CWI

Combining distributional semantics and structured data to study lexical change

Astrid van Aggelen, Laura Hollink, Jacco van Ossenbruggen

VRE4EIC

scores of lexical change derived using distributional NLP

+

PRINCETON UNIVERSITY

WordNet
A lexical database for English

Outline

- WHY this integration?
- WHAT NLP lexical change data do we have?
- WHAT does Wordnet contain?
- HOW did we integrate the two?
- WHAT can this integrated source be used FOR?

[writings, yellow, four, woods, preface, aggression, marching, looking, granting, eligible, electricity, rouse, originality, lord, meadows, sinking, hormone, regional, pierce, appropriation, foul, politician, bringing, disturb, recollections, prize, wooden, persisted, succession, immunities, reliable, charter, specially, nigh, tired, hanging, bacon, pulse, empirical, elegant, second, valiant, sustaining, sailed, errors, relieving, thunder, cooking, contributed, fingers, vassals, fossil, designing, increasing, admiral, hero, avert, reporter, error, atoms, reported, china, burgesses, pancreas, natured, substance, pretensions, climbed, reports, controversy, natures, military, numerical, criticism, golden, divide, classification, owed, explained, replace, brought, remnant, stern, unit, opponents, painters, spoke, occupying, symphony, music, therefore, strike, sermons, females, holy, populations, successful, brings, hereby, hurt, glass, harmless, midst, hold, circumstances, morally, locked, pursue, accomplishment, plunged, temperatures, concepts, revenues, example, misfortunes, triple, unjust, household, artillery, organized, currency, caution, british, want, absolute, provincial, complaining, travel, drying, feature, machine, hot, significance, symposium, preferable, dignified, oceans, beauty, shores, wrong, destined, types, profess, effective, youths, revolt, headquarters, presiding, baggage, keeps, democratic, wing, wind, wine, senators, welcomed, dreamed, concurrence, reforms, vary, quakers, fidelity, wrought, admirably, fit, heretofore, fix, occupations, survivors, distinguishing, fig, nobler, wales, hidden, admirable, easier, glorify, grievous, detachment, effects, schools, township, sixteen, silver, structural, represents, clothed, arrow, addicted, interfering, burial, preceded, financial, telescope, concord, series, displacement, commons, contracting, fortnight, substantially, cathedral, message, whip, borne, toleration, misfortune, excepting, mason, re, encourage, adapt, engineer, foundation, assured, threatened, strata, sensory, assures, faculties, grapes, crowned, estimate, universally, chlorine, enormous, ate, exposing, heading, shipped, musicians, speedy, repealed, appreciable, nouns, channels, wash, instruct, olds, exchequer, service, similarly, engagement, cooling, needed, master, listed, legs, bitter, ranging, listen, danish, rewards, collapse, bounty, wisdom, motionless, sulphur, positively, peril, showed, coward, tree, nations, project, pneumonia, idle, exclaimed, endure, seminary, feeling, acquisition, willingness, spectrum, shrubs, notwithstanding, dozen, affairs, wholesome, person, responsible, eagerly, metallic, recommended, causing, absorbed, amusing, doors, committing, transactions, belligerent, object, diminishing, wells, swiss, affirmation, mouth, letter, conceded, retaining, shalt, singer, episode, grove, professor, camp, fugitives, detriment, nineteenth, incomplete, saying, bomb, insects, meetings, nominated, schism, undue, soluble, gauge, participate, tempted, lessons, touches, busy, liberated, holder, bush, bliss, touched, rich, heartily, rice, plate, remotest, terrors, foremost, pocket, altogether, relish, societies, contributes, patch, release, hasten, respond, blew, disaster, fair, unanimously, expediency, consummation, sensitivity, radius, result, fail, resigned, hammer, best, lots, rings, solicitude, pressures, score, scorn, propagated, occupational, magnesium, preserve, discipline, men, extend, nature, rolled, felony, impetus, extent, defiance, carbon, debt, tyranny, accident, sacrificing, disdain, country, readers, adventures, demanded, estates, planned, logic, argue, adapted, asked, alternate, …]

NLP data of lexical change are often at the level of strings… :-(

scores of lexical change derived using distributional NLP

PRINCETON UNIVERSITY

WordNet

A lexical database for English

scores of lexical change derived using distributional NLP

PRINCETON UNIVERSITY

WordNet

A lexical database for English

Distributional NLP

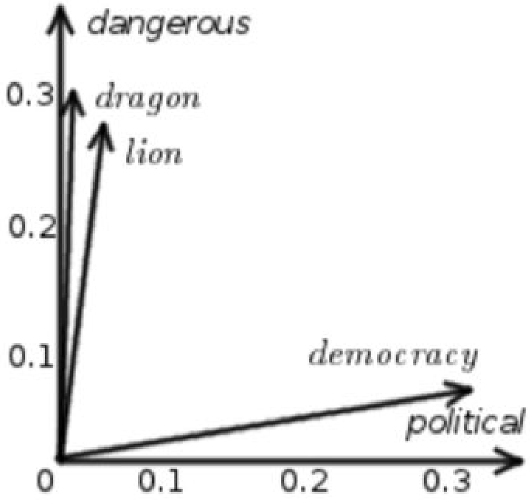
from text corpus to word vector

tanglements begin to have a significant effect on the relaxation times. The undiluted sy
even more doses. Although its effect on the circulation of wild polioviruses ha
their properties would have a beneficial effect on the overall scheme, members heard.
as rabbits or sheep, has a devastating effect on the fine-leaved bouncy turf rich in spe
st, such groups must have had a major effect on the structure of the forest. The v
ish whether artemether has a beneficial effect on the objective and unambiguous prima
ernment and that has inevitably had an effect on the level of the charge. 'This is fi
og-meat and biscuits had had a ruinous effect on the housekeeping. Happily Herbert ha
were talking about had had a very bad effect on the Quigleys. Mrs Quigley was hyper
oleoresins of the dipterocarps have an effect on the bacteria of the fore-stomach of col
n but progressive and compensatory in effect. On the circumference of that circle are n
bility of charging for more services. The effect on the demography of the inner cities co
ce in April 1988 have had a devastating effect on young people. At the stroke of a pen t
ur to her to worry about the devastating effect Paula was having on Edward. Behir
and for public health activities. Thus in effect reference centres are indistinguishable f
matrix between 'knowledge of a cause/effect relationship between participation progra
nds, detecting a marked distance decay effect. Research p
rease in blood volume in the lungs I an effect shown by transthoracic impedance techn
ime. It is this delay between cause and effect that is fundamental to the observed visc
so great variety") give an overall effect that the conclusion is a promotional, or u
e per se , there is some authority to the effect that trespass to goods requires proof of
Tc interval confirming a largely additive effect; the dose response curves for salbutam
lal solution are further examples of this effect. The fundamentals of light scatteri
v up together than the cross-cousins. In effect, the parallel cousins are as familiar as s
hat if a placebo is to have a therapeutic effect, the patient must believe that it will. Nev

	I	*like*	*enjoy*	*deep*	*learning*	*NLP*	*flying*	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

