Wiktionary and NLP: Improving synonymy networks

ACL-IJCNLP

Singapore, 7 Aug 2009

Emmanuel Navarro	IRIT, CNRS & Univ. of Toulouse
Franck Sajous	**CLLE-ERSS, CNRS & Univ. of Toulouse**
Bruno Gaume	CLLE-ERSS & IRIT, CNRS & Univ. of Toulouse
Laurent Prévot	LPL, CNRS & Univ. of Provence
ShuKai Hsieh	English Department, NTNU, Taiwan
Tzu-Yi Kuo	Graduate Institute of Linguistics, NTU, Taiwan
Pierre Magistry	TIGP, CLCLP, Academia Sinica, GIL, NTU, Taiwan
Chu-Ren Huang	Dept. of Chinese and Bilingual Studies, Hong Kong Poly U., Hong Kong.

Goals

- giving a method for improving synonymy networks;
- applying it to Wiktionary;
- in the meanwhile, investigate the possibilities of:
  - using Wiktionary as a resource for NLP;
  - using NLP for improving Wiktionary.

Summary

Wiktionary as a lexical resource

Lexical resources

- NLP requires lexical resources
- English: Princeton WordNet
- Some other languages (eg. French): non-satisfaying and/or non-free
- Some others: purely under-resourced

Wiktionary as a lexical resource

Lexical resources

- NLP requires lexical resources
- English: Princeton WordNet
- Some other languages (eg. French): non-satisfaying and/or non-free
- Some others: purely under-resourced

Wiktionary

- multilingual
- freely available
- $\longrightarrow$ a perfect candidate?

Wiktionary: (very) short description

Collaborative editing

Non experts-led

Wiktionary: (very) short description

Collaborative editing

Non experts-led

OK, we know but it's worth taking a chance, remaining aware of it

Wiktionary: (very) short description

Collaborative editing

Non experts-led
OK, we know but it's worth taking a chance, remaining aware of it

Articles content

- etymology
- parts of speech
- definitions, examples
- translations
- synonyms/antonyms
- hypernyms/hyponyms

Wiktionary: (very) short description

Collaborative editing

Non experts-led

OK, we know but it's worth taking a chance, remaining aware of it

Articles content

- etymology
- parts of speech
- definitions, examples
- translations
- synonyms/antonyms
- hypernyms/hyponyms

boot

Etymology 1

Middle English, from Old French *bote*

Noun

1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms

* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations

shoe

* French: botte
* Spanish : bota

kick - see kick

Verb

1. To kick

   *I **booted** the ball*

2. To disconnect

   *I got **booted** from the chatroom*

Synonyms

* (*kick*): kick
* (disconnect): kick

Translations [...]

Etymology 2

Akin to Old Norse *bót*

Wiktionary: (very) short description

Collaborative editing

Non experts-led

OK, we know but it's worth taking a chance, remaining aware of it

Articles content

- etymology
- parts of speech
- definitions, examples
- translations
- synonyms/antonyms
- hypernyms/hyponyms

boot

Etymology 1 *etymology*

Middle English, from Old French *bote*

Noun

1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms

* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations

shoe

* French: botte
* Spanish : bota

kick - see kick

Verb

1. To kick

   *I **booted** the ball*

2. To disconnect

   *I got **booted** from the chatroom*

Synonyms

* (*kick*): kick
* (disconnect): kick

Translations [...]

Etymology 2

Akin to Old Norse *bót*

Wiktionary: (very) short description

Collaborative editing

Non experts-led

OK, we know but it's worth taking a chance, remaining aware of it

Articles content

- etymology
- parts of speech
- definitions, examples
- translations
- synonyms/antonyms
- hypernyms/hyponyms

boot

Etymology 1 *etymology*

Middle English, from Old French *bote*

Noun *part of speech (noun)*

1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms

* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations

shoe

* French: botte
* Spanish : bota

kick - see kick

Verb *part of speech (verb)*

1. To kick

   *I **booted** the ball*

2. To disconnect

   *I got **booted** from the chatroom*

Synonyms

* (*kick*): kick
* (disconnect): kick

Translations [...]

Etymology 2

Akin to Old Norse *bót*

Wiktionary: (very) short description

Collaborative editing

Non experts-led

OK, we know but it's worth taking a chance, remaining aware of it

Articles content

- etymology
- parts of speech
- definitions, examples
- translations
- synonyms/antonyms
- hypernyms/hyponyms

boot

Etymology 1 *etymology*
Middle English, from Old French *bote*

Noun *part of speech (noun)*
1. A heavy shoe *wordsenses*
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations
shoe
* French: botte
* Spanish : bota

kick - see kick

Verb *part of speech (verb)*
1. To kick *wordsense #1*
*I **booted** the ball*
2. To disconnect *wordsense #2*
*I got **booted** from the chatroom*

Synonyms
* (*kick*): kick
* (disconnect): kick

Translations [...]

Etymology 2
Akin to Old Norse *bót*

Wiktionary: (very) short description

Collaborative editing

Non experts-led

OK, we know but it's worth taking a chance, remaining aware of it

Articles content

- etymology
- parts of speech
- definitions, examples
- translations
- synonyms/antonyms
- hypernyms/hyponyms

boot

Etymology 1 *etymology*
Middle English, from Old French *bote*

Noun *part of speech (noun)*
1. A heavy shoe *wordsenses*
2. A blow with the foot; a kick.

Synonyms *synonyms*
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations
shoe
* French: botte
* Spanish : bota

kick - see kick

Verb *part of speech (verb)*
1. To kick *wordsense #1*
*I **booted** the ball*
2. To disconnect *wordsense #2*
*I got **booted** from the chatroom*

Synonyms
* (*kick*): kick
* (disconnect): kick

Translations [...]

Etymology 2
Akin to Old Norse *bót*

Wiktionary: (very) short description

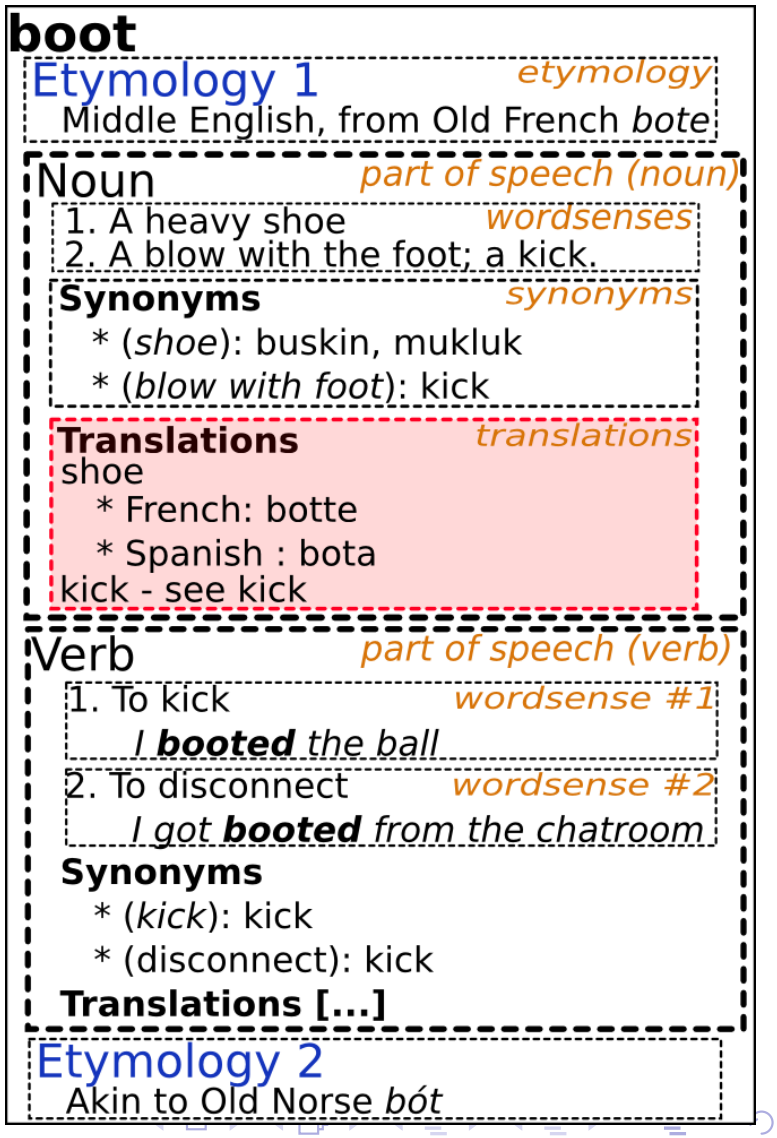
Collaborative editing

Non experts-led

OK, we know but it's worth taking a chance, remaining aware of it

Articles content

- etymology
- parts of speech
- definitions, examples
- translations
- synonyms/antonyms
- hypernyms/hyponyms



Wiktionary: (very) short description

Collaborative editing

Non experts-led

OK, we know but it's worth taking a chance, remaining aware of it

Articles content

- etymology
- parts of speech
- definitions, examples
- translations
- synonyms/antonyms
- hypernyms/hyponyms

The 'regular' case, but...

content&layout heterogeneous over languages and even within a given language

boot

Etymology 1

Middle English, from Old French *bote*

Noun

1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms

* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations

shoe

* French: botte
* Spanish : bota

kick - see kick

Verb

1. To kick

   *I **booted** the ball*

2. To disconnect

   *I got **booted** from the chatroom*

Synonyms

* (*kick*): kick
* (disconnect): kick

Translations [...]

Etymology 2

Akin to Old Norse *bót*

1 Wiktionary

2 Synonymy networks
- Wiktionary graph
- Gold standards
- Comparison

3 Improving Wiktionary's network
- Exploiting its Small World structure
- Using translation links

Extracting Wiktionary's graph of synonymy

Modeling

Extracting Wiktionary's graph of synonymy

Modeling

- one English entry's POS
  $\longrightarrow$ one vertex

Extracting Wiktionary's graph of synonymy

Modeling

- one English entry's POS
  $\longrightarrow$ one vertex
- synonymy links among same POS

Extracting Wiktionary's graph of synonymy

Modeling

- one English entry's POS
  $\longrightarrow$ one vertex
- synonymy links among same POS
- wordsenses flattened

Extracting Wiktionary's graph of synonymy

Modeling

- one English entry's POS
  $\longrightarrow$ one vertex
- synonymy links among same POS
- wordsenses flattened
- edges symmetrized

Extracting Wiktionary's graph of synonymy

Modeling

- one English entry's POS
  $\longrightarrow$ one vertex
- synonymy links among same POS
- wordsenses flattened
- edges symmetrized

boot

English

Noun

1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms

* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Verb

1. To kick
2. To disconnect

Synonyms

* (*kick*): kick
* (disconnect): kick

Extracting Wiktionary's graph of synonymy

Modeling

- one English entry's POS
  $\longrightarrow$ one vertex
- synonymy links among same POS
- wordsenses flattened
- edges symmetrized

boot

English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Verb
1. To kick
2. To disconnect

Synonyms
* (*kick*): kick
* (disconnect): kick

en.N.boot

en.V.boot

Extracting Wiktionary's graph of synonymy

Modeling

- one English entry's POS
  $\longrightarrow$ one vertex
- synonymy links among same POS
- wordsenses flattened
- edges symmetrized

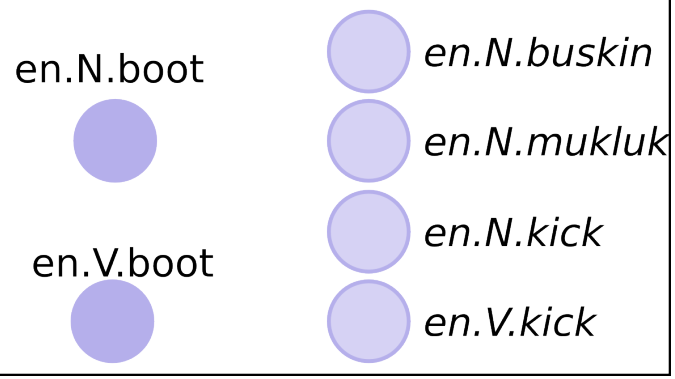
boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Verb
1. To kick
2. To disconnect

Synonyms
* (*kick*): kick
* (disconnect): kick



Extracting Wiktionary's graph of synonymy

Modeling

- one English entry's POS
  $\longrightarrow$ one vertex
- synonymy links among same POS
- wordsenses flattened
- edges symmetrized

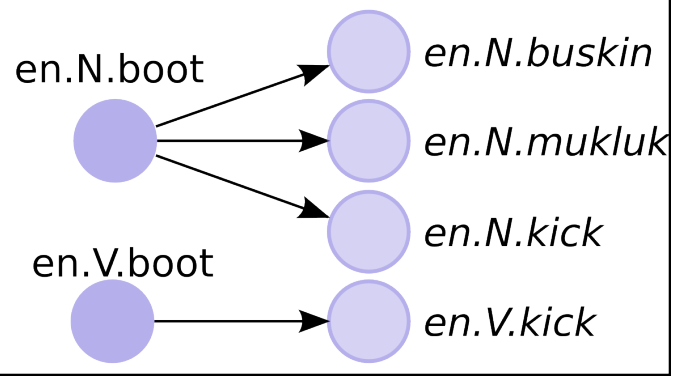
boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Verb
1. To kick
2. To disconnect

Synonyms
* (*kick*): kick
* (disconnect): kick



Extracting Wiktionary's graph of synonymy

Modeling

- one English entry's POS
  $\longrightarrow$ one vertex
- synonymy links among same POS
- wordsenses flattened
- edges symmetrized

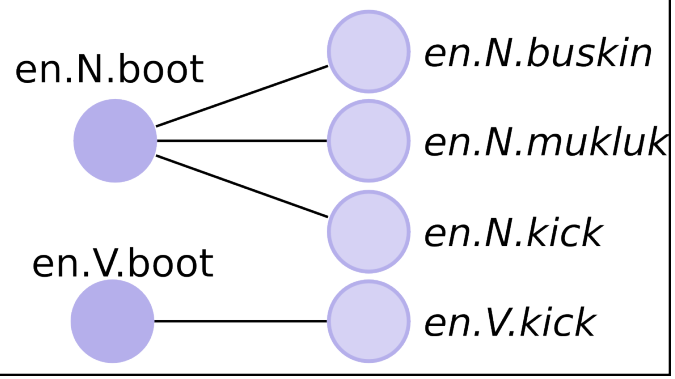
boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Verb
1. To kick
2. To disconnect

Synonyms
* (*kick*): kick
* (disconnect): kick



Wordsenses flattened...

Why?

Wordsenses flattened. . .

Why?

- no underlying linguistic theory, but:

Wordsenses flattened. . .

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource

Wordsenses flattened...

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations

Wordsenses flattened...

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Wordsenses flattened...

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations

Wordsenses flattened...

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* buskin, mukluk
* kick

?

Wordsenses flattened...

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations
- wordsenses of the target vertices?

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* buskin, mukluk
* kick

?

Wordsenses flattened...

Why?

