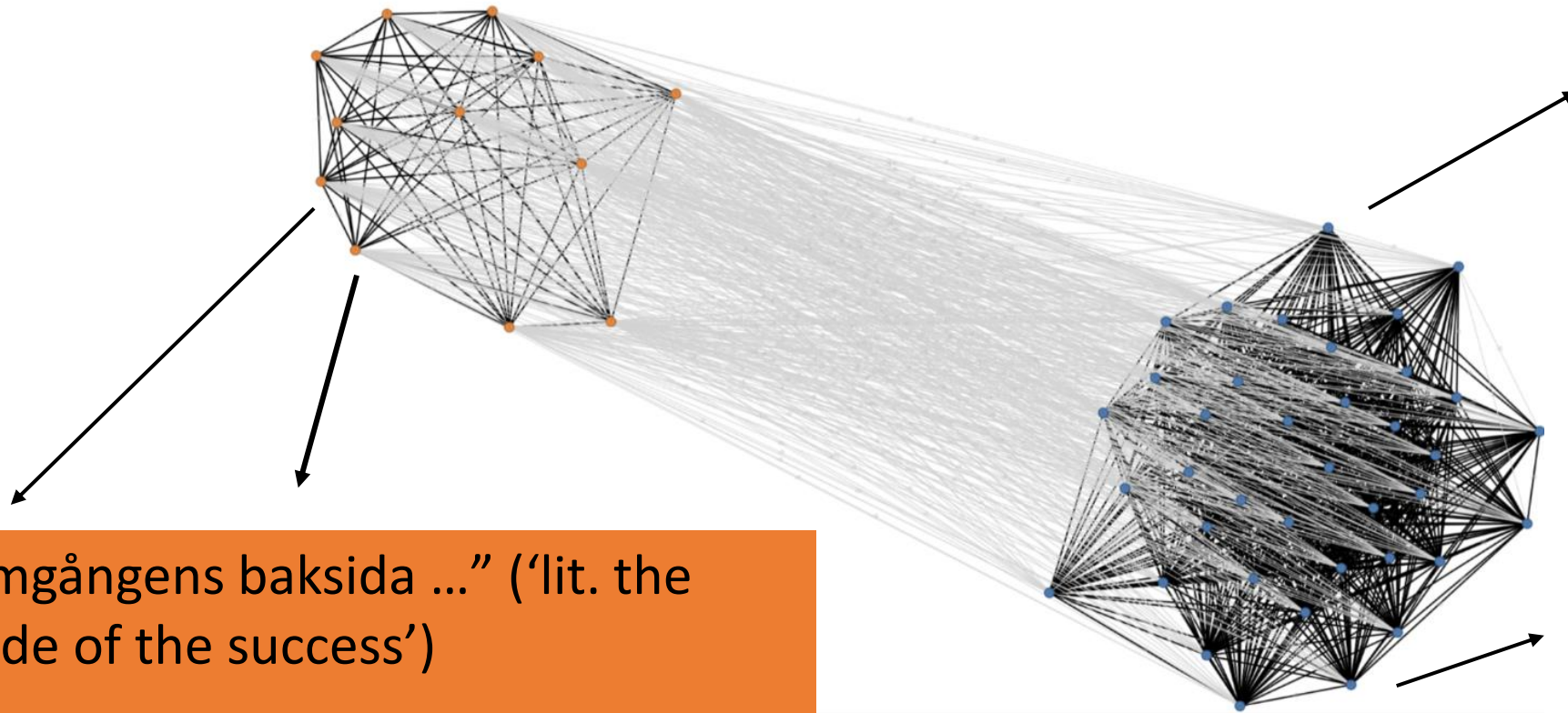
SVT corpus (texts published by the Swedish public service television company, 2004-2021).

About 200 million tokens

Available through Korp, Språkbanken's word research platform



**Example:** the noun *baksida* ('back; downside, disadvantage') in DUrel



“... framgångens baksida ...” (‘lit. the back side of the success’)

“... baksidan med droger ...” (lit. ‘the back side with drugs’) (SVT)

“... polishusets baksida ...” (‘the back of the police house’)

“... baksidan av låret ...” (‘the back of the thigh’) (SVT)

# Experiment 1

We started from SO headwords with more than one sense and evaluated whether DUREl managed to cluster these senses among the uses in a corpus.

18 SO-headwords with one or more subsenses.

**Example:** *enkelspårig* - 'one-track; 'simplistic, superficial, narrow-minded'

50 random usages (sentences) of the headwords were extracted from the SVT corpus.