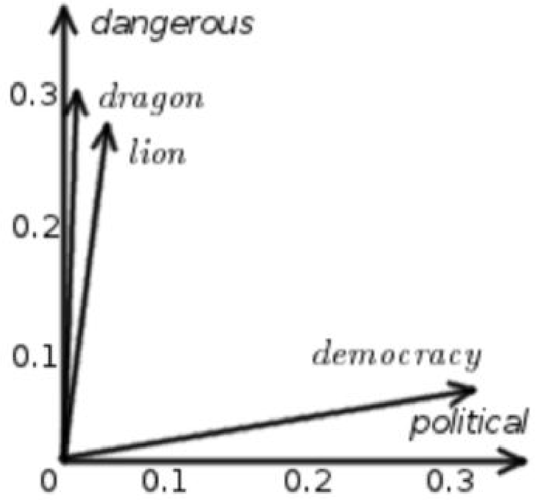
Distributional NLP

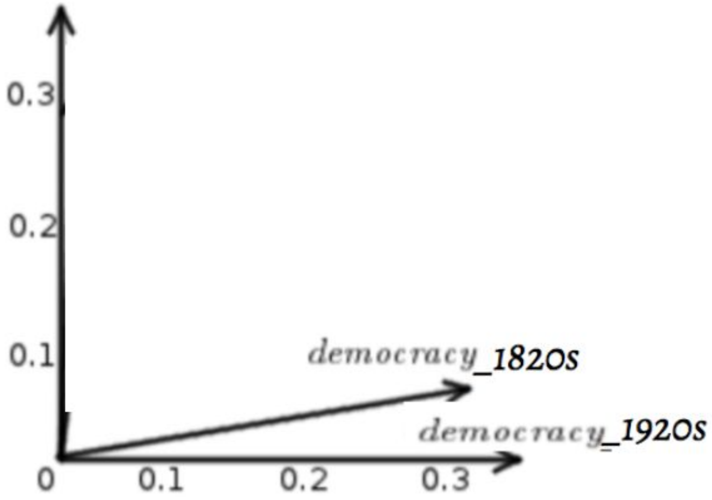
from word vector to similarities



Distributional NLP

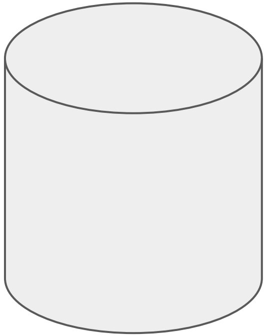
from word vector to similarities over time





HistWords

The NLP data we use



10k English words (w)

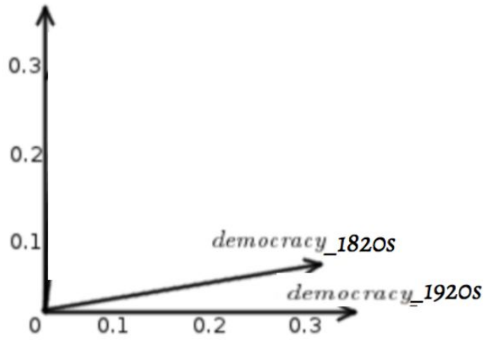
x

37 cross-decade

cosine sim's:

cos-sim(w_t, w_{t+1}) 1810s-1820s, …, 1990s-2000s

cos-sim (w_t, w_{1990s}) 1810s-1990s, …, 1980s-1990s

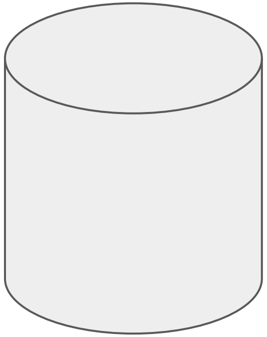


HistWords:
Word Embeddings for Historical Text

William L. Hamilton, Jure Leskovec, Dan Jurafsky

HistWords

The NLP data we use



10k English words (w)

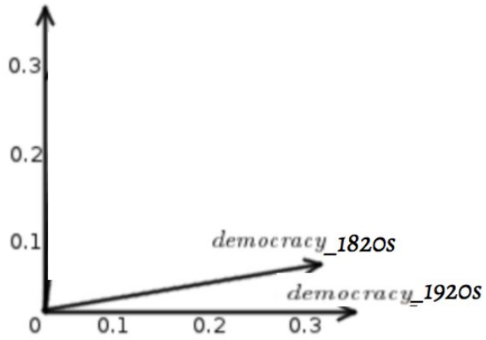
x **not POS-tagged!**

37 cross-decade

cosine sim's:

cos-sim(w_t, w_{t+1}) 1810s-1820s, …, 1990s-2000s

cos-sim (w_t, w_{1990s}) 1810s-1990s, …, 1980s-1990s



HistWords:
Word Embeddings for Historical Text

William L. Hamilton, Jure Leskovec, Dan Jurafsky

scores of lexical change derived using distributional NLP

PRINCETON UNIVERSITY

WordNet

A lexical database for English

PRINCETON UNIVERSITY

WordNet

A lexical database for English

WordNet Search - 3.1

- [WordNet home page] - [Glossary] - [Help]

Word to search for: web | 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

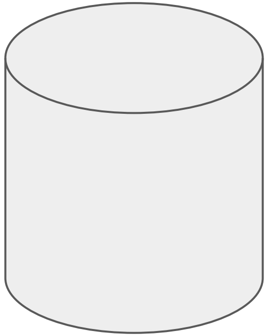
- S: (n) **web** (an intricate network suggesting something that was formed by weaving or interweaving) *"the trees cast a delicate web of shadows over the lawn"*
- S: (n) **web**, entanglement (an intricate trap that entangles or ensnares its victim)
- S: (n) vane, **web** (the flattened weblike part of a feather consisting of a series of barbs on either side of the shaft)
- S: (n) network, **web** (an interconnected system of things or people) *"he owned a network of shops"; "retirement meant dropping out of a whole network of people who had been part of my life"; "tangled in a web of cloth"*
- S: (n) World Wide Web, WWW, **web** (computer network consisting of a collection of internet sites that offer text and graphics and sound and animation resources through the hypertext transfer protocol)
- S: (n) **web** (a fabric (especially a fabric in the process of being woven))
- S: (n) **web** (membrane connecting the toes of some aquatic birds and mammals)

Verb

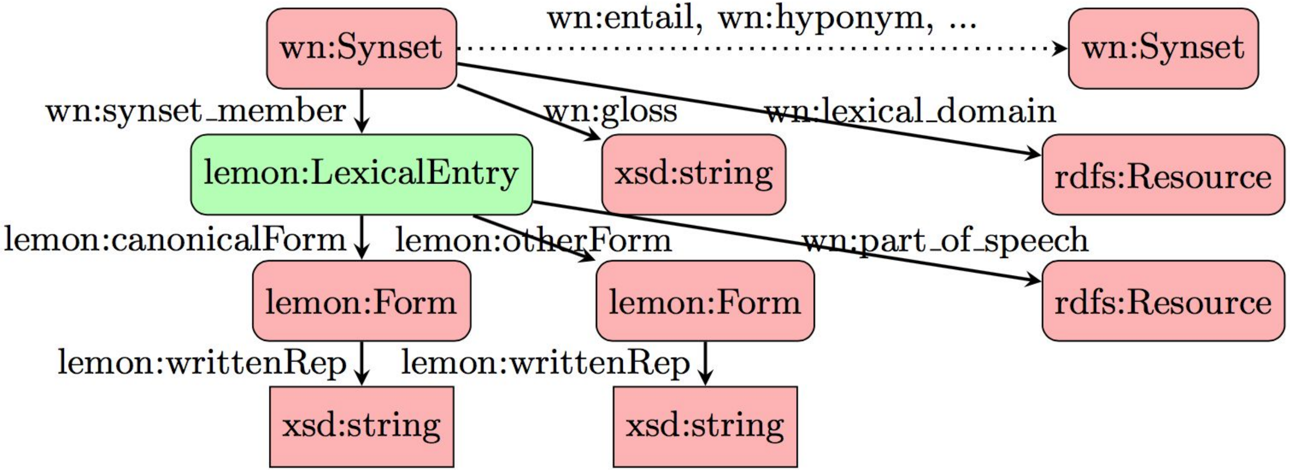
- S: (v) **web**, net (construct or form a web, as if by weaving)