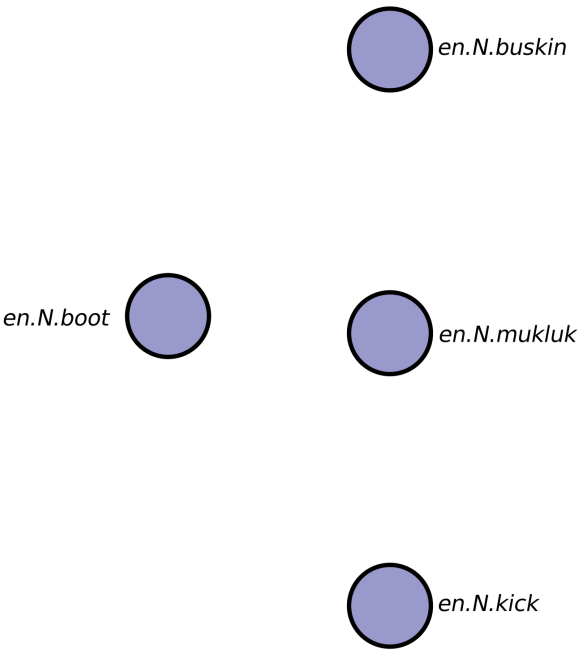
- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations
- wordsenses of the target vertices?

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* buskin, mukluk
* kick

?



Wordsenses flattened...

Why?

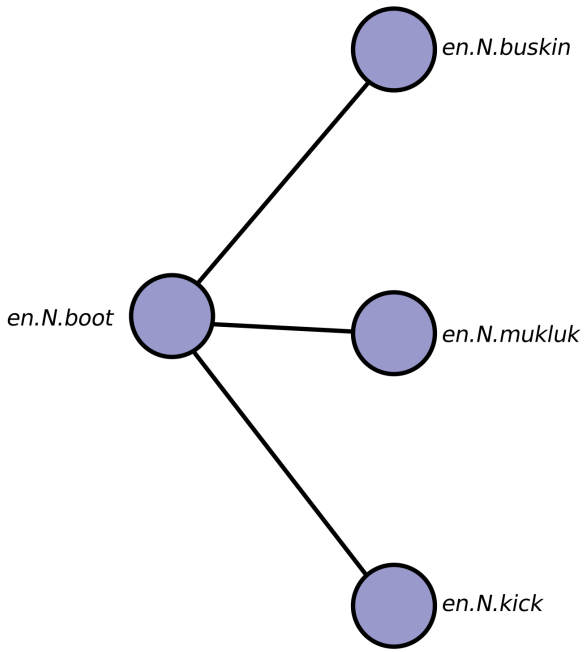
- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations
- wordsenses of the target vertices?

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* buskin, mukluk
* kick

?



Wordsenses flattened...

Why?

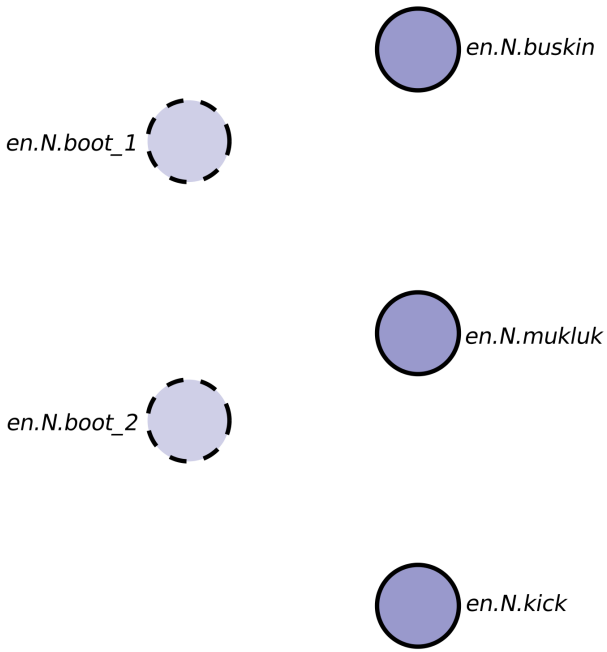
- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations
- wordsenses of the target vertices?

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* buskin, mukluk
* kick

?



Wordsenses flattened...

Why?

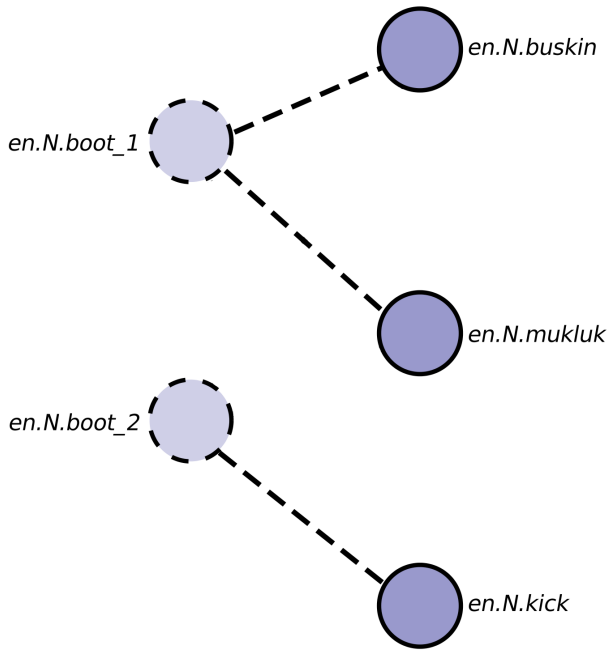
- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations
- wordsenses of the target vertices?

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* buskin, mukluk
* kick

?



Wordsenses flattened...

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations
- wordsenses of the target vertices?

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
? * buskin, mukluk
* kick

buskin (N)

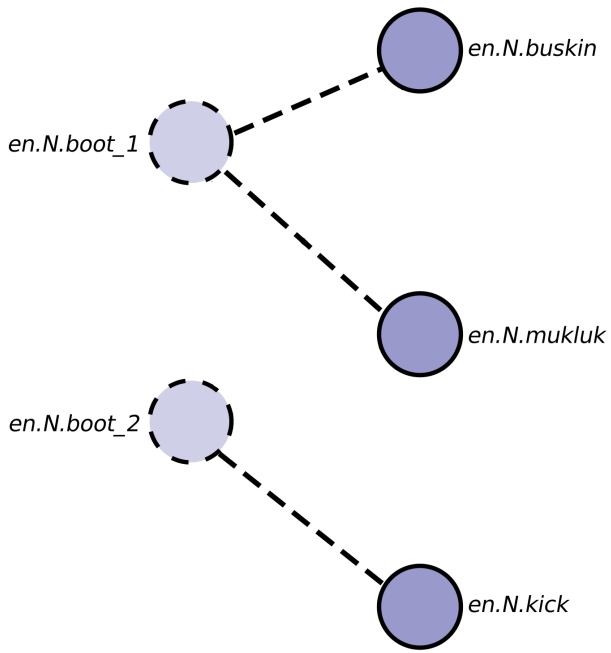
1. A half-boot
2. A type of boot worn by the ancient Athenian tragic actors

mukluk (N)

1. A soft boot made of reindeer skin or sealskin and worn by Inuit.

kick (N)

1. A hit or strike with the leg or foot
2. The action of swinging a foot or leg
3. Sth that tickles the fancy
4. (Internet) The removal of a person from an online activity
5. (figuratively) Any bucking motion of an object that lacks legs or feet



Wordsenses flattened...

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations
- wordsenses of the target vertices?

buskin (N)

1. A half-boot
2. A type of boot worn by the ancient Athenian tragic actors

mukluk (N)

1. A soft boot made of reindeer skin or sealskin and worn by Inuit.

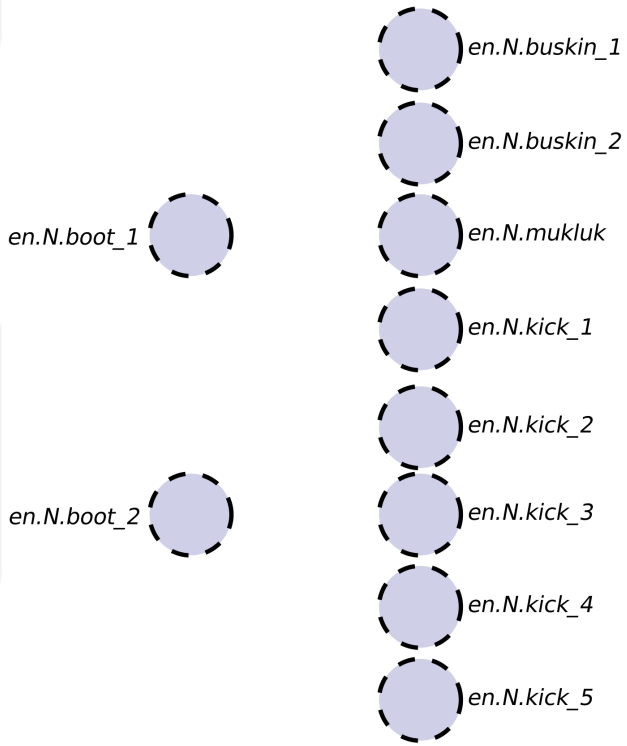
kick (N)

1. A hit or strike with the leg or foot
2. The action of swinging a foot or leg
3. Sth that tickles the fancy
4. (Internet) The removal of a person from an online activity
5. (figuratively) Any bucking motion of an object that lacks legs or feet

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
? * buskin, mukluk
* kick



Wordsenses flattened...

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations
- wordsenses of the target vertices?

buskin (N)

1. A half-boot
2. A type of boot worn by the ancient Athenian tragic actors

mukluk (N)

1. A soft boot made of reindeer skin or sealskin and worn by Inuit.

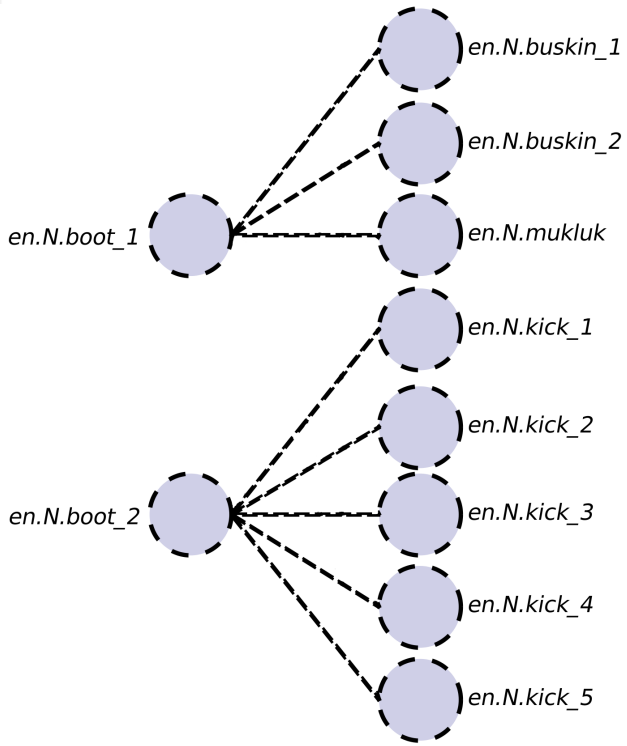
kick (N)

1. A hit or strike with the leg or foot
2. The action of swinging a foot or leg
3. Sth that tickles the fancy
4. (Internet) The removal of a person from an online activity
5. (figuratively) Any bucking motion of an object that lacks legs or feet

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
? * buskin, mukluk
* kick



Wordsenses flattened...

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations
- wordsenses of the target vertices?

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* buskin, mukluk
* kick

?

buskin (N)

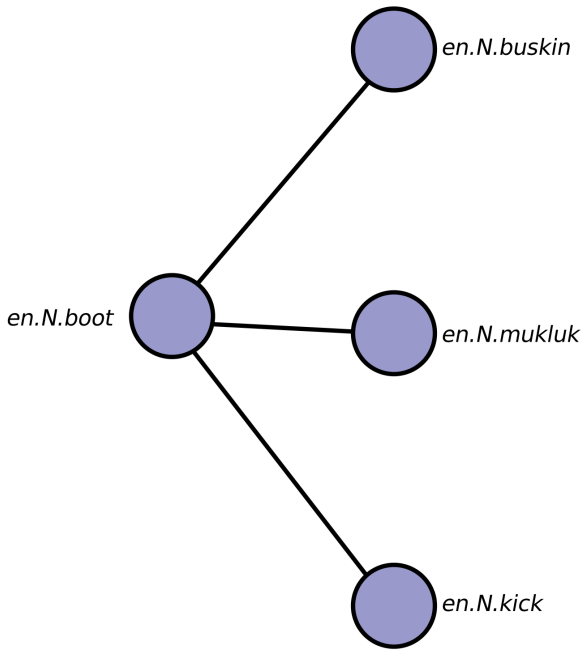
1. A half-boot
2. A type of boot worn by the ancient Athenian tragic actors

mukluk (N)

1. A soft boot made of reindeer skin or sealskin and worn by Inuit.

kick (N)

1. A hit or strike with the leg or foot
2. The action of swinging a foot or leg
3. Sth that tickles the fancy
4. (Internet) The removal of a person from an online activity
5. (figuratively) Any bucking motion of an object that lacks legs or feet



Wordsenses flattened...

Why?

- no underlying linguistic theory, but:
- constraints stem from the resource
- wordsenses may appear in the definitions and not in the semantic relations
- wordsenses of the target vertices?

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

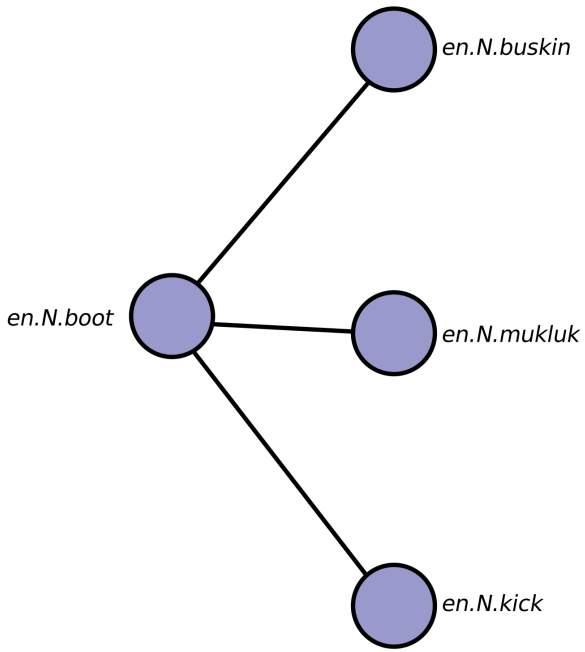
Synonyms
* buskin, mukluk
* kick

?

buskin (N)
1. A half-boot
2. A type of boot worn by the ancient Athenian tragic actors

mukluk (N)
1. A soft boot made of reindeer skin or sealskin and worn by Inuit.

kick (N)
1. A hit or strike with the leg or foot
2. The action of swinging a foot or leg
3. Sth that tickles the fancy
4. (Internet) The removal of a person from an online activity
5. (figuratively) Any bucking motion of an object that lacks legs or feet