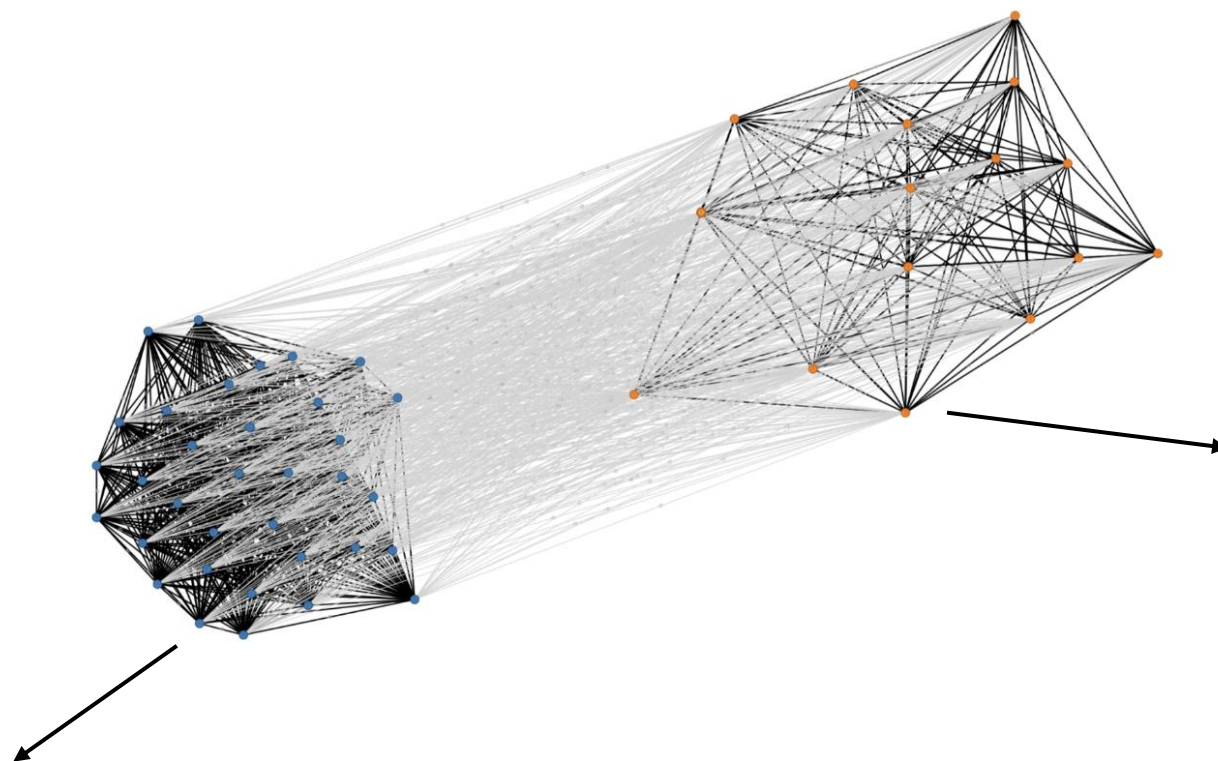
SO lexicographers assessed all the sentences and classified if the senses of every word was covered in the dictionary and, if that was the case, which sense it was (the main sense, a subsense etc.).

Goal: 1) get an indication of the semantic quality of the entries in SO  
2) create a manually curated gold standard.

**Example:** DUREl sense cluster for the headword *enkelspårig* ('one-track; 'simplistic, superficial, narrow-minded')

"Den enkelspåriga järnvägen mellan Motala och Hallsberg är idag en flaskhals ..."

('The one-track railway between Motala and Hallsberg is a bottleneck today ...').



"De tror att vi är enkelspåriga lantisar, de tror att vi är trångsynta, att vi är rasister och homofober."

('They think we're narrow-minded peasants, they think we're bigoted, that we're racists and homophobes.').