Wordnet 3.1 RDF

RDF-WN containing +/- 150k English lexical entries



PRINCETON UNIVERSITY
WordNet
A lexical database for English



scores of lexical change derived using distributional NLP

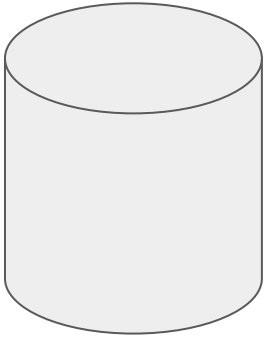
+

PRINCETON UNIVERSITY

WordNet
A lexical database for English

Similarities to distances

The NLP data we use



10k English words (w)

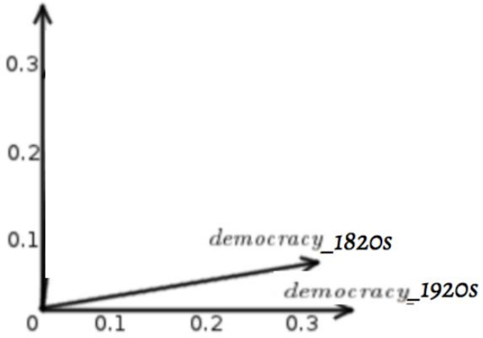
x

37 cross-decade

cosine dist's:

cos-dist(w_t, w_{t+1}) 1810s-1820s, …, 1990s-2000s

cos-dist(w_t, w_{1990s}) 1810s-1990s, …, 1980s-1990s

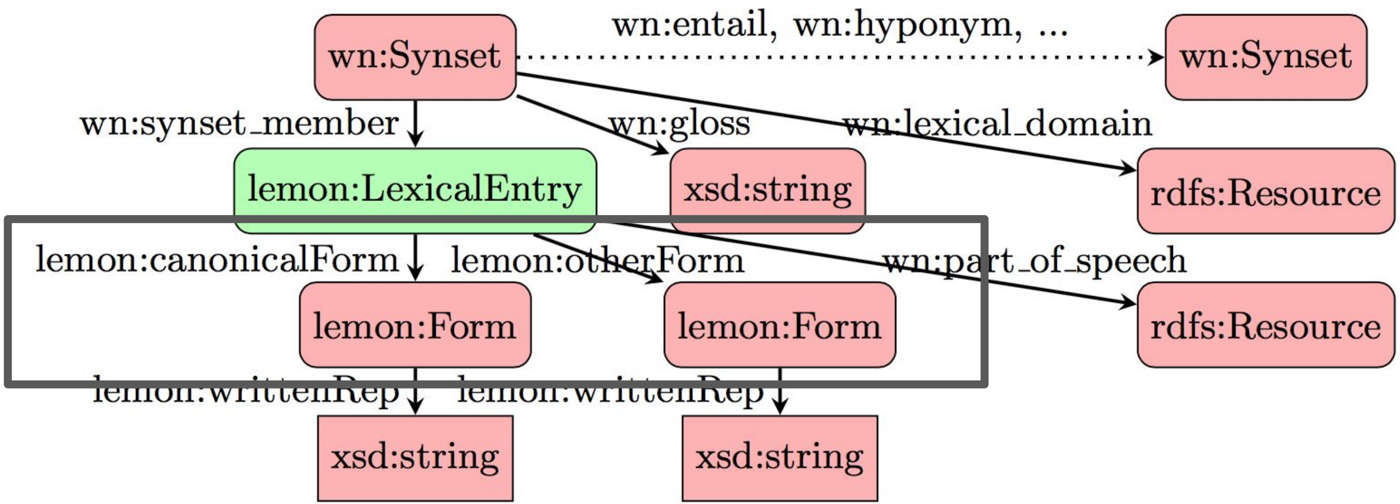


HistWords:
Word Embeddings for Historical Text

William L. Hamilton, Jure Leskovec, Dan Jurafsky

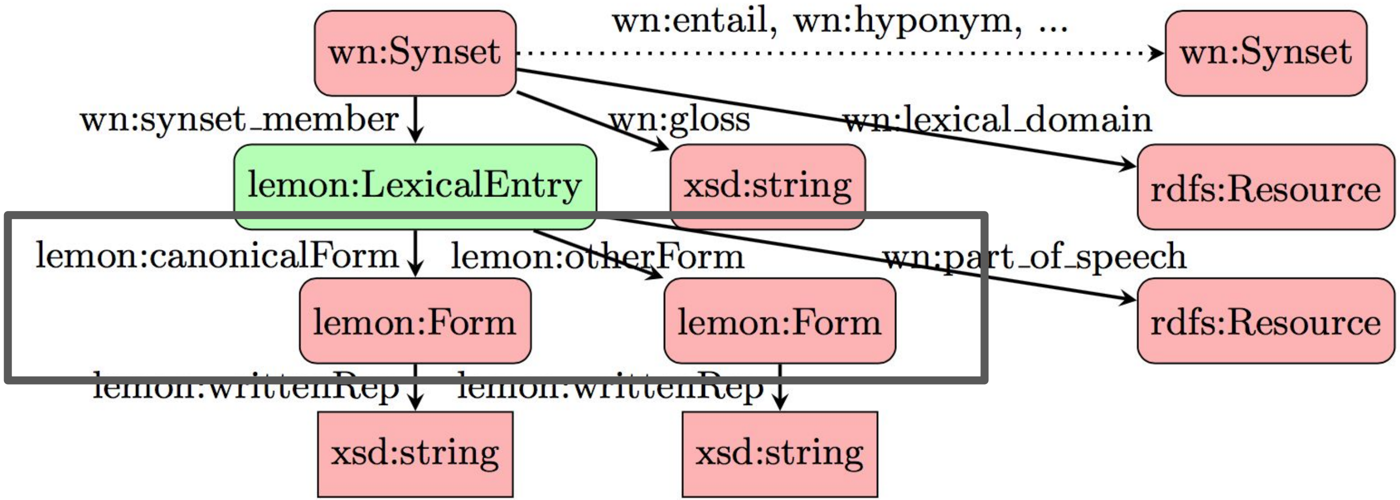
Linking HistWords to Wordnet

- What WN instance level to annotate with change scores?