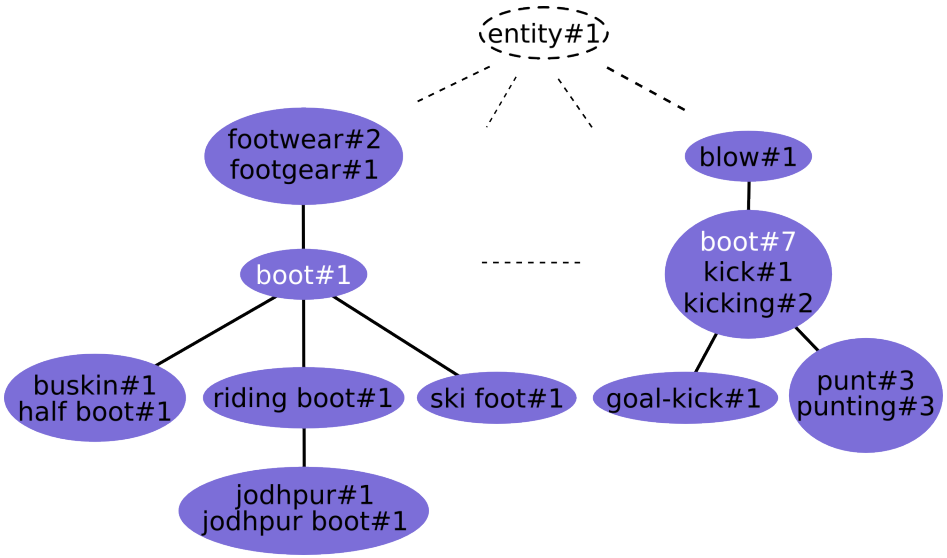
Another reason:

One of our gold standard (Dicosyn) has its wordsenses flattened

Extracting WordNet's synonymy network

WordNet

- synonymy between wordsenses
- relations already symmetric
- same POS in a given synset



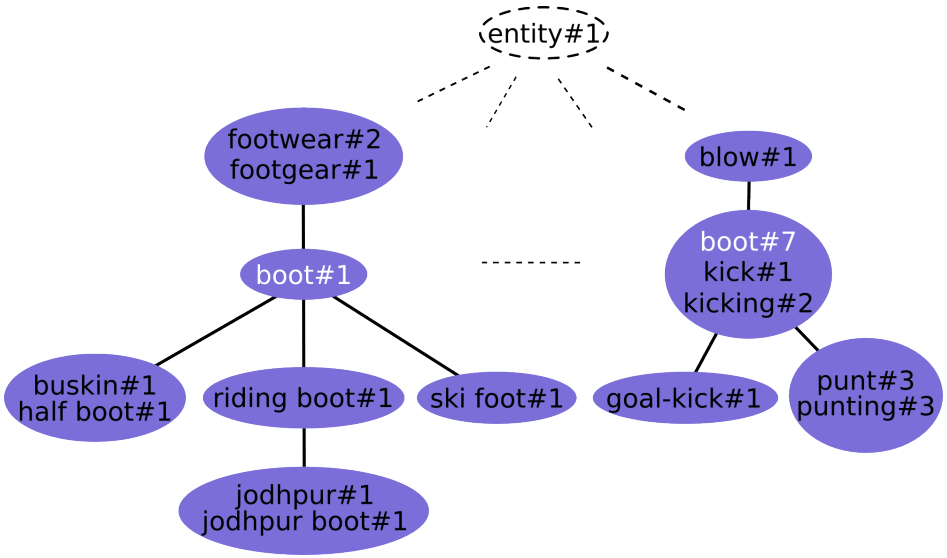
Extracting WordNet's synonymy network

WordNet

- synonymy between wordsenses
- relations already symmetric
- same POS in a given synset

Modeling

- vertices: words
- edges between all words in a given synset



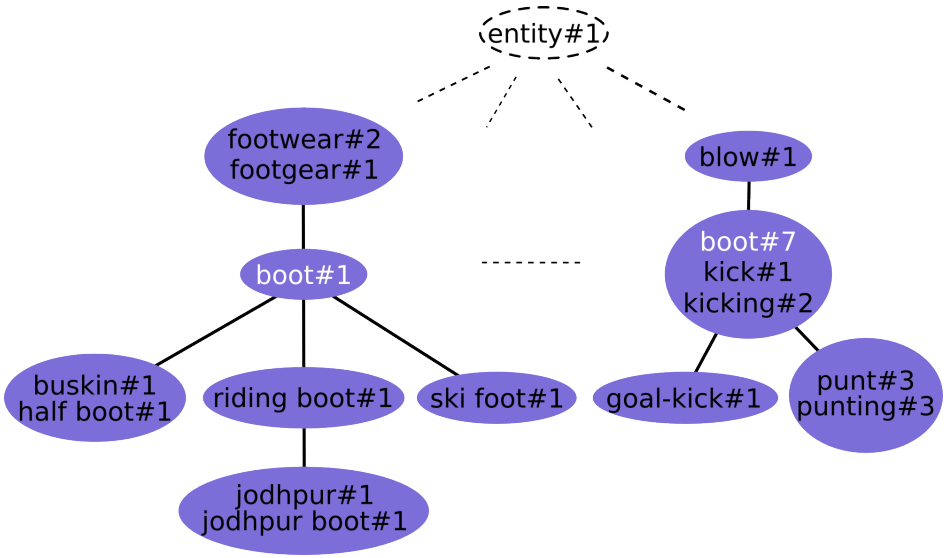
Extracting WordNet's synonymy network

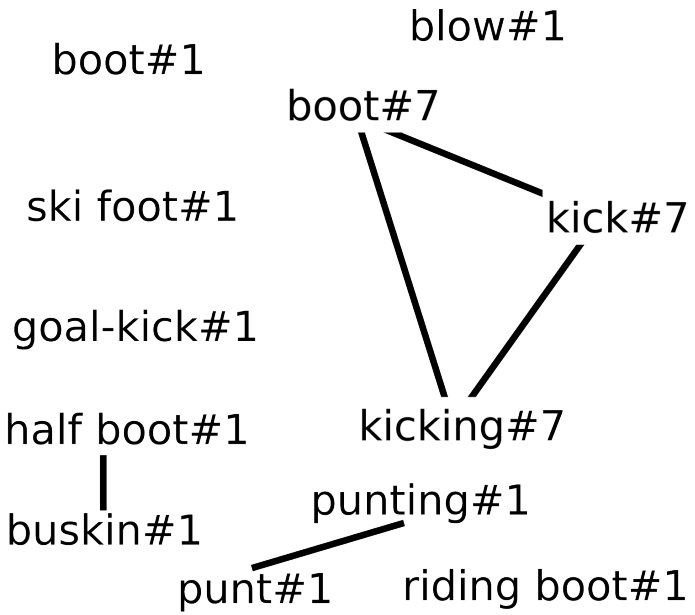
WordNet

- synonymy between wordsenses
- relations already symmetric
- same POS in a given synset

Modeling

- vertices: words
- edges between all words in a given synset





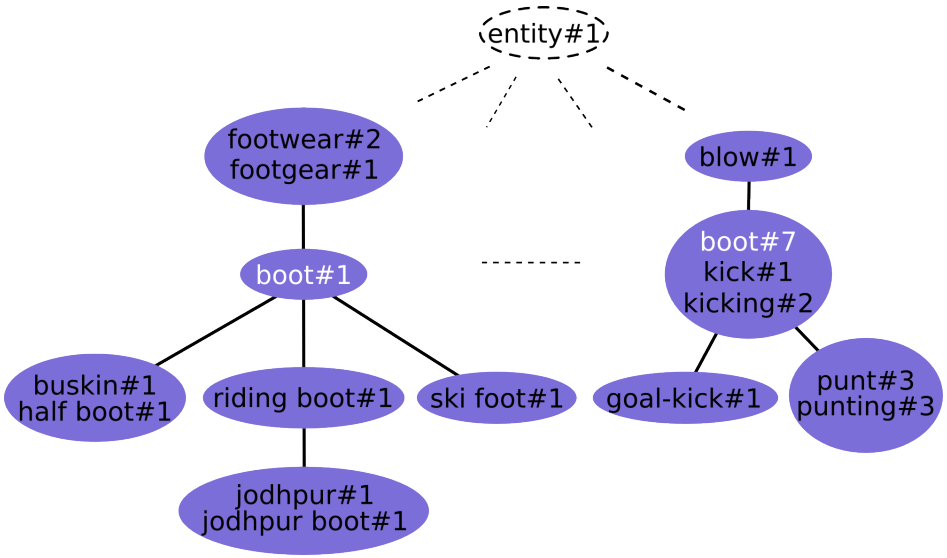
Extracting WordNet's synonymy network

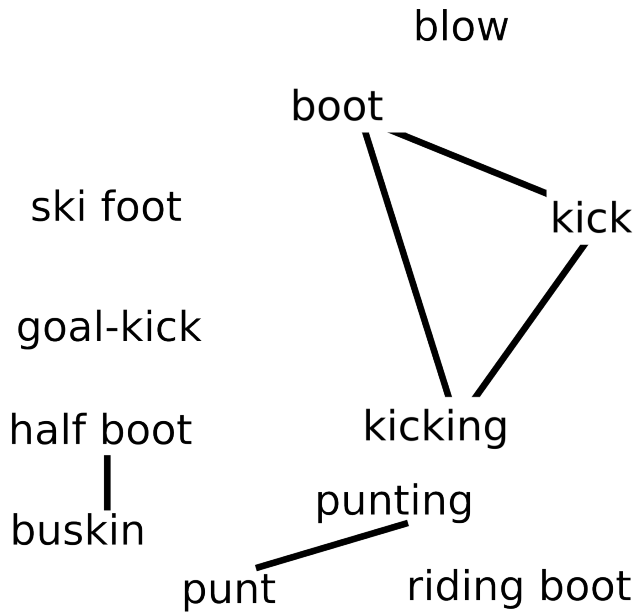
WordNet

- synonymy between wordsenses
- relations already symmetric
- same POS in a given synset

Modeling

- vertices: words
- edges between all words in a given synset





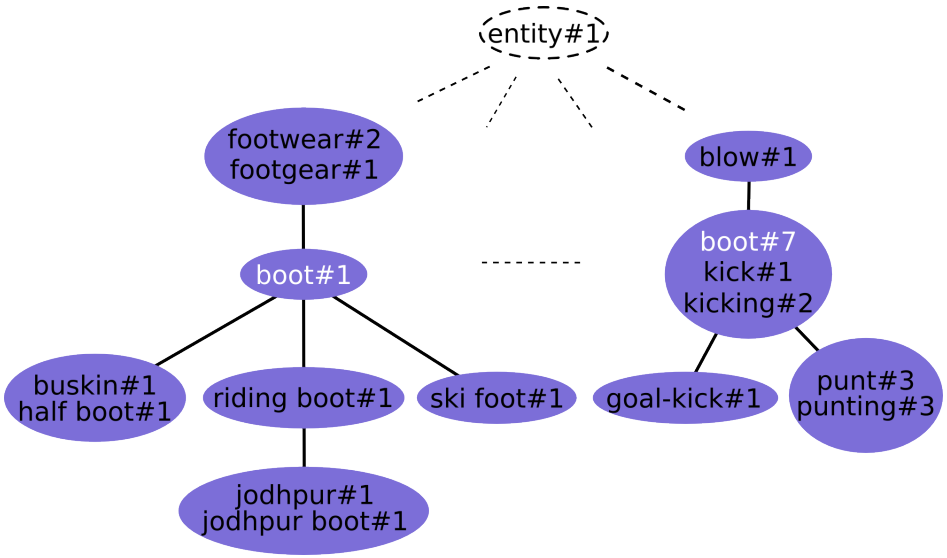
Extracting WordNet's synonymy network

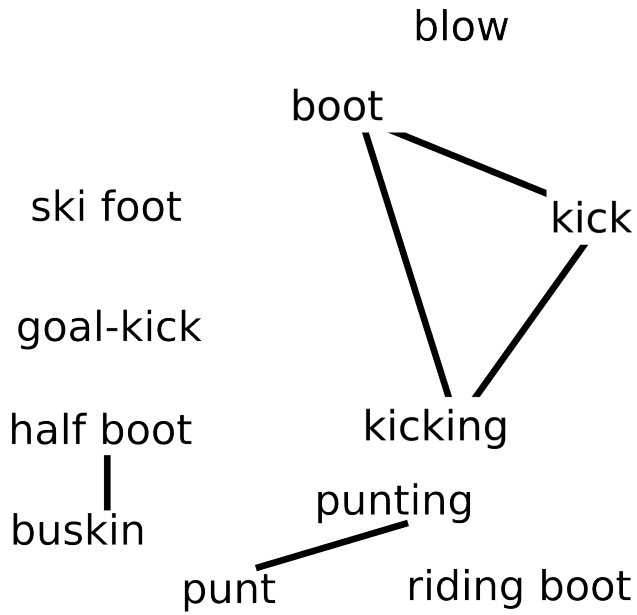
WordNet

- synonymy between wordsenses
- relations already symmetric
- same POS in a given synset

Modeling

- vertices: words
- edges between all words in a given synset
- + using hyponymy with leave synsets containing single-words





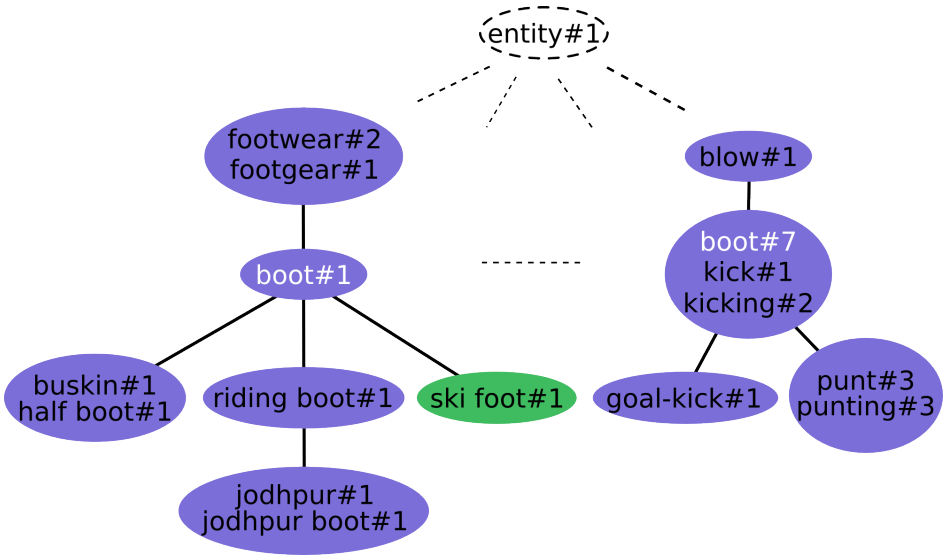
Extracting WordNet's synonymy network

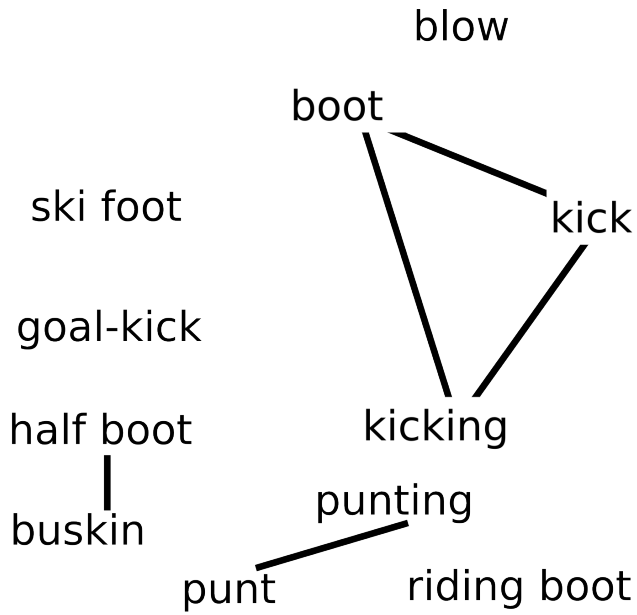
WordNet

- synonymy between wordsenses
- relations already symmetric
- same POS in a given synset