# Cluster evaluation based on ARI

Headword	ARI	Headword	ARI
enkelspårig ('one-track, simplistic')	1.0	kriga ('make war')	0.614
fasad ('facade')	1.0	rutten ('rotten')	0.299
ofantlig ('immense')	1.0	ventilera ('ventilate')	0.291
baksida ('back')	0.863	lirka ('tinker, coax')	0.251
fotavtryck ('footprint')	0.84	vansinnig ('insane')	0.237
klimat ('climate')	0.772	hagla ('fall hail')	0.228
bagage ('baggage')	0.758	skör ('fragile')	0.068
vissen ('withered')	0.645	hemmaplan ('setting')	0.0
tvärnita ('jam on the breaks')	0.642	kapitulera ('capitulate')	-0.008
<b>Average</b>		<b>0.528</b>	

## Qualitative analysis of the DUREl clusters

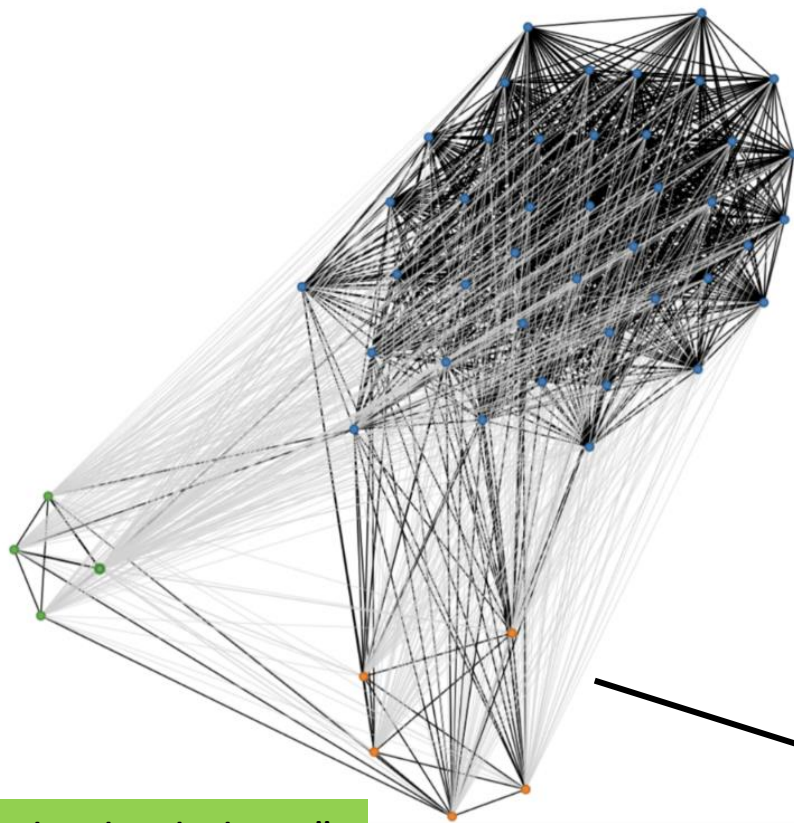
In addition, the DUREl clusters were analyzed qualitatively by the lexicographers, which provided useful insights.

Example: in SO, the headword *hagla* (lit. 'fall hail') has two senses, the main sense 'fall hail' and the figurative subsense 'appear (over someone) in large quantity // mostly about abstract phenomena'.

DUREl predicts *three* main semantic clusters which might be an indication that the word has three senses.



DURel sense cluster for *hagla* ('fall hail; appear (over someone) in large quantity // mostly about abstract phenomena')



"När vi kom ut så haglade det ganska kraftigt, och blåste och spöregnade."

('When we came out, it was hailing quite heavily, and blowing and pouring down') (SVT)

"Smädeorden haglade ..."

('the vituperations rained/came thick and fast' ... ). (SVT)

"De [ungdomarna] står ofta på avstånd och kastar och för polisen blir det då att springa mot dem medan stenar haglar."

('They [the young people] often stand at a distance and throw, and the police then have to run towards them while stones are hailing'.) (SVT)

Lexicographers: based on the editorial principles, the current subsense in the SO entry *hagla* is too wide

## Experiment 2

We focused on the number of senses recorded in SO and compared that number with the number of clusters inferred in the corpus using DUReL.

A discrepancy in the number of SO senses and the recorded clusters can indicate that the entry is outdated. Such SO headwords can be prioritized when revising the dictionary.

About 39,000 nouns, adjectives and verbs have only one sense recorded in SO. We selected (randomly) 281 of them.

All of the 281 headwords have at least 25 occurrences in the corpus. The usages were extracted from the corpus and DUReL clusters were inferred.

## Experiment 2: some results

215 out of 281 words were predicted to have only *one* DUREl cluster (and hence one sense), e.g. *brevinkast* ('letterbox'), *tunisisk* ('Tunisian') and *sukta* ('long (in vain)').