Linking HistWords to Wordnet

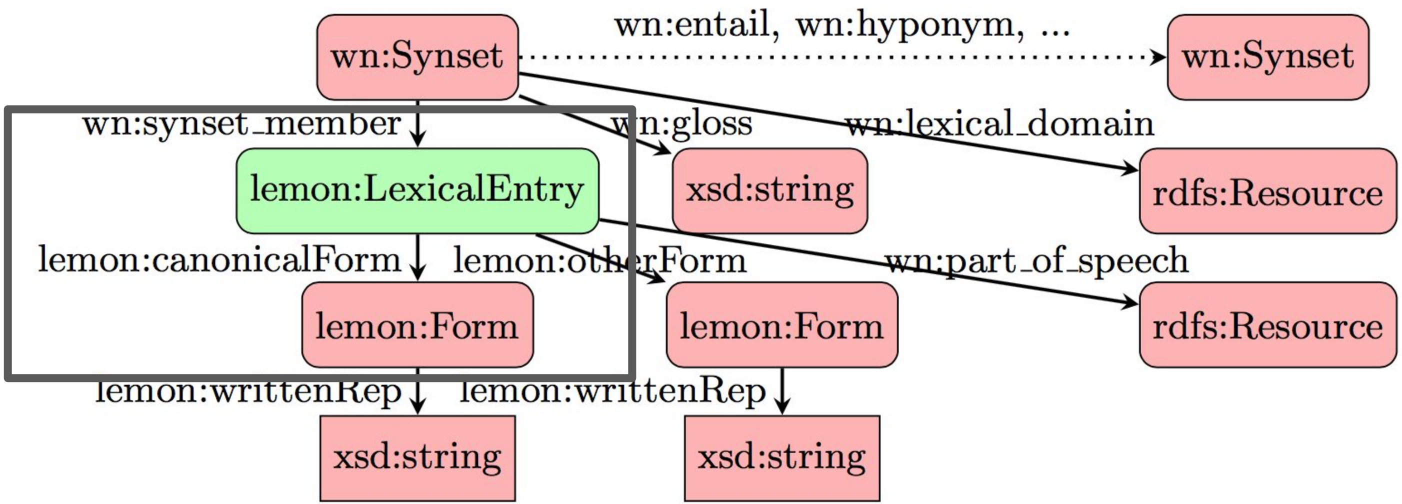
- What WN instance level to annotate with change scores?



Problem:
queries relating change scores and lexical entries need a complicated UNION operation

Linking HistWords to Wordnet

- What WN instance level to annotate with change scores?



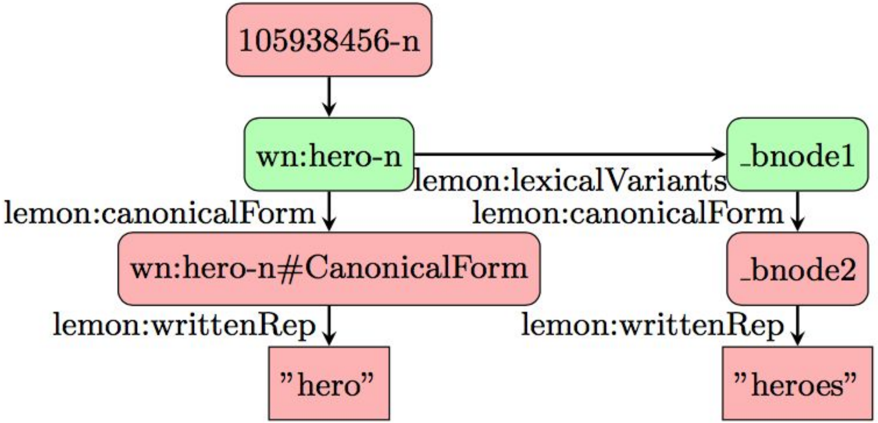
Pragmatic solution: use just the canonical forms of LEs, making the relation between LE and label one-to-one. Now the change can be attached to LE.

Linking HistWords and Wordnet entries

1. **Match** HistWords words on canonical form of lexical entries
   => 7.365 matches (out of 10.000)
2. **Stem** HistWords words **and match** on canonical forms
   => 8.878 matches (out of 10.000)

Linking HistWords and Wordnet entries

1. **Match** HistWords words on canonical form of lexical entries
   => 7.365 matches (out of 10.000)
2. **Stem** HistWords words **and match** on canonical forms
   => 8.878 matches (out of 10.000)



Linking HistWords and Wordnet entries

1. **Match** HistWords on canonical form
   => 7.365 matches (out of 10.000)
2. **Stem** HistWords words **and match** on canonical forms
   => 8.878 matches (out of 10.000)

Important: one word in HistWords can have match on multiple lexical entries with the same canonical form but with different parts of speech!

E.g. “web” matches on WN lexical entries web-V and web-N

Linking HistWords and Wordnet entries

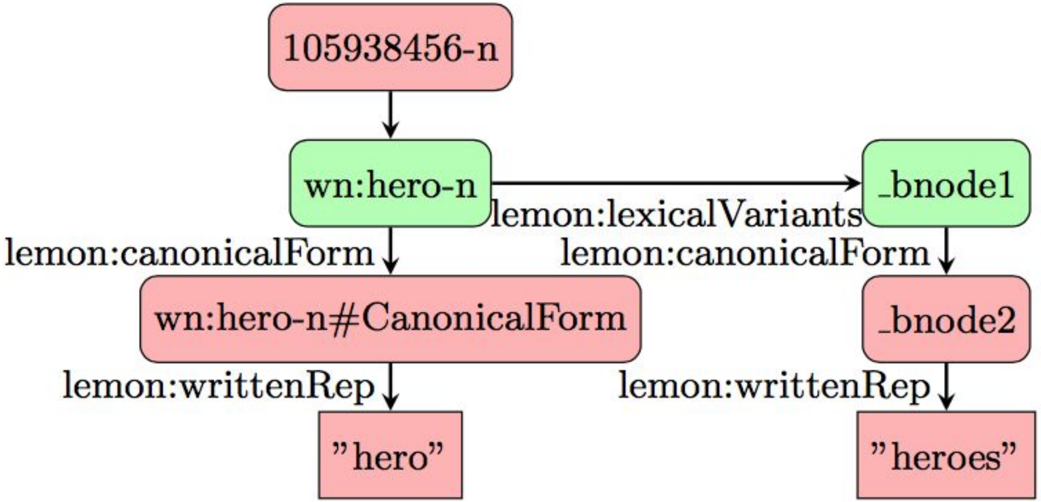
1. **Match** HistWords on canonical form
   => 7.365 matches (out of 10.000)
2. **Stem** HistWords words **and match** on canonical forms
   => 8.878 matches (out of 10.000)
   mapped on 12.469 lexical entries

Important: one word in HistWords can have match on multiple lexical entries with the same canonical form but with different parts of speech!

E.g. “web” matches on WN lexical entries web-v and web-n

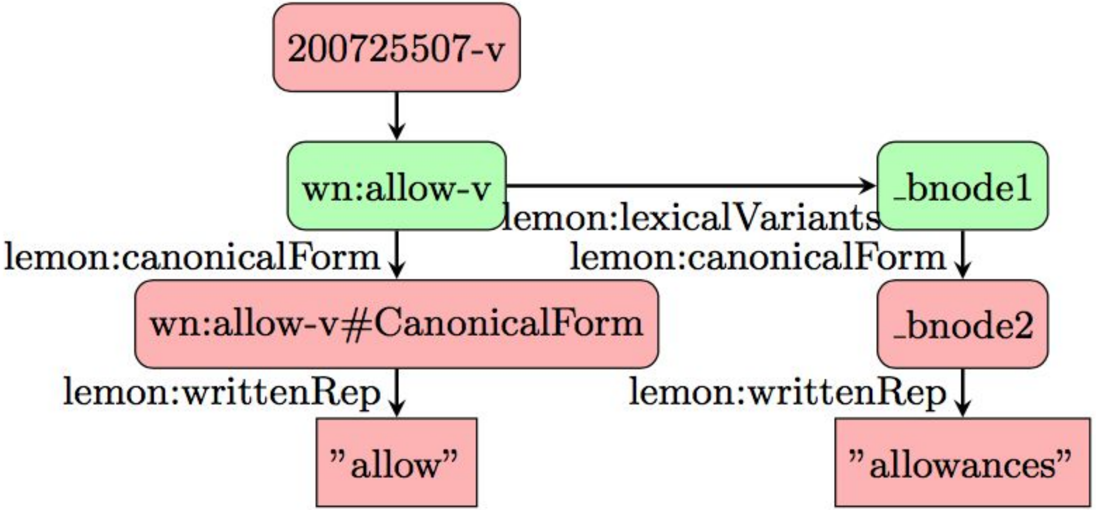
Data model

How we represented matches by stem-and-match:



Data model

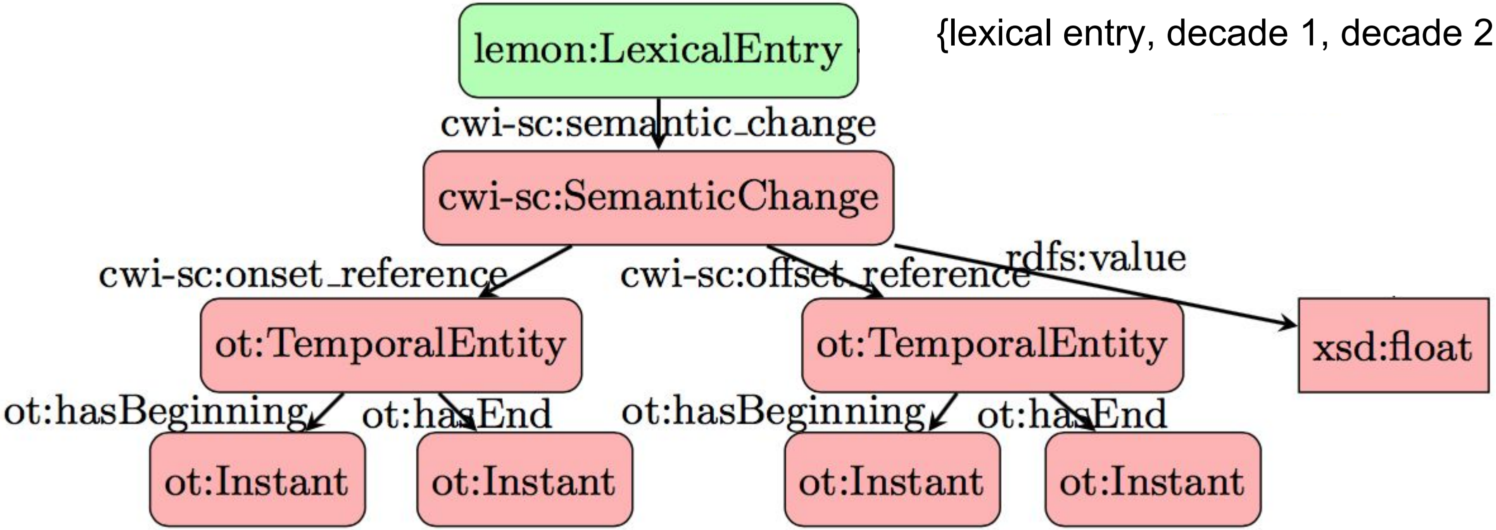
How we represented matches by stem-and-match:



Side note:
another reason for adding the change scores to LEs and not forms is conservativeness: otherwise we would have declared "allowances" to be a verb and to have the same synset!

Data model

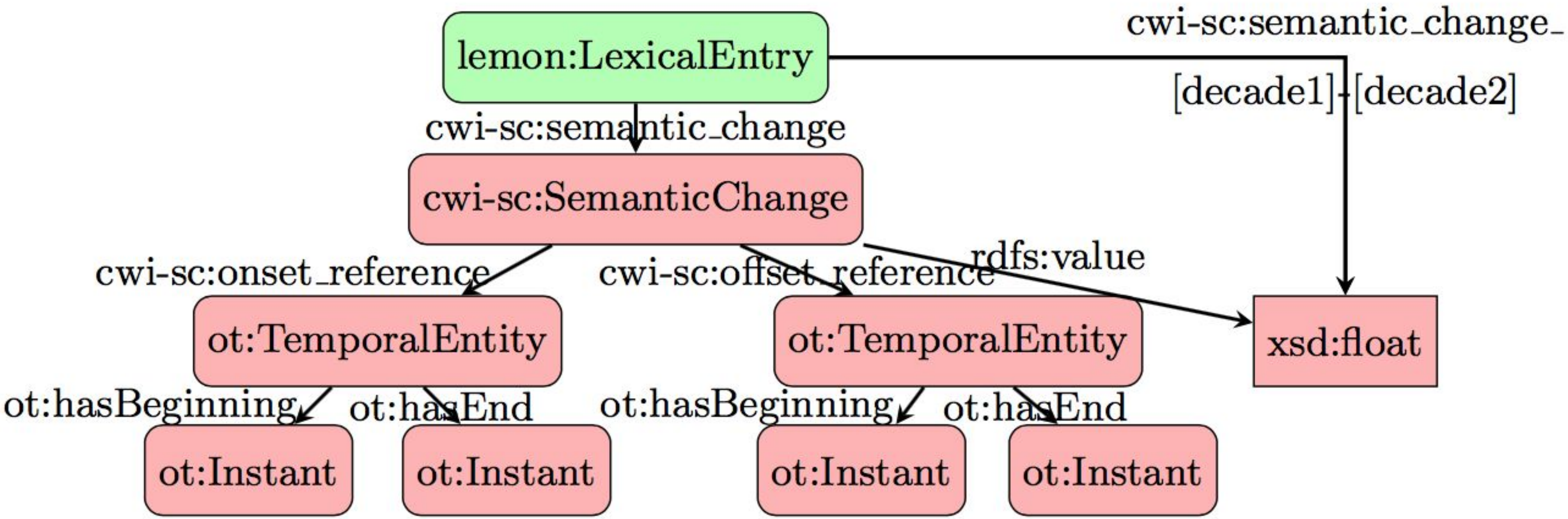
How we connected the change scores to the lexical entries:



{lexical entry, decade 1, decade 2, change score}

Data model

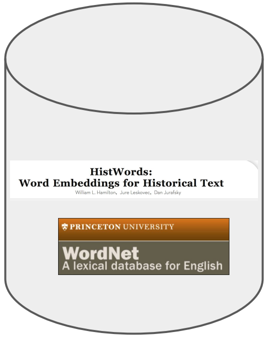
How we connected the change scores to the lexical entries:



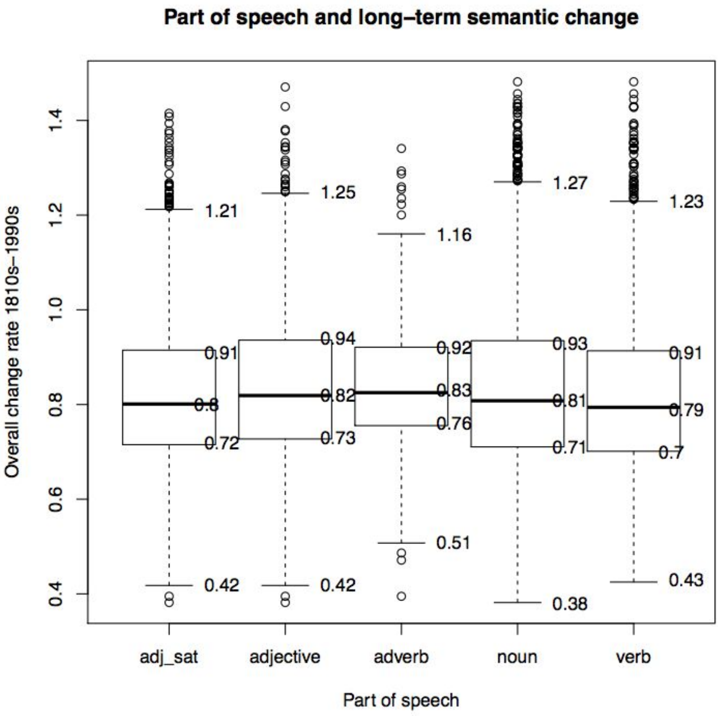
Resulting dataset

- Downloadable (.ttl) from http://github.com/aan680/SemanticChange + WN-RDF from http://wordnet-rdf.princeton.edu
- Queryable using SPARQL