Modeling

- vertices: words
- edges between all words in a given synset
- + using hyponymy with leave synsets containing single-words





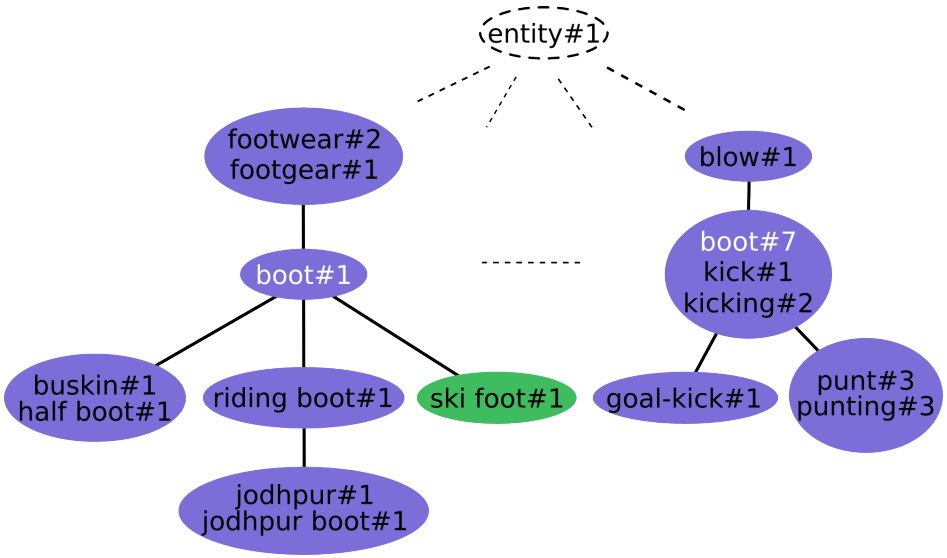
Extracting WordNet's synonymy network

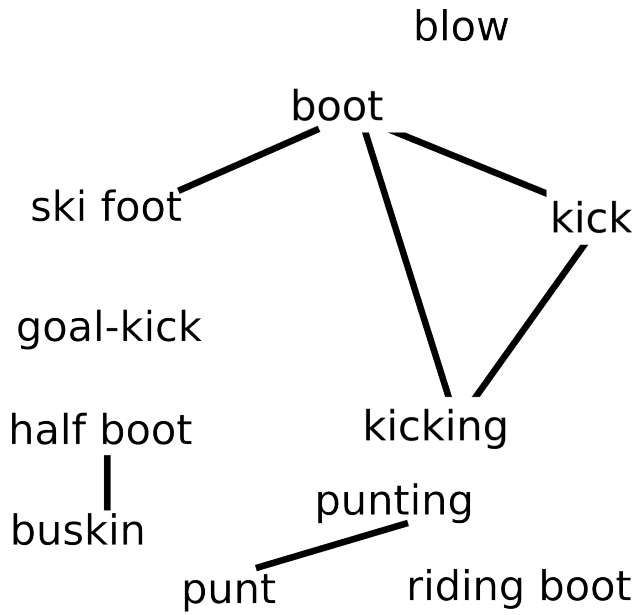
WordNet

- synonymy between wordsenses
- relations already symmetric
- same POS in a given synset

Modeling

- vertices: words
- edges between all words in a given synset
- + using hyponymy with leave synsets containing single-words





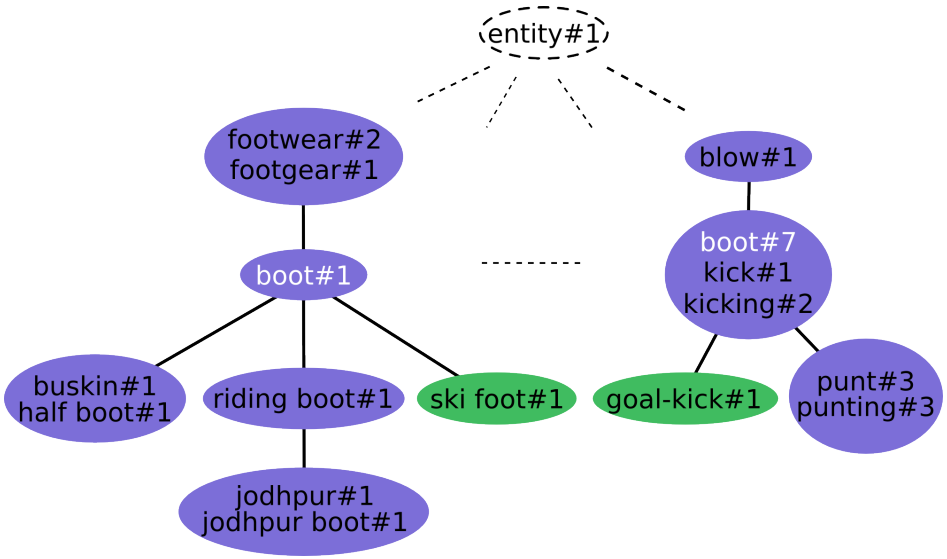
Extracting WordNet's synonymy network

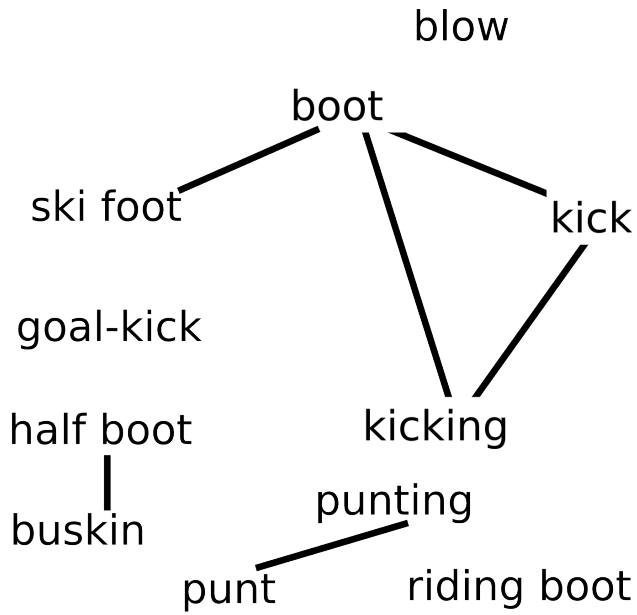
WordNet

- synonymy between wordsenses
- relations already symmetric
- same POS in a given synset

Modeling

- vertices: words
- edges between all words in a given synset
- + using hyponymy with leave synsets containing single-words





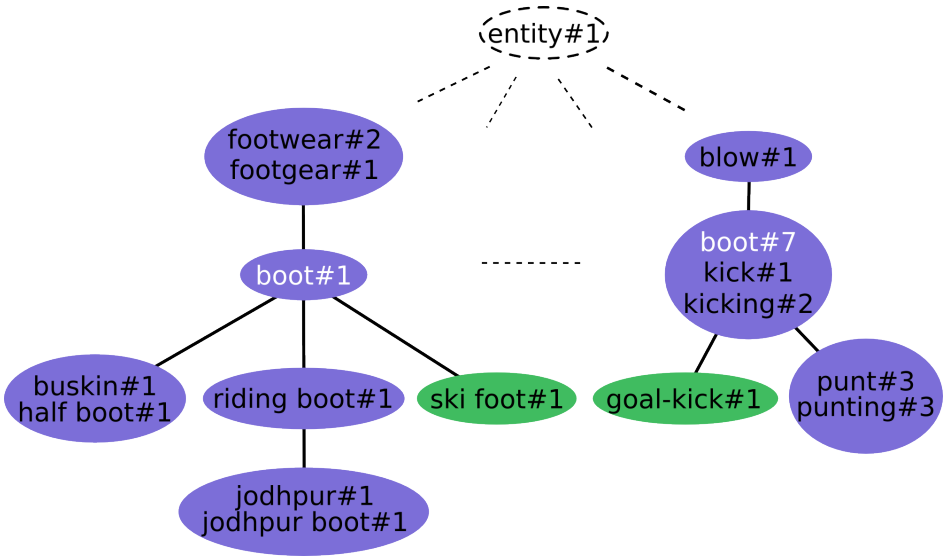
Extracting WordNet's synonymy network

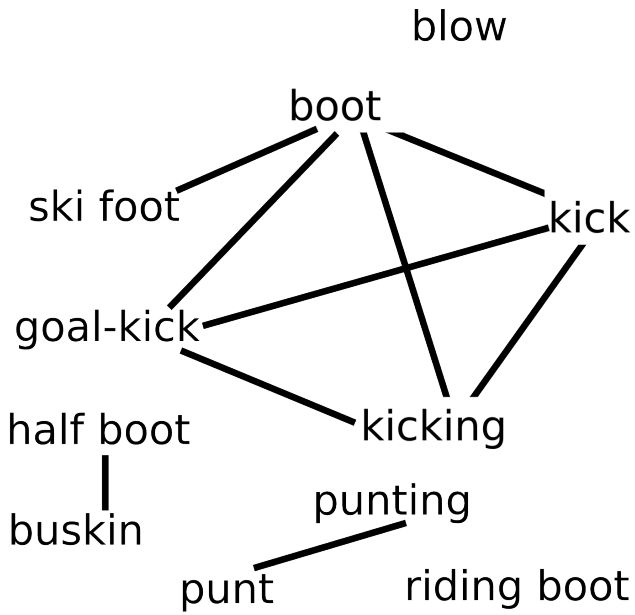
WordNet

- synonymy between wordsenses
- relations already symmetric
- same POS in a given synset

Modeling

- vertices: words
- edges between all words in a given synset
- + using hyponymy with leave synsets containing single-words





Extracting Dicosyn synonymy network

Dicosyn

- compilation of synonymy relations extracted from 7 dictionaries *(Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse and Robert)* ;
- produced at ATILF, corrected at CRISCO lab: `http://elsap1.unicaen.fr/dicosyn.html`
- wordsenses are flattened ;
- network already built ;
- just need to be symmetrized.

Small Worlds (SW)

Lexical resources are (often) SW

Small Worlds (SW)

Lexical resources are (often) SW

- globally sparses in edges, locally dense (high clustering) ;

Small Worlds (SW)

Lexical resources are (often) SW

- globally sparses in edges, locally dense (high clustering) ;
- average path length is short ;

Small Worlds (SW)

Lexical resources are (often) SW

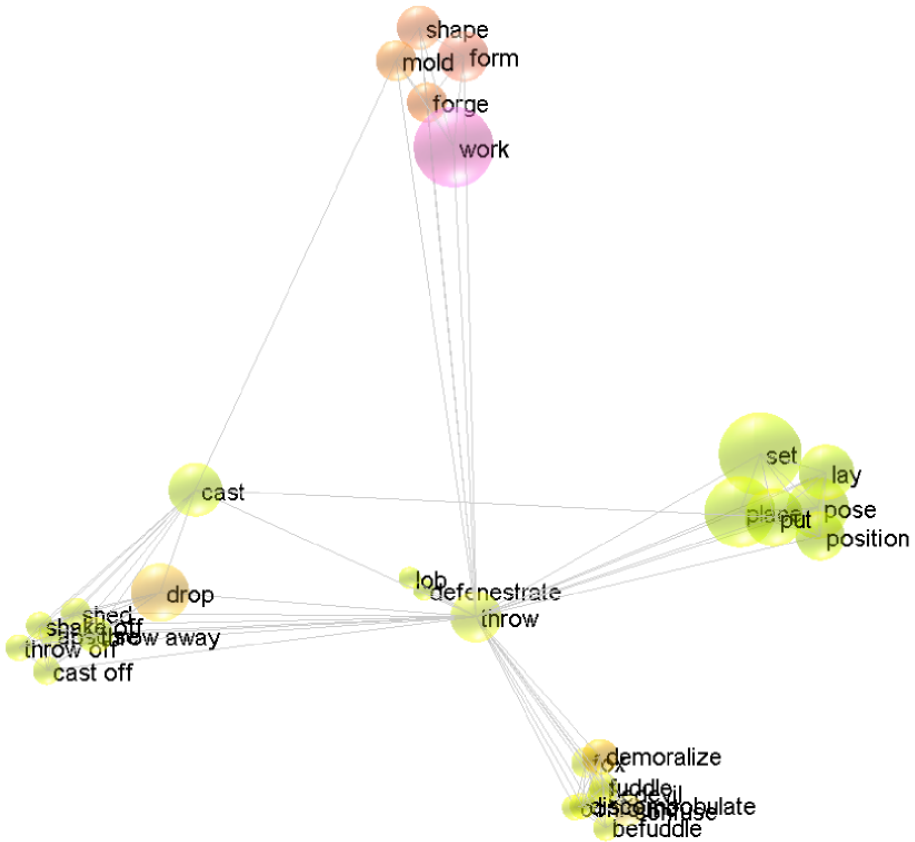
- globally sparses in edges, locally dense (high clustering) ;
- average path length is short ;
- heavy-tailed degree distribution.

Small Worlds (SW)

Lexical resources are (often) SW

- globally sparses in edges, locally dense (high clustering) ;
- average path length is short ;
- heavy-tailed degree distribution.

Eg.: to throw (WordNet)

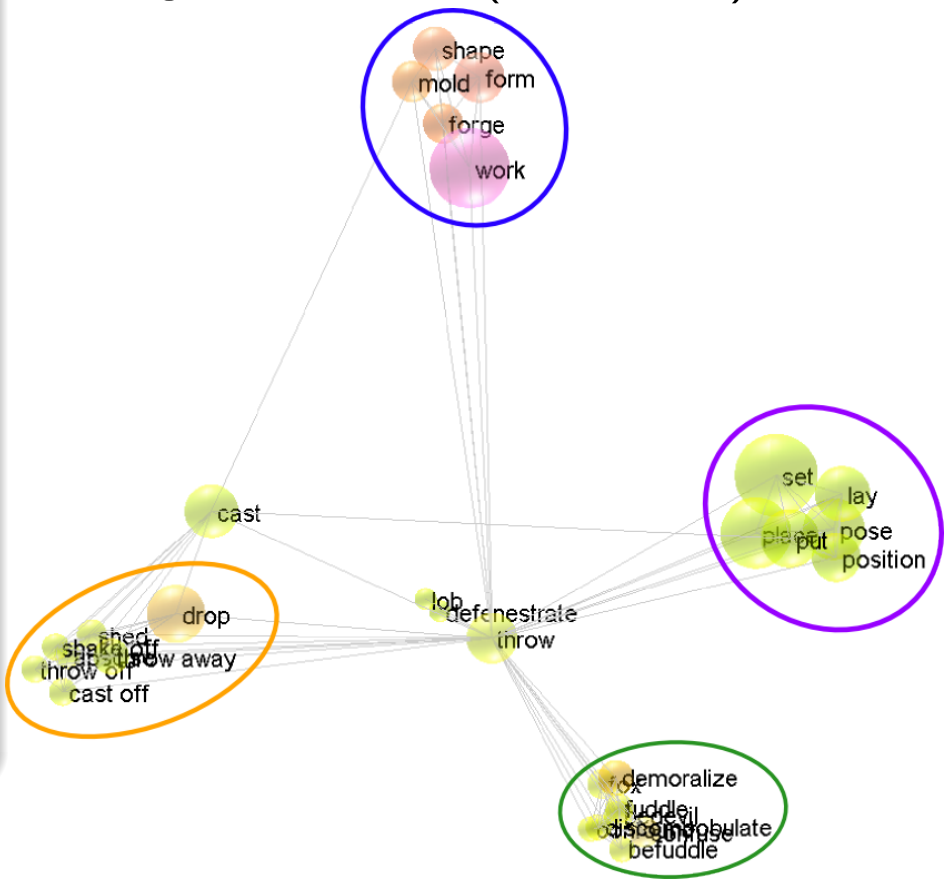


Small Worlds (SW)

Lexical resources are (often) SW

- globally sparses in edges, locally dense (high clustering) ;
- average path length is short ;
- heavy-tailed degree distribution.