Assessment by the SO lexicographers: these headwords have only one sense in the corpus samples.

→ the semantic description of those headwords in SO is good enough.

## Experiment 2: more results

49 of the 281 headwords with one sense had 2 clusters in DUREl,

9 of them had 3 clusters in DUREl,

6 had 4 clusters in DUREl,

1 had 5 clusters in DUREl,

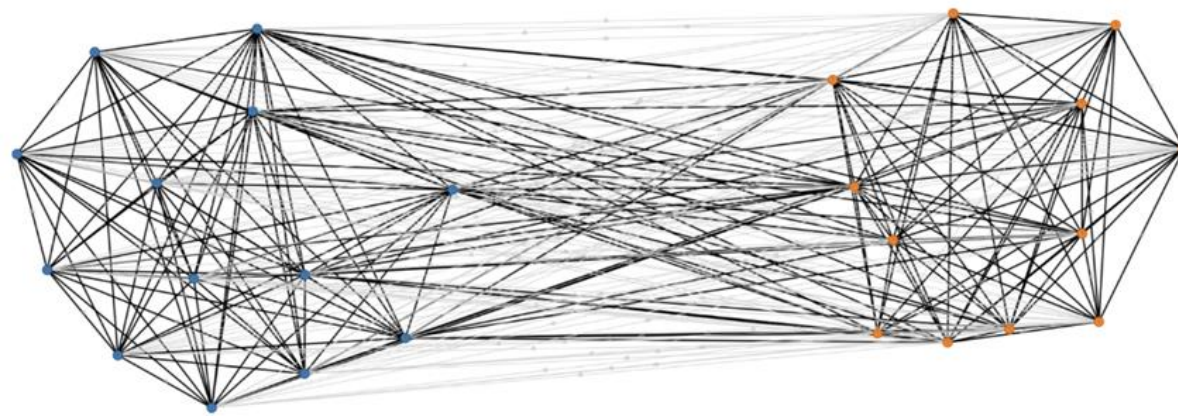
1 had 6 clusters in DUREl

Hence, according to the annotation tool, these headwords were predicted to have between 2 and 6 senses.

# Example: two DUrel clusters for the noun *lydnad* ('obedience')

“tävla i lydnad”  
(‘compete in  
obedience’);

“nordiska  
mästerskap i  
lydnad”  
(‘Nordic  
championships  
in obedience’).  
(SVT)



“Nunnor följer de tre  
klosterlöftena: att  
leva i celibat,  
fattigdom och lydnad.

(‘Nuns follow the  
three monastic vows:  
to live in celibacy,  
poverty and  
obedience.’) (SVT)

The word has only one sense in SO but two senses in the corpora

## More SO headwords that lack a sense

Some more headwords with only *one* sense in SO that have two or more confirmed semantic clusters by DUReI:

- *slutspurt* ('final spurt' but also 'finish'),
- *brotningsmatch* ('wrestling-match'),
- *avstickare* ('detour' but also 'digression')

The new senses are now included in the SO database and will be public in the next edition.

# Summary and discussion

The experiments are still small-scaled, but the results are promising.

The studies has brought the lexicographers' attention to

- missing main senses,
- subsenses (figurative uses, meaning extensions and specializations) that should be added and
- meaning descriptions that are too general and should be split in accordance with the principles for the semantic descriptions in the dictionary.

So far we have mainly found well-established senses that should already have been included in SO, but we have also found completely new senses in Swedish.

## Summary and discussion (2)

Based on the results of the experiment, it is also clear that the DUREl can be improved.

For example, when it comes to nouns like *fransos* ('Frenchman') and *kulstötare* ('shot-putter'), DUREl shows two clusters, but the words have only one sense in SO and in the corpora.



# Thank you for your attention!

The noun  
*slutspurt* ('final  
spurt'; finish')

A figurative  
subsense have  
been added to  
the SO  
database,  
October 2024

**slutspurt** *slutspurten slutspurter*

**ORDKLASS:** substantiv

**UTTAL:** slu`tspurt

- sista del av spurt i hastighetstävling

**SYN**

→ finish *substantiv* → sista, avgörande del av (hastighets)tävling

**EXEMPEL:** *en rafflande slutspurt mot mållinjen*

- äv. bildligt

**EXEMPEL:** *slutspurt på terminen; vad kan vi förvänta oss av slutspurten inför valet?*

**HISTORIK:** belagt sedan 1890

..