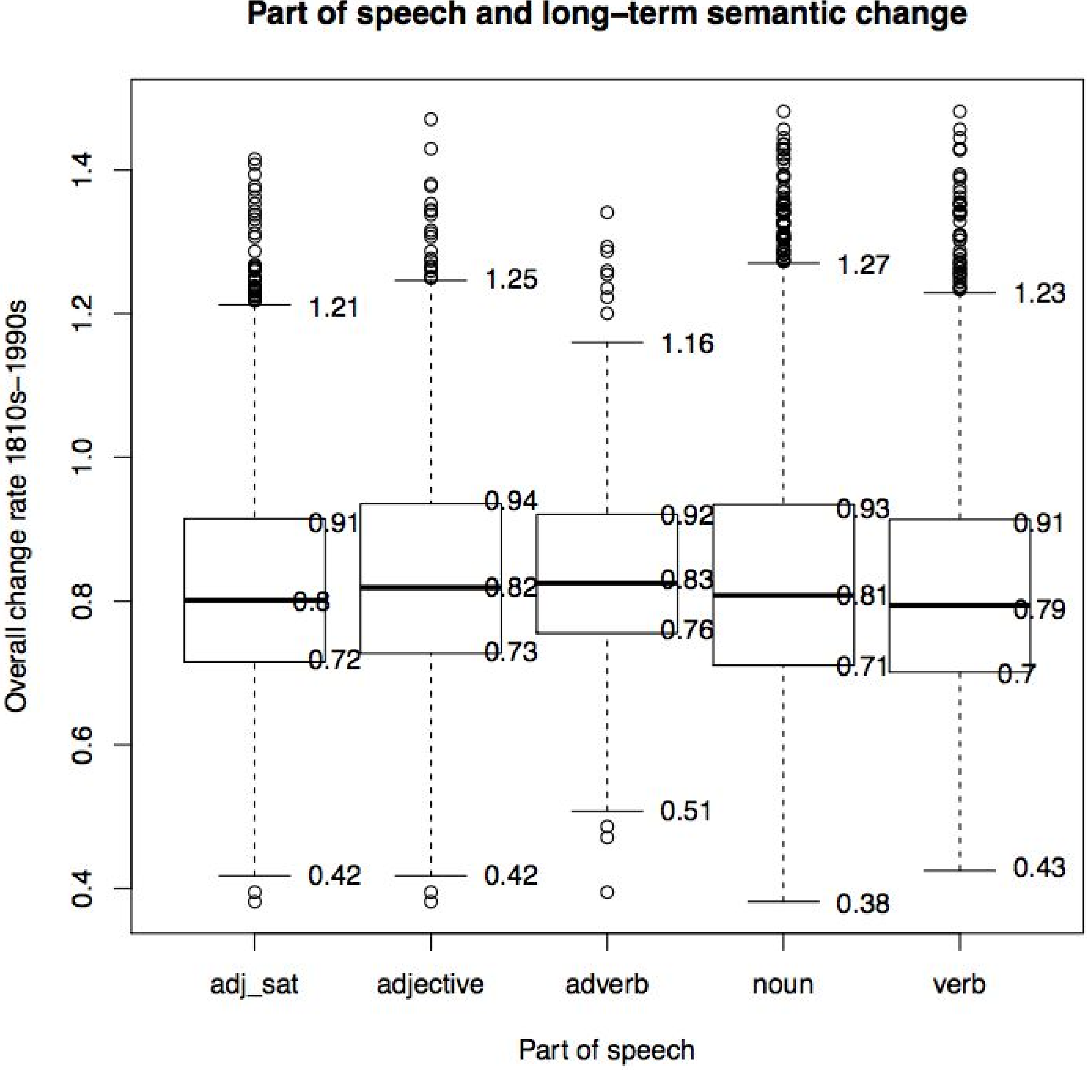
```
PREFIX cwi: <http://project.ia.cwi.nl/semanticChange/>
SELECT * WHERE {
?le cwi:semantic_change_1980s-1990s ?value.
} ORDER BY DESC(?value) LIMIT 5
```



Example applications



Do words of different linguistic categories show different degrees of change?



Part of speech and long–term semantic change
Overall change rate 1810s–1990s
1.4
1.2
1.0
0.8
0.6
0.4
1.21
0.91
0.8
0.72
0.42
1.25
0.94
0.82
0.73
0.42
1.16
0.92
0.83
0.76
0.51
1.27
0.93
0.81
0.71
0.38
1.23
0.91
0.79
0.7
0.43
adj_sat
adjective
adverb
noun
verb
Part of speech

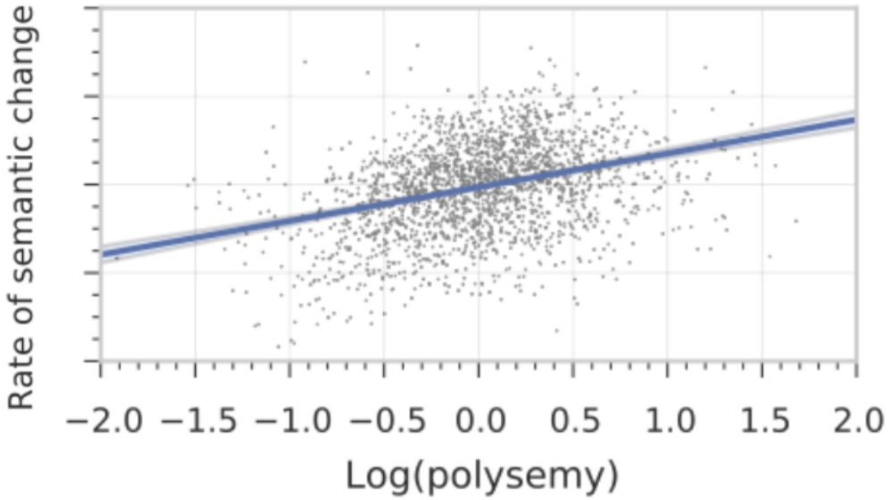
Example applications

Are words of some semantic categories more prone to change than others?