Eg.: to throw (WordNet)

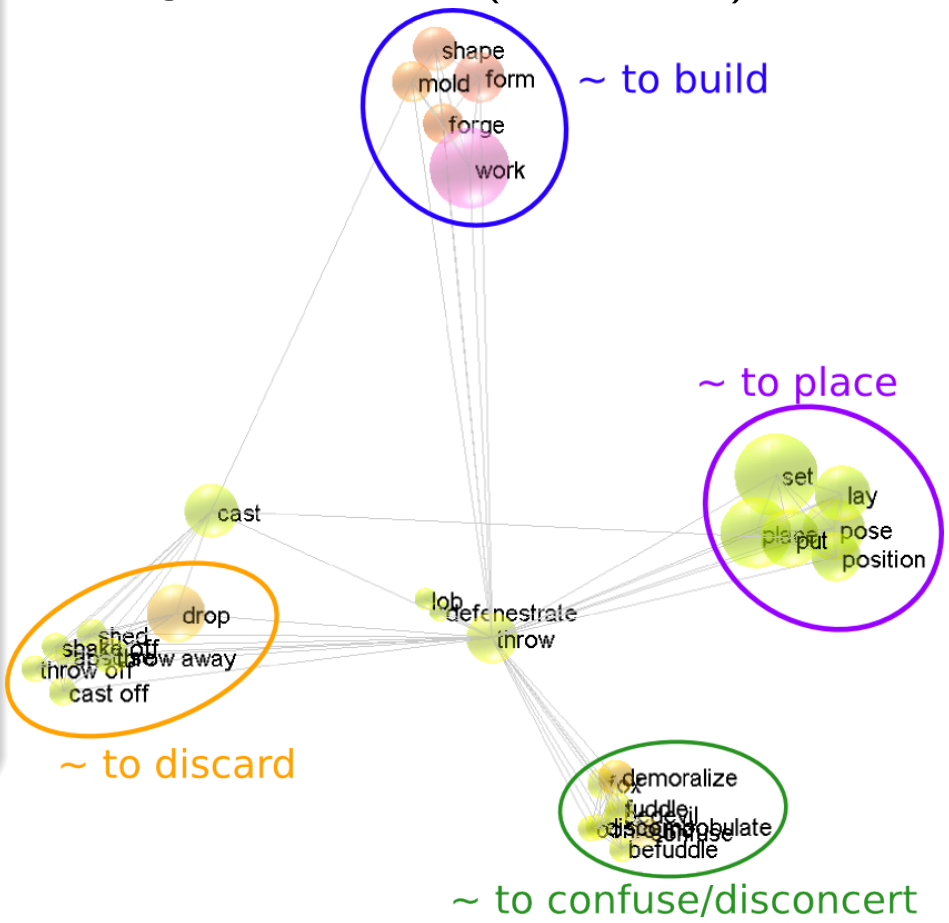


Small Worlds (SW)

Lexical resources are (often) SW

- globally sparses in edges, locally dense (high clustering) ;
- average path length is short ;
- heavy-tailed degree distribution.

Eg.: to throw (WordNet)

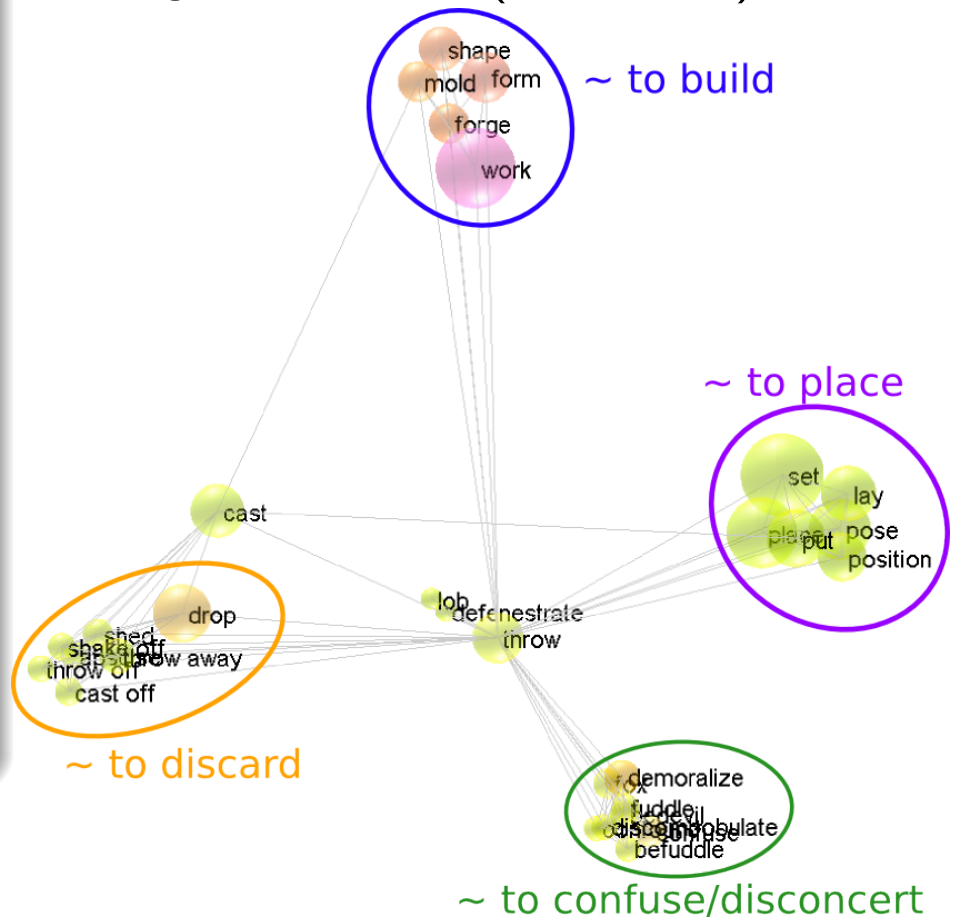


Small Worlds (SW)

Lexical resources are (often) SW

- globally sparses in edges, locally dense (high clustering) ;
- average path length is short ;
- heavy-tailed degree distribution.
- studying graphs' properties shows that Wiktionary, WordNet and Dicosyn are SW

Eg.: to throw (WordNet)

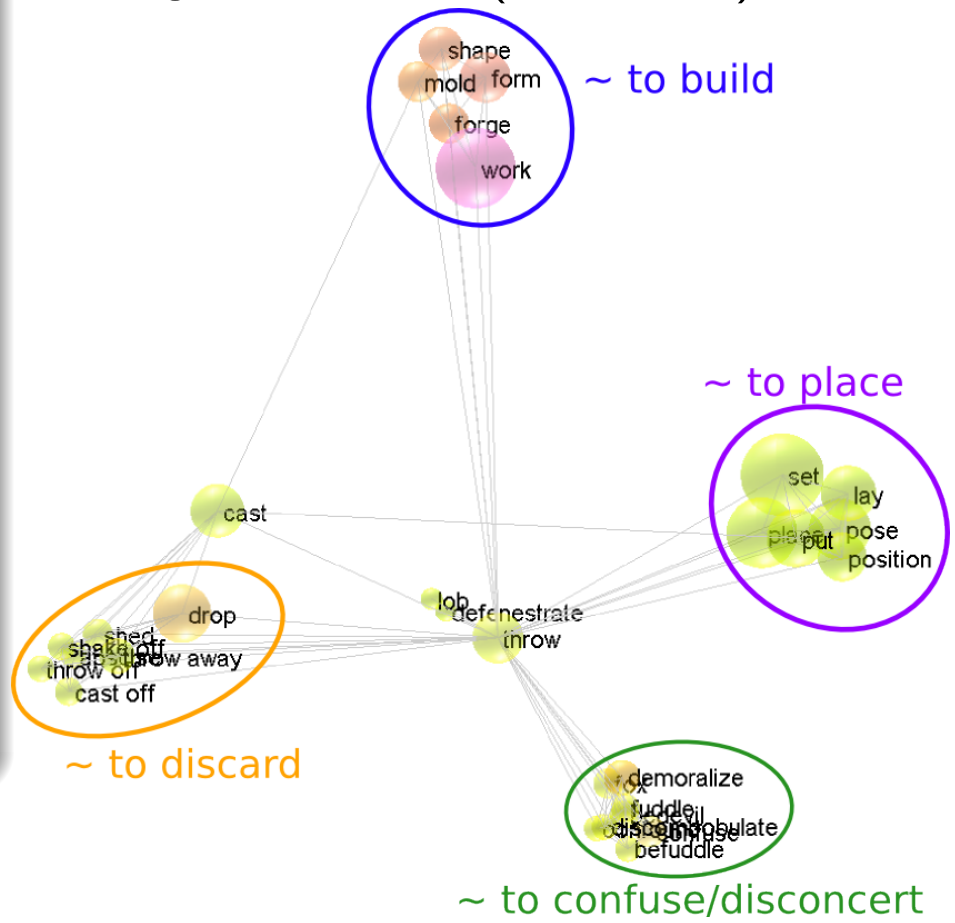


Small Worlds (SW)

Lexical resources are (often) SW

- globally sparses in edges, locally dense (high clustering) ;
- average path length is short ;
- heavy-tailed degree distribution.
- studying graphs' properties shows that Wiktionary, WordNet and Dicosyn are SW
- $\longrightarrow$ we can take advantage of SW's characteristics!

Eg.: to throw (WordNet)

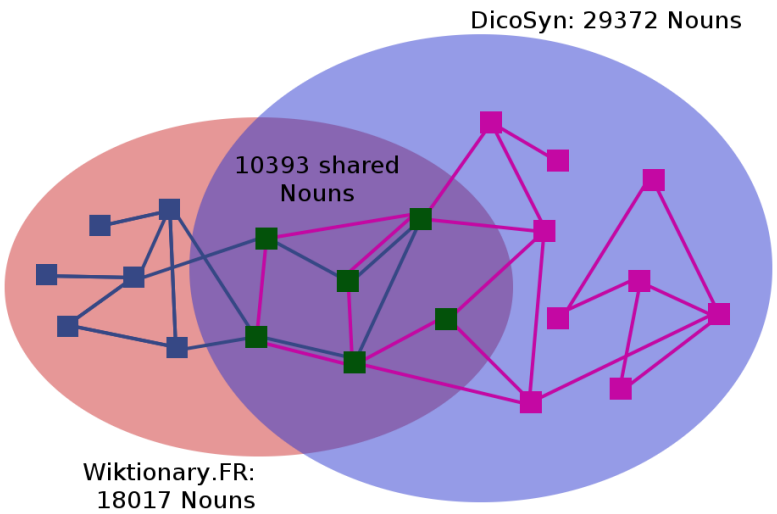


Wiktionary FR/Dicosyn

Lexical coverage/Synonymy network

	Words				
	Wikt.	DicoSyn	Shared	P	R
N	18017	29372	10393	58%	35%
A	5411	9452	3076	57%	33%
V	3897	9147	2966	76%	32%

Exemple: Nouns synonymy network



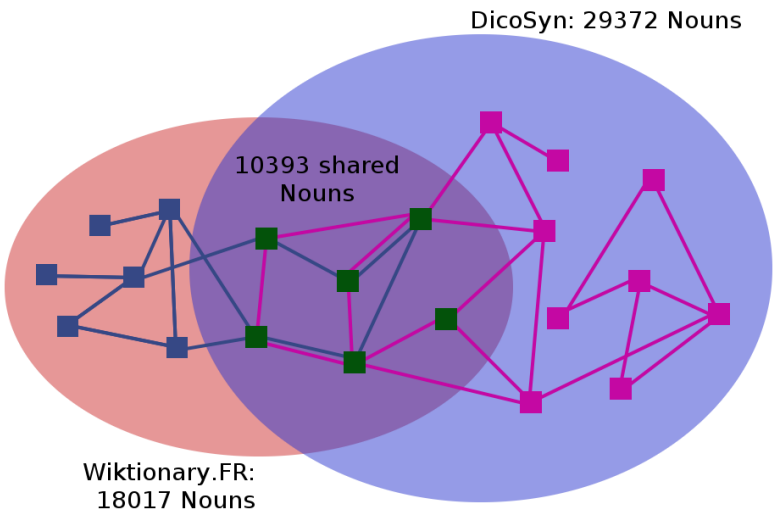
Wiktionary FR/Dicosyn

Lexical coverage/Synonymy network

	Words				
	Wikt.	DicoSyn	Shared	P	R
N	18017	29372	10393	58%	35%
A	5411	9452	3076	57%	33%
V	3897	9147	2966	76%	32%

Relations

Exemple: Nouns synonymy network



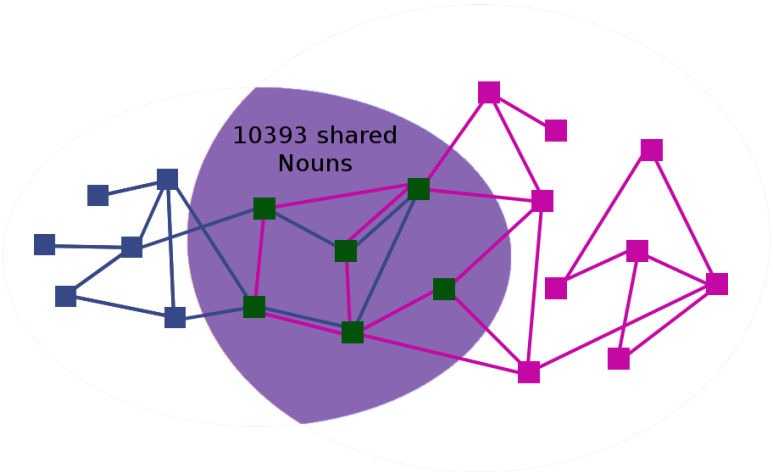
Wiktionary FR/Dicosyn

Lexical coverage/Synonymy network

	Words				
	Wikt.	DicoSyn	Shared	P	R
N	18017	29372	10393	58%	35%
A	5411	9452	3076	57%	33%
V	3897	9147	2966	76%	32%

Relations

Exemple: Nouns synonymy network



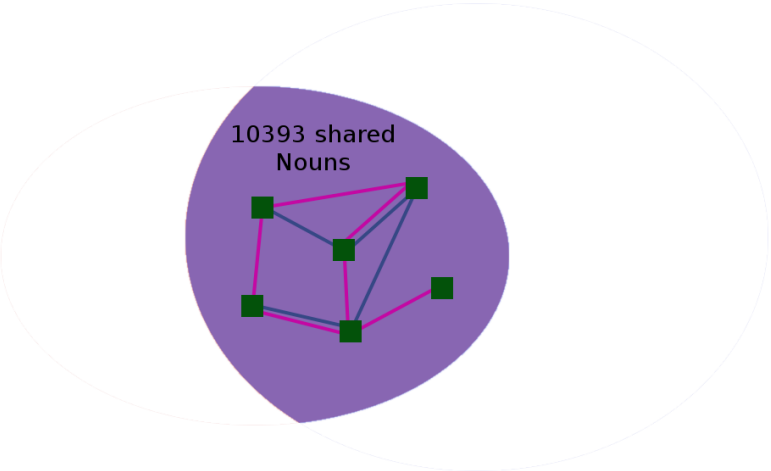
Wiktionary FR/Dicosyn

Lexical coverage/Synonymy network

	Words				
	Wikt.	DicoSyn	Shared	P	R
N	18017	29372	10393	58%	35%
A	5411	9452	3076	57%	33%
V	3897	9147	2966	76%	32%

Relations

Exemple: Nouns synonymy network



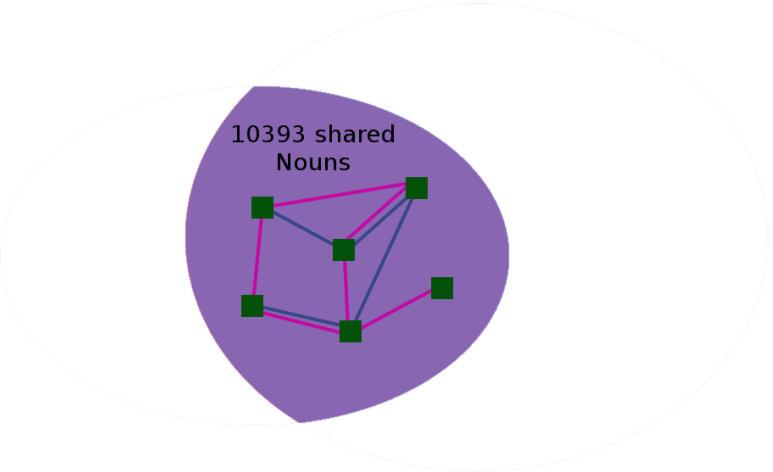
Wiktionary FR/Dicosyn

Lexical coverage/Synonymy network

	Words				
	Wikt.	DicoSyn	Shared	P	R
N	18017	29372	10393	58%	35%
A	5411	9452	3076	57%	33%
V	3897	9147	2966	76%	32%

	Relations				
	Wikt.	DicoSyn	Shared	P	R
N	3510	44501	3510	69%	8%
A	1300	17404	1677	78%	7%
V	899	23968	1267	71%	4%

Exemple: Nouns synonymy network



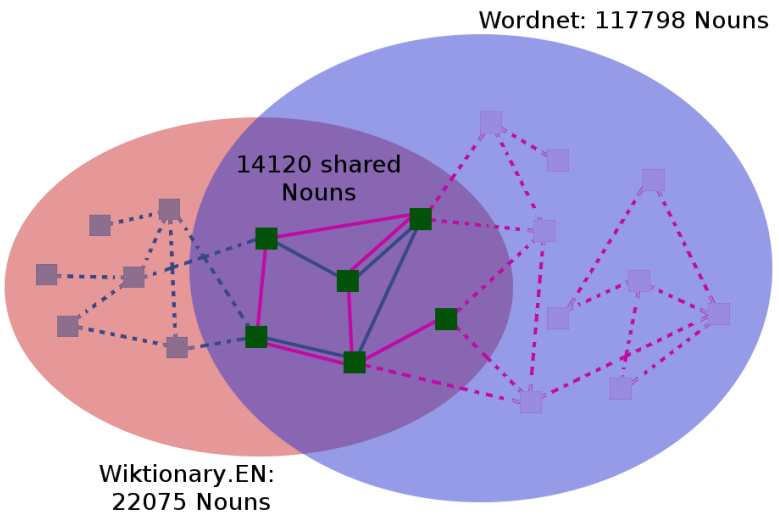
Wiktionary EN/WordNet

Lexical coverage/Synonymy network

	Words				
	Wikt.	WordNet	Shared	P	R
N	22075	117798	14120	64%	12%
A	8437	21479	5874	70%	27%
V	6368	11529	5157	81%	45%

	Relations				
	Wikt.	Wordnet	Shared	P	R
N	6453	18440	2763	43%	15%
A	3139	12792	1314	42%	10%
V	2667	18725	993	37%	5%

Exemple: Nouns synonymy network



Comments. . .

Gold standards, precision&recall

- a rough comparison, but:

Comments. . .

Gold standards, precision&recall

- a rough comparison, but:
- how meaningful are these measures (in our case)?

Comments. . .

Gold standards, precision&recall

- a rough comparison, but:
- how meaningful are these measures (in our case)?
- $\leftrightarrow$ how "perfect" are our gold standards?

Comments. . .

Gold standards, precision&recall

- a rough comparison, but:
- how meaningful are these measures (in our case)?
- $\leftrightarrow$ how "perfect" are our gold standards?
- examples:
  - *'to horn'*, in WN, absent from Wikt.en.N in our experiment, appeared in Wikt. in 2009

Comments. . .

Gold standards, precision&recall

- a rough comparison, but:
- how meaningful are these measures (in our case)?
- $\leftrightarrow$ how "perfect" are our gold standards?
- examples:
  - *'to horn'*, in WN, absent from Wikt.en.N in our experiment, appeared in Wikt. in 2009
  - *'to poo'* (childish), *'to prefetch'*, *'to google'*, (technical neologisms) are in Wikt and not in WN $\rightarrow$ noise?

Comments. . .

Gold standards, precision&recall

- a rough comparison, but:
- how meaningful are these measures (in our case)?
- $\leftrightarrow$ how "perfect" are our gold standards?
- examples:
  - *'to horn'*, in WN, absent from Wikt.en.N in our experiment, appeared in Wikt. in 2009
  - *'to poo'* (childish), *'to prefetch'*, *'to google'*, (technical neologisms) are in Wikt and not in WN $\rightarrow$ noise?
  - Wikt really misses some WN's relations: *'to act*$\leftrightarrow$*to play'*

Comments. . .

Gold standards, precision&recall

- a rough comparison, but:
- how meaningful are these measures (in our case)?
- $\leftrightarrow$ how "perfect" are our gold standards?
- examples:
  - *'to horn'*, in WN, absent from Wikt.en.N in our experiment, appeared in Wikt. in 2009
  - *'to poo'* (childish), *'to prefetch'*, *'to google'*, (technical neologisms) are in Wikt and not in WN $\rightarrow$ noise?
  - Wikt really misses some WN's relations: *'to act*$\leftrightarrow$*to play'*
  - but are relations from Wikt and absent from WN nessecarily errors: *'to reduce*$\leftrightarrow$*to decrease'*, *'to cook*$\leftrightarrow$*to microwave'* (all words appear in WN) $\rightarrow$ noise?

Comments. . .

Gold standards, precision&recall

- a rough comparison, but:
- how meaningful are these measures (in our case)?
- $\leftrightarrow$ how "perfect" are our gold standards?
- examples:
  - *'to horn'*, in WN, absent from Wikt.en.N in our experiment, appeared in Wikt. in 2009
  - *'to poo'* (childish), *'to prefetch'*, *'to google'*, (technical neologisms) are in Wikt and not in WN $\rightarrow$ noise?
  - Wikt really misses some WN's relations: *'to act*$\leftrightarrow$*to play'*
  - but are relations from Wikt and absent from WN nessecarily errors: *'to reduce*$\leftrightarrow$*to decrease'*, *'to cook*$\leftrightarrow$*to microwave'* (all words appear in WN) $\rightarrow$ noise?
- $\rightarrow$ *(we assume that)* with time, recall will grow

Comments. . .

Gold standards, precision&recall

- a rough comparison, but:
- how meaningful are these measures (in our case)?
- $\leftrightarrow$ how "perfect" are our gold standards?
- examples:
  - *'to horn'*, in WN, absent from Wikt.en.N in our experiment, appeared in Wikt. in 2009
  - *'to poo'* (childish), *'to prefetch'*, *'to google'*, (technical neologisms) are in Wikt and not in WN $\rightarrow$ noise?
  - Wikt really misses some WN's relations: *'to act*$\leftrightarrow$*to play'*
  - but are relations from Wikt and absent from WN nessecarily errors: *'to reduce*$\leftrightarrow$*to decrease'*, *'to cook*$\leftrightarrow$*to microwave'* (all words appear in WN) $\rightarrow$ noise?
- $\rightarrow$ *(we assume that)* with time, recall will grow
- $\rightarrow$ is it possible to (automatically) measure precision?

Neighbours method

Intuition of transitivity

- *"Neighbours of my neighbours should be in my neighbourhood"*

Neighbours method

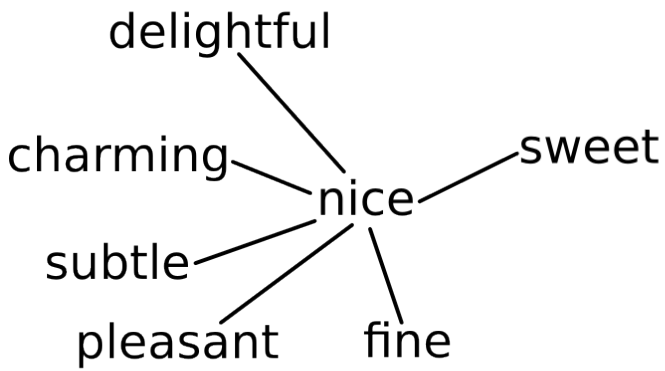
Intuition of transitivity

- *"Neighbours of my neighbours should be in my neighbourhood"*
- $\longrightarrow$ a neighbour of my neighbours which is not in my neighbourhood should be a good neighbour candidate.

Neighbours method

Intuition of transitivity

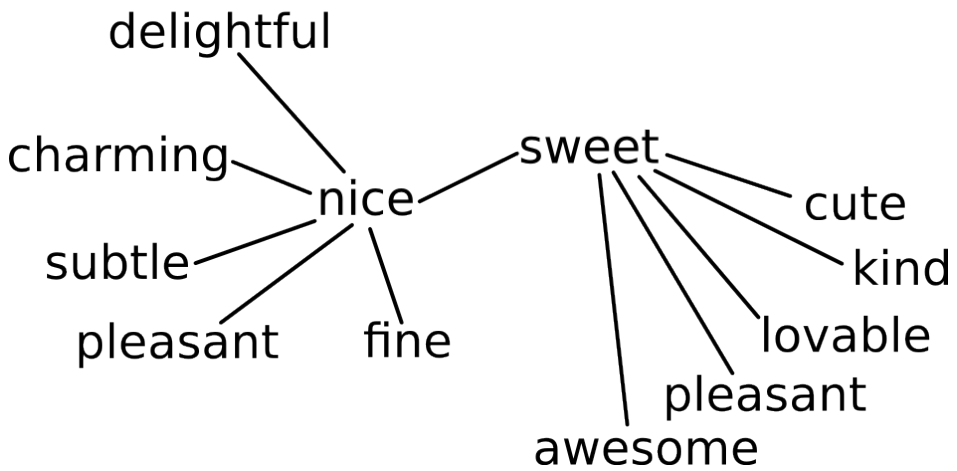
- *"Neighbours of my neighbours should be in my neighbourhood"*
- $\longrightarrow$ a neighbour of my neighbours which is not in my neighbourhood should be a good neighbour candidate.



Neighbours method

Intuition of transitivity

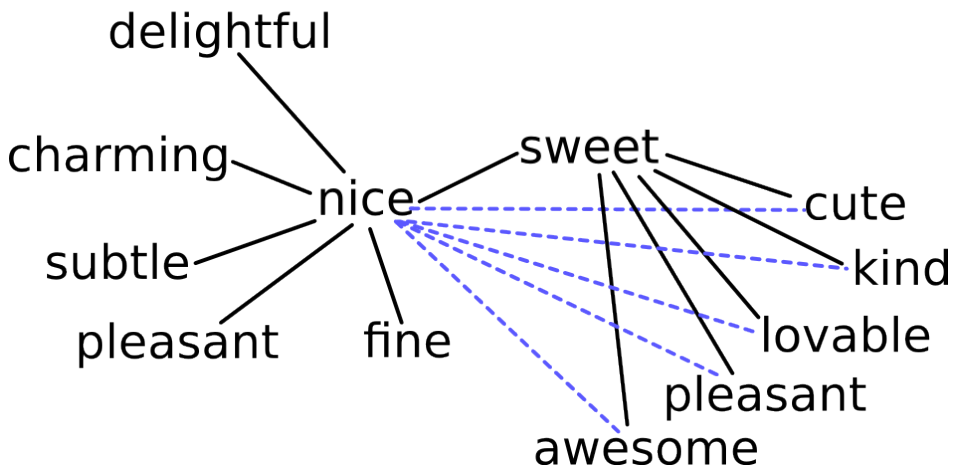
- *"Neighbours of my neighbours should be in my neighbourhood"*
- $\longrightarrow$ a neighbour of my neighbours which is not in my neighbourhood should be a good neighbour candidate.



Neighbours method

Intuition of transitivity

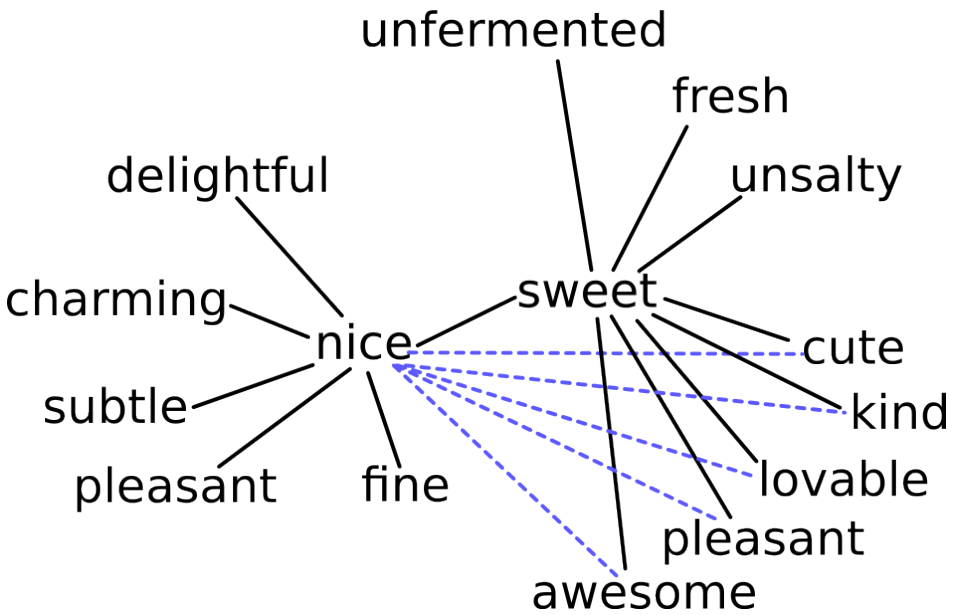
- *"Neighbours of my neighbours should be in my neighbourhood"*
- $\longrightarrow$ a neighbour of my neighbours which is not in my neighbourhood should be a good neighbour candidate.