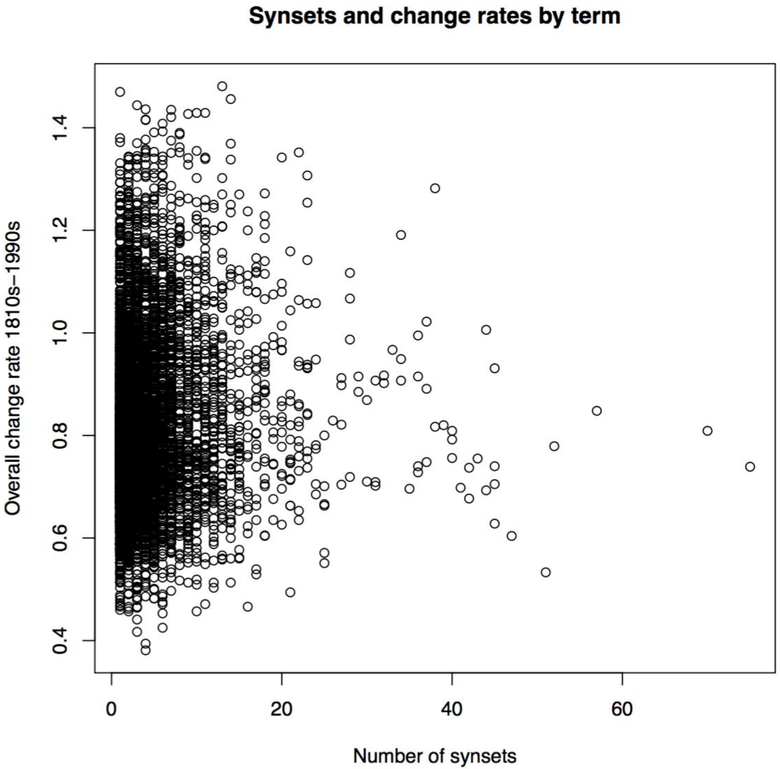
Mean change score	Domain	Mean change score	Domain
0.909	noun.process	0.814	verb.body
0.872	noun.phenomenon	0.791	noun.animal
0.869	noun.event	0.784	noun.food
0.867	noun.act	0.778	noun.feeling
0.86	noun.possession	0.737	verb.weather

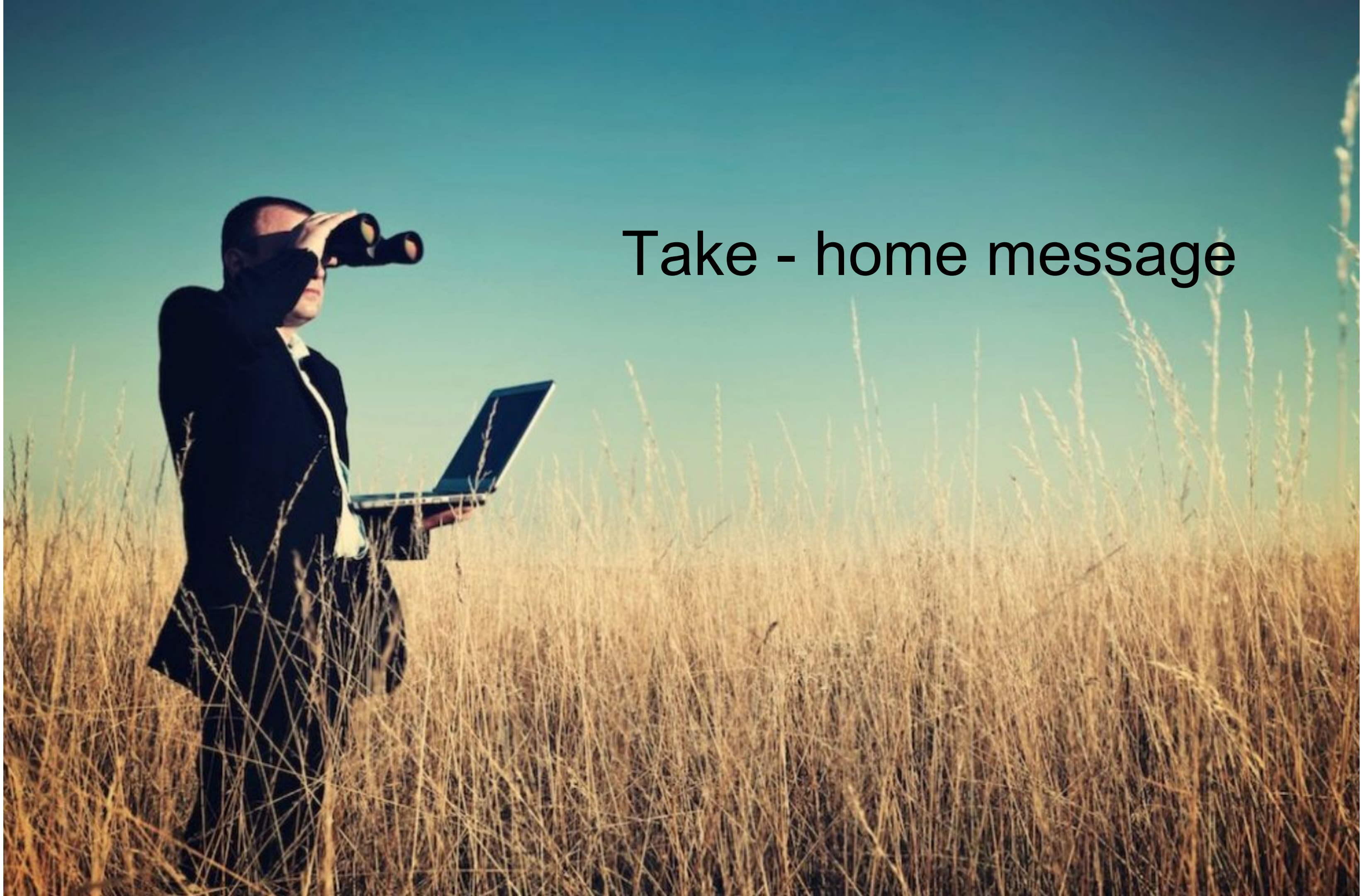
Example applications

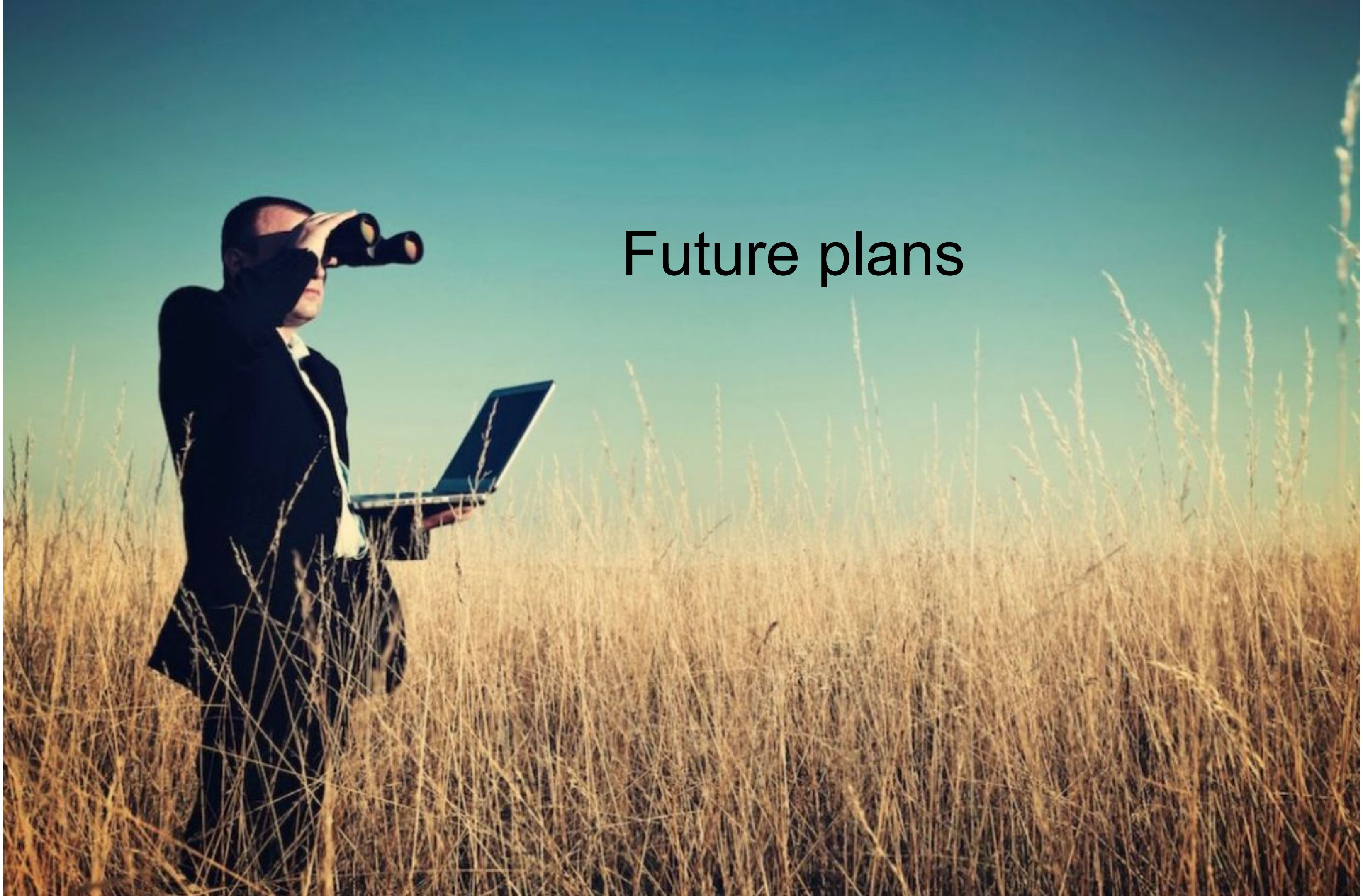
Do more polysemous words and less polysemous words change at a different rate?