Neighbours method

Intuition of transitivity

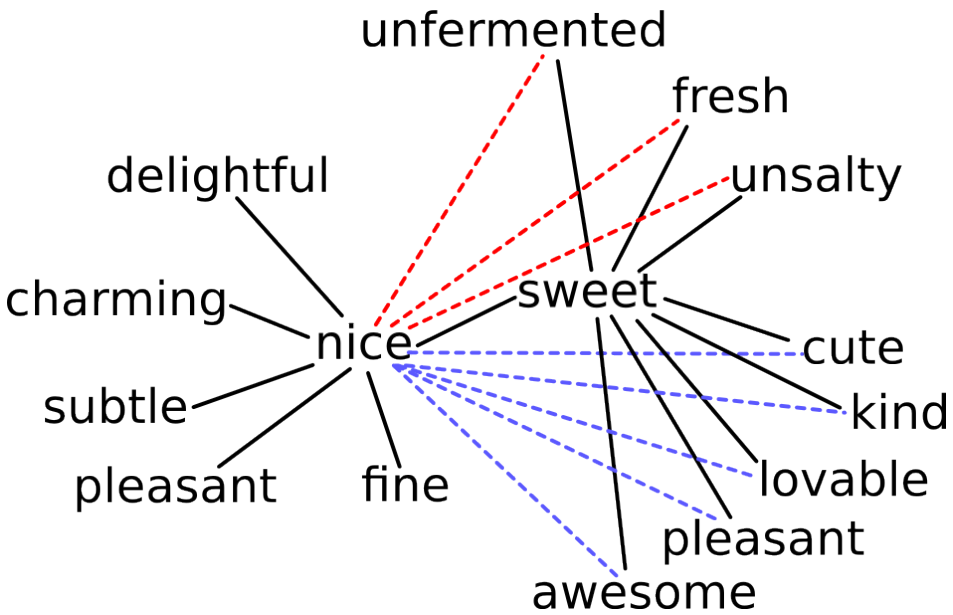
- *"Neighbours of my neighbours should be in my neighbourhood"*
- $\longrightarrow$ a neighbour of my neighbours which is not in my neighbourhood should be a good neighbour candidate.



Neighbours method

Intuition of transitivity

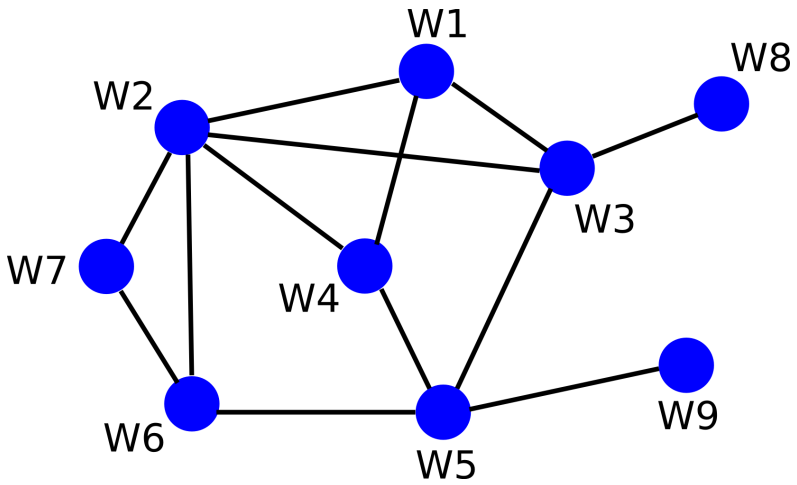
- *"Neighbours of my neighbours should be in my neighbourhood"*
- $\longrightarrow$ a neighbour of my neighbours which is not in my neighbourhood should be a good neighbour candidate.



Prox method

Prox (Gaume et al.)

- A stochastic method designed to study Hierarchical SW
- Metrics: for any 2 vertices (`u`, `v`), computes *"the probability that a randomly wandering particle starting from u stands in v after k steps."*

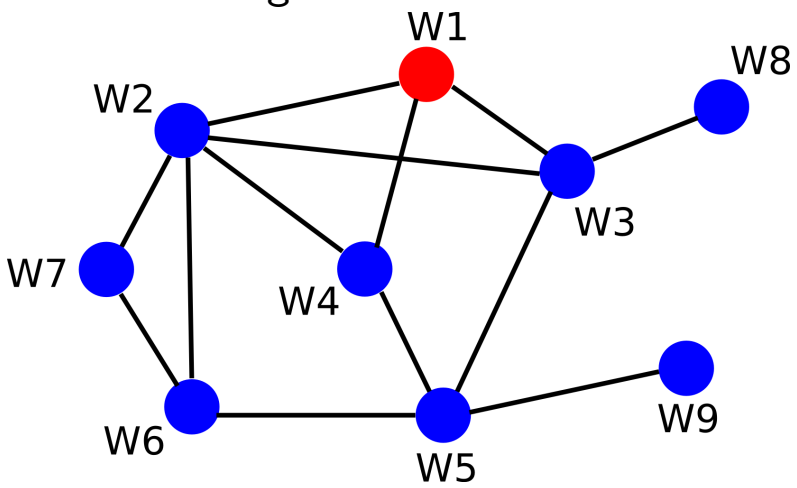


Prox method

Prox (Gaume et al.)

- A stochastic method designed to study Hierarchical SW
- Metrics: for any 2 vertices (u, v), computes *"the probability that a randomly wandering particle starting from u stands in v after k steps."*

Eg. from Word1

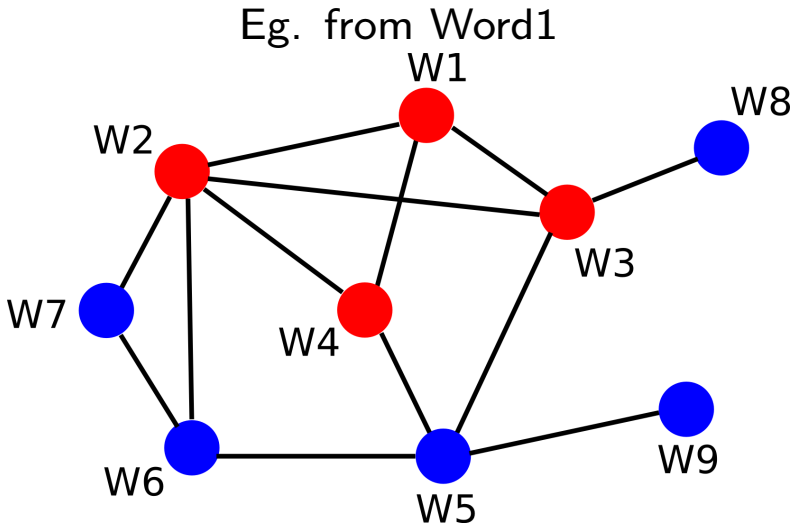


	1	2	3	4	5	6	7	8	9
Initial	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Prox method

Prox (Gaume et al.)

- A stochastic method designed to study Hierarchical SW
- Metrics: for any 2 vertices (u, v), computes *"the probability that a randomly wandering particle starting from u stands in v after k steps."*

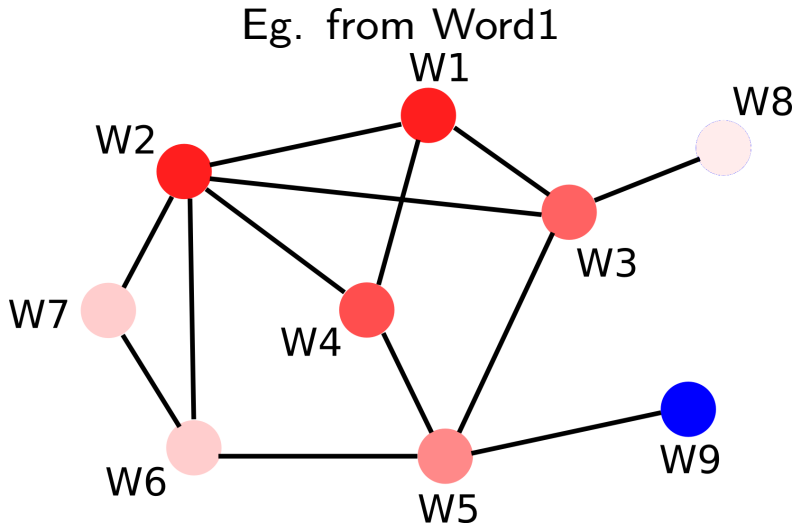


	1	2	3	4	5	6	7	8	9
Step 1	0.25	0.25	0.25	0.25	0.00	0.00	0.00	0.00	0.00

Prox method

Prox (Gaume et al.)

- A stochastic method designed to study Hierarchical SW
- Metrics: for any 2 vertices (u, v), computes *"the probability that a randomly wandering particle starting from u stands in v after k steps."*

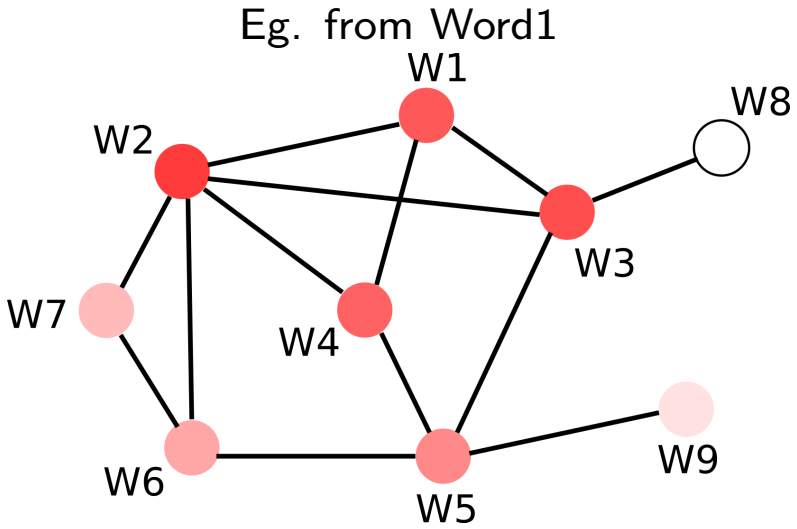


	1	2	3	4	5	6	7	8	9
Step 2	0.22	0.22	0.15	0.17	0.11	0.04	0.04	0.01	0.00

Prox method

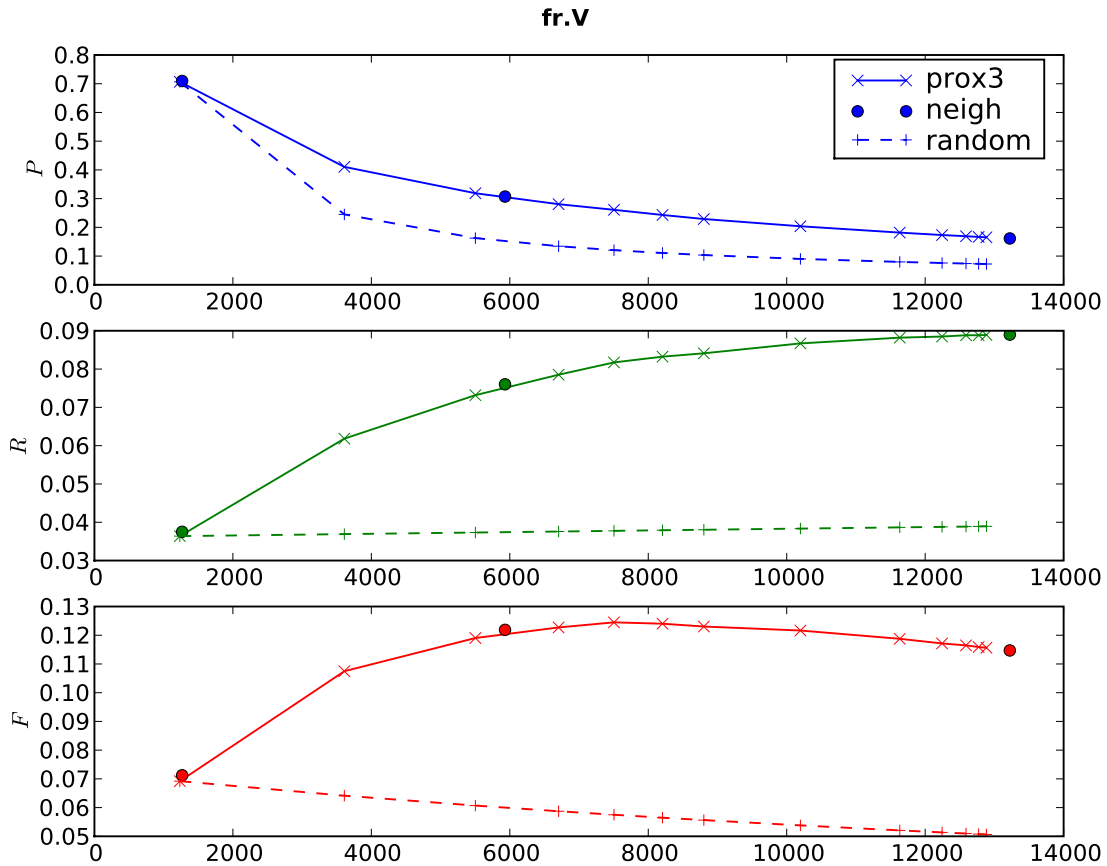
Prox (Gaume et al.)

- A stochastic method designed to study Hierarchical SW
- Metrics: for any 2 vertices (u, v), computes *"the probability that a randomly wandering particle starting from u stands in v after k steps."*

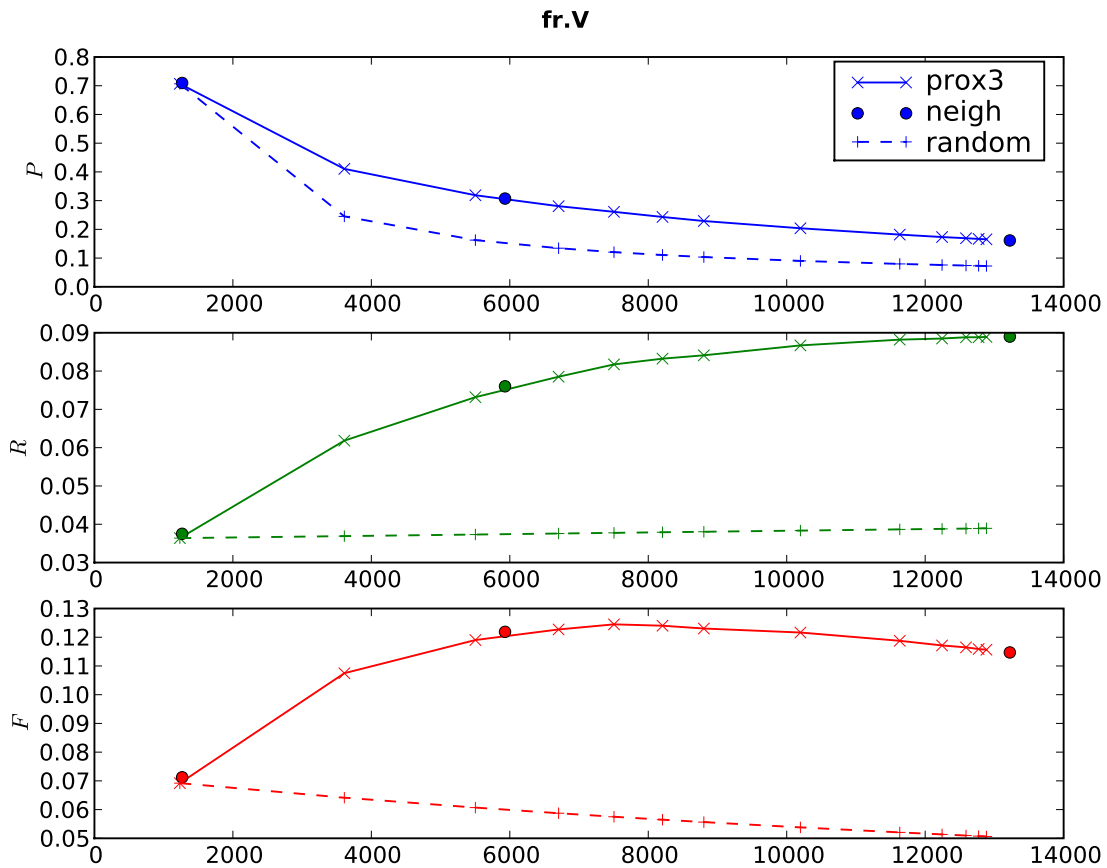


	1	2	3	4	5	6	7	8	9
Step 3	0.16	0.19	0.17	0.15	0.11	0.08	0.06	0.00	0.02

Results


fr.V
prox3
neigh
random
P
R
F
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0
2000
4000
6000
8000
10000
12000
14000

Results



Comments

- Prox method provides (ordered) relevant links eg. *'to absolve'*↔*'to forgive'*, absent from WN
- false positives may be intersting to consider:
  - *'to uncover'*↔*'to peel'* (hypernymy)
  - *'to skin'*↔*'to peel'* ('inter-domain synonymy')

Translation links method

Intuition

2 words sharing many translations in different languages are likely to be synonymous

Translation links method

Intuition

2 words sharing many translations in different languages are likely to be synonymous

Method

- let T_w be the set of a word w's translations
- for every pair of words (w,w'):

$$Jaccard(w, w') = \frac{|T_w \cap T_{w'}|}{|T_w \cup T_{w'}|}$$

- incrementally add relations, according to the Jaccard rank, up to a given threshold

Translation links method

Intuition

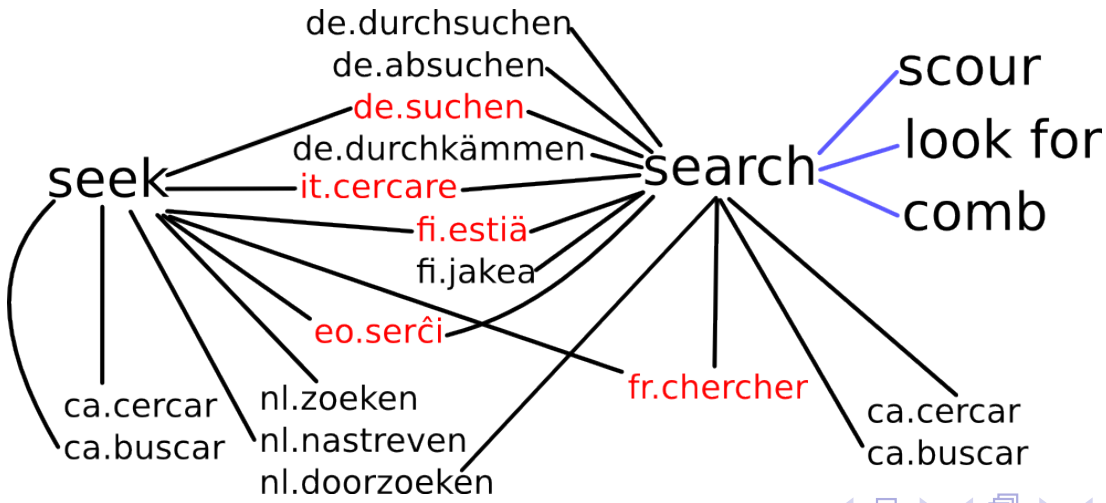
2 words sharing many translations in different languages are likely to be synonymous

Method

- let T_w be the set of a word w's translations
- for every pair of words (w, w'):

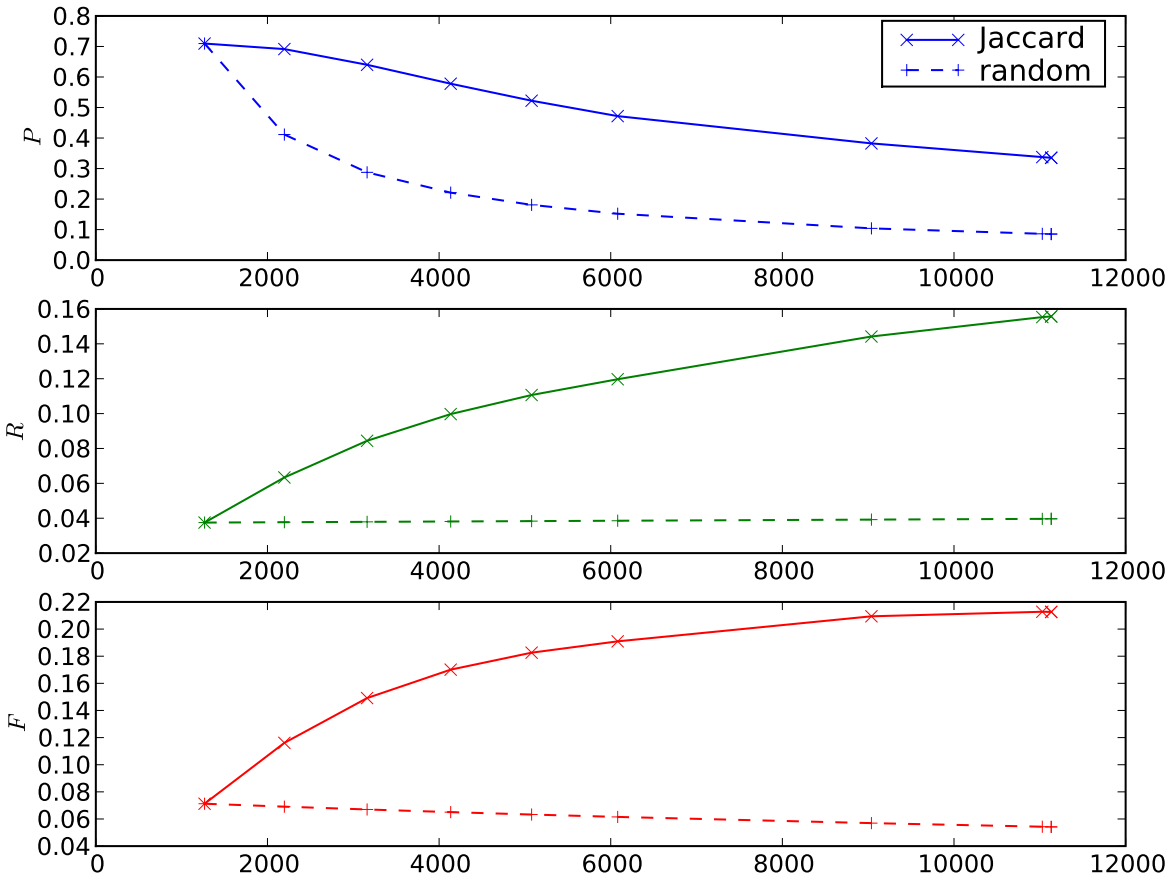
$$Jaccard(w, w') = \frac{|T_w \cap T_{w'}|}{|T_w \cup T_{w'}|}$$

- incrementally add relations, according to the Jaccard rank, up to a given threshold

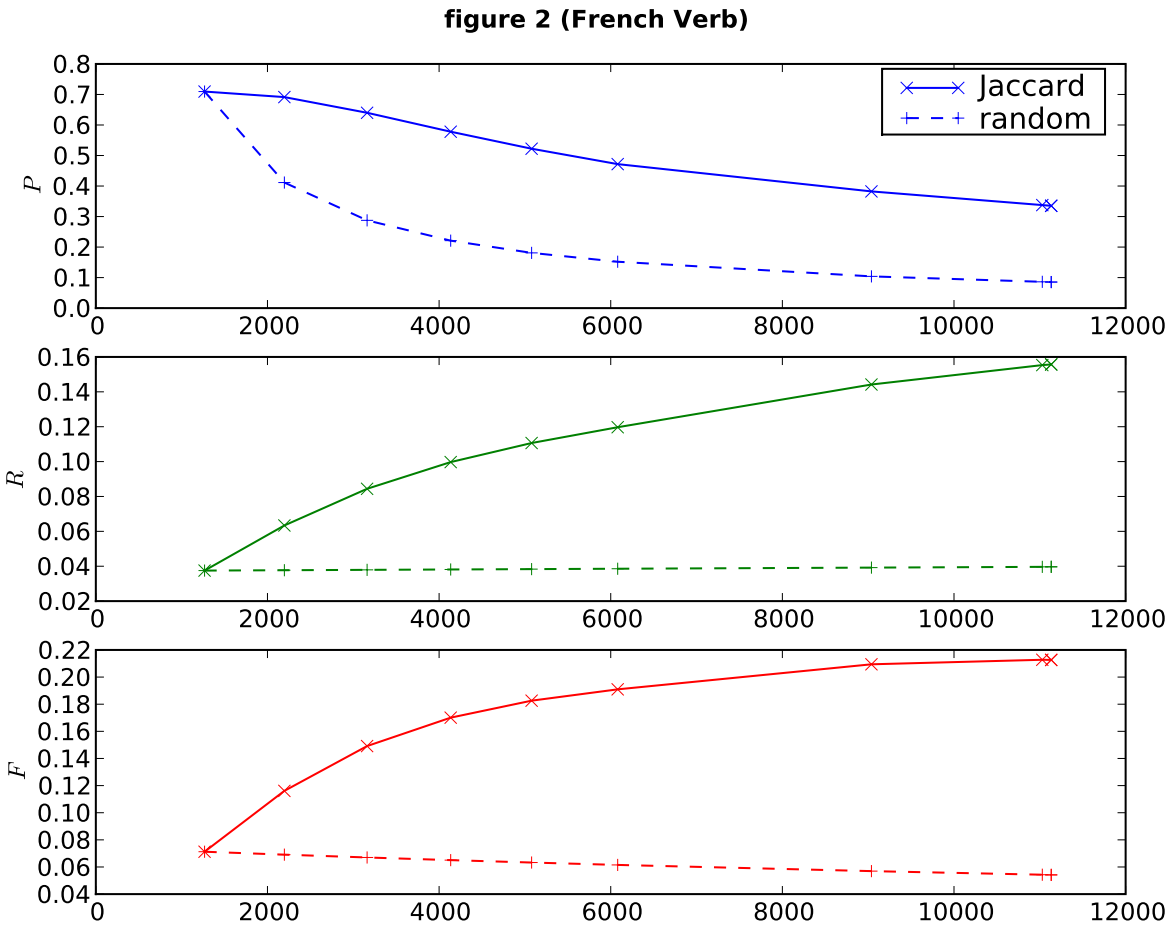


Results

figure 2 (French Verb)



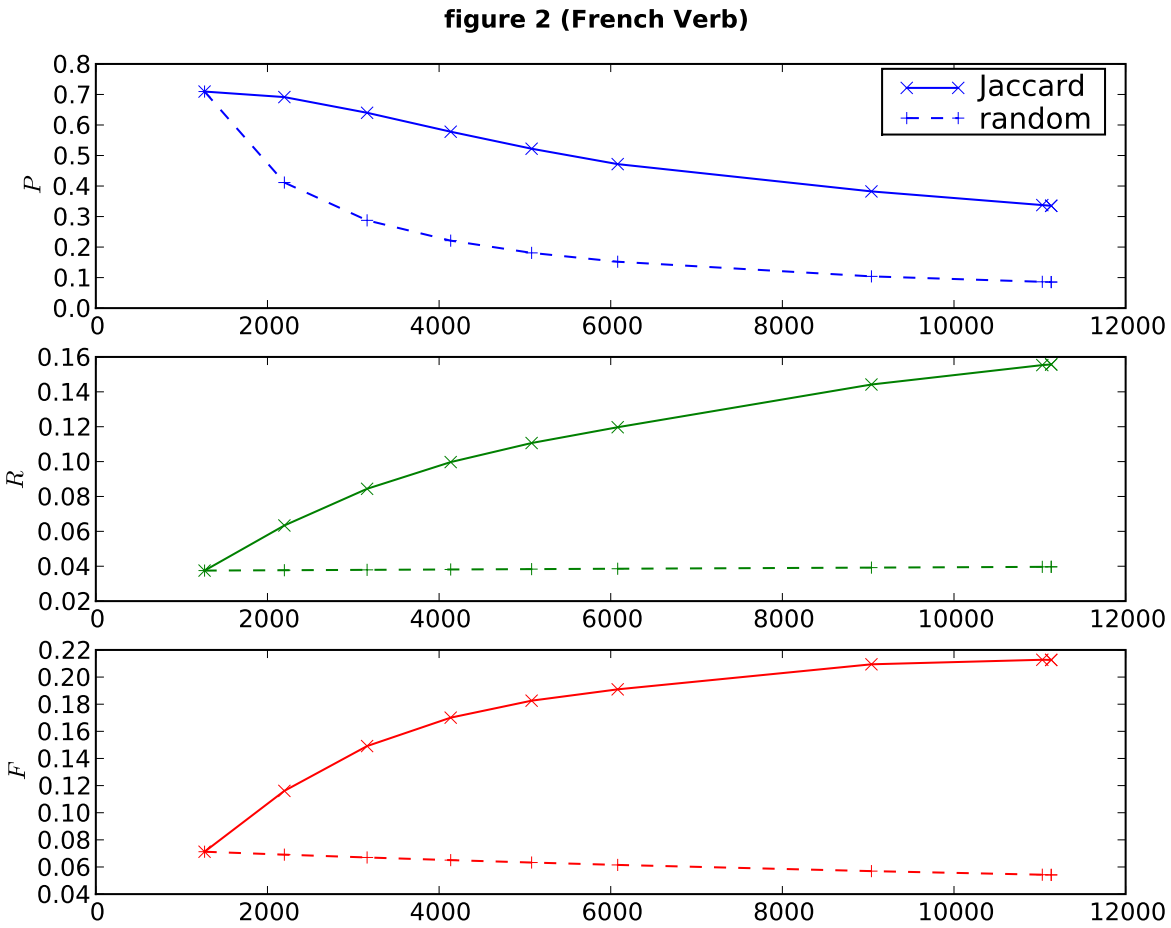
Results



Comments

- adding first 1000 edges (+55%) → loss of only 2% precision
- added links are not the same as with Prox method

Results



Comments

- adding first 1000 edges (+55%) → loss of only 2% precision
- added links are not the same as with Prox method

Idea

- use translations method to densify the graph
- then use the clusters' structure (Prox)

Conclusion

Hypothesis are confirmed

- many missing links should be added among members of the same cluster
- words sharing many translations are likely to be synonymous

Conclusion

Hypothesis are confirmed

- many missing links should be added among members of the same cluster
- words sharing many translations are likely to be synonymous

Our methods work...

- but there is room for improvement
- $\longrightarrow$ combine both methods should give better results

Conclusion

Hypothesis are confirmed

- many missing links should be added among members of the same cluster
- words sharing many translations are likely to be synonymous

Our methods work...

- but there is room for improvement
- $\longrightarrow$ combine both methods should give better results

Direct application

- support for collaborative editing
- $\longrightarrow$ module to be included in Wiktionary's framework?
- a list of synonyms, ordered by relevancy may be provided to the contributor

Future work

Diachronic study

- study how wiktionaries evolve
- $\longrightarrow$ forsee contributors' NLP needs
- eg. when to apply the methods presented here

Future work

Diachronic study

- study how wiktionaries evolve
- $\longrightarrow$ forsee contributors' NLP needs
- eg. when to apply the methods presented here

Invariants and variability

- study of the (in)varibility of semantic pairings (Wiktionary as a multilingual synonymy networks)
- eg. *house/family*, *child/fruit*, *feel/know*

Thank you!

Questions?