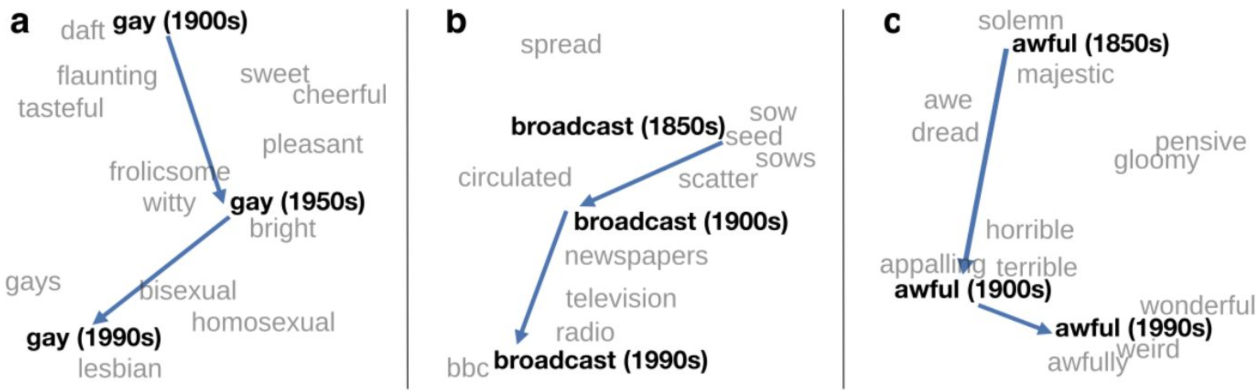
Source: Hamilton et al. 2016

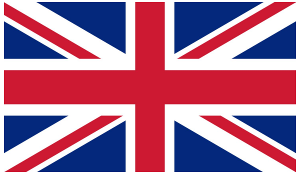



Take - home message

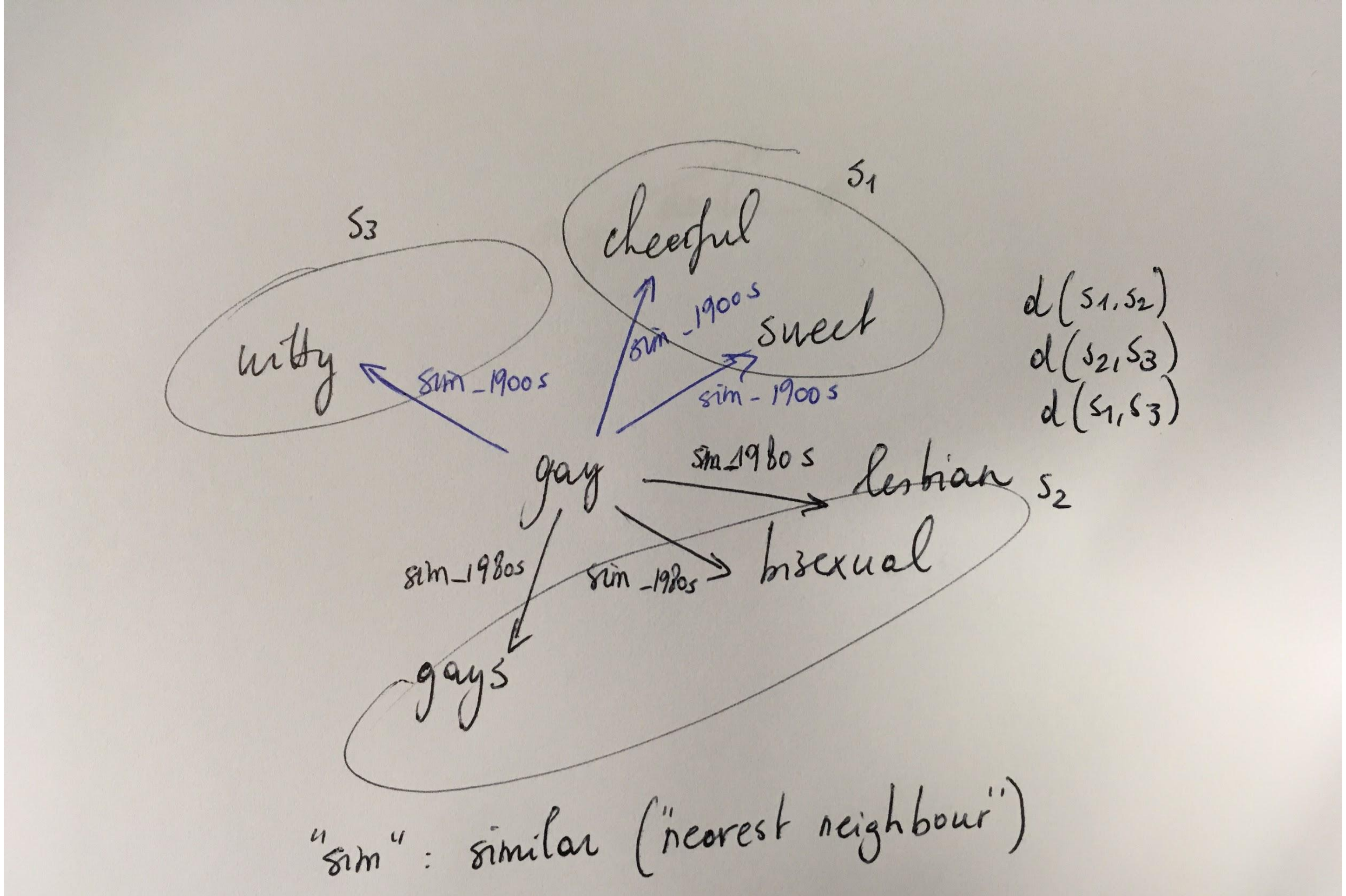


Future plans





Compare lexical change across languages, aiming to distinguish between lexical and conceptual change



Induce the dominant sense of each word per decade, using nearest neighbours and grouping their synsets

Question time!!!

Acknowledgments:

The VRE4EIC project has received funding
from the European Union's Horizon 2020 research and innovation
programme under grant agreement No 676247