

# A Method for Grouping Synonyms

Ingrid Falk<sup>1</sup>, Claire Gardent<sup>2</sup>, Evelyne Jacquy<sup>3</sup>, Fabienne Venant<sup>4</sup>

<sup>1</sup>INRIA / Université Nancy 2

<sup>2</sup>CNRS / INRIA Nancy Grand-Est, Nancy

<sup>3</sup>CNRS / ATILF, Nancy

<sup>4</sup>Université Nancy 2 / INRIA, Nancy Grand-Est, Nancy

eLexicography in the 21st century

# Outline

## 1 Introduction and motivation

Example

## 2 Approach

## 3 Resources

## 4 Method

Extracting indexes.

Similarity of two indexes.

Reference sample.

## 5 Results

Evaluation measures

Reflexive usage.

## 6 Conclusion

## 7 Outlook

# Outline

## Objective

Build a dictionary which assigns to each meaning of a word (given by a definition)

the group of synonyms of that word which correspond to this meaning.

A method to merge synonym dictionaries and a large coverage general purpose dictionary.

- ▶ 5 synonym dictionaries from the synonym base of the ATILF and
- ▶ the TLFi (Trésor de la Langue Française informatisé).

Result :

- ▶ A large coverage **synonym** dictionary where **groups of synonyms** of a word are associated with **definitions**.
- ▶ A **general purpose** dictionary assigning a **group of synonyms** to each meaning (**definition**) of a word.

# Desired results

Example : **achever** (finish, accomplish)

## Synonym dictionaries:

abattre, aboutir, accomplir, aiguïser, améliorer, anéantir, assommer, boucler, cesser, clore, clôturer, compléter, conclure, conduire, consommer, continuer, couronner, estoquer, expédier, exécuter, finir, parachever, parfaire, perfectionner, raser, ruiner, réaliser, réussir, se taire, terminer, tuer.

## TLFi definitions:

1. *Mettre la dernière main pour perfectionner* (Finalize to improve.)
2. *Porter un coup mortel à un animal déjà atteint physiquement; donner le coup de grâce.* (Give a mortal blow to an animal already physically damaged;)
3. *Mener à sa fin, compléter l'action de.* (Lead to an end, complete the action.)

# Desired results

Synonym groupings, attached to the meaning(s) given by the TLFi definitions:

## Synonym dictionaries

abattre, aboutir, accomplir, accomplir, aiguïser, améliorer, anéantir, assommer, boucler, cesser, clore, clôturer, compléter, compléter, conclure, conduire, consommer, continuer, couronner, estoquer, expédier, exécuter, finir, parachever, parfaire, perfectionner, raser, ruiner, réaliser, réussir, se taire, terminer, tuer.

## TLFi definitions:

1. Mettre la dernière main pour perfectionner. (Finalize to improve.)
2. Porter un coup mortel à un animal déjà atteint physiquement; donner le coup de grâce. (Give a mortal blow to an animal already achieved physically;)
3. Mener à sa fin, compléter l'action de. (Lead to an end, complete the action.)

# Outline

- 1 Introduction and motivation
  - Example
- 2 Approach
- 3 Resources
- 4 Method
  - Extracting indexes.
  - Similarity of two indexes.
  - Reference sample.
- 5 Results
  - Evaluation measures
  - Reflexive usage.
- 6 Conclusion
- 7 Outlook

# Approach

## Input

Synonym dictionaries: Synonym base from the ATILF

General purpose dictionary: TLFi

TLFi  $\rightsquigarrow$  definitions  $\approx$  meaning (sense)

## Output

A synonym dictionary which associates to each sense (definition) of a word the corresponding synonym group.

## Related work

- ▶ DicoSyn [Manguin et Al., 2004]
- ▶ WOLF [Sagot and Fišer, 2008]

### Differences

**DicoSyn:** no synonym groupings, no associated definitions.

**WOLF:** synonym groupings are obtained by translation to existing BALKANET synsets, links to WORDNET synsets.



# Outline

- 1 Introduction and motivation
  - Example
- 2 Approach
- 3 Resources
- 4 Method
  - Extracting indexes.
  - Similarity of two indexes.
  - Reference sample.
- 5 Results
  - Evaluation measures
  - Reflexive usage.
- 6 Conclusion
- 7 Outlook

## Resources used

5 of 7 synonym dictionaries from the ATILF.

Syn. Dic.	verbs	Syn/verb
<b>Bailly</b>	2600	1.
<b>Benac</b>	2656	1.5
<b>Du Chazaud</b>	3808	5.25
<b>Larousse</b>	3835	4.7
<b>Le Petit Robert</b>	5027	6.
<b>total</b>	5736	11.

but: no part of speech information, no definitions.

### TLFi

- ▶ 54 280 entries, 92 997 lemmas, 271 166 definitions.
- ▶ digitized, available online (<http://atilf.atilf.fr/>), XML format
- ▶ glosses have been lemmatised and POS-tagged.

but: few synonyms, information is not systematic.

# Outline

- 1 Introduction and motivation
  - Example
- 2 Approach
- 3 Resources
- 4 **Method**
  - Extracting indexes.
  - Similarity of two indexes.
  - Reference sample.
- 5 Results
  - Evaluation measures
  - Reflexive usage.
- 6 Conclusion
- 7 Outlook

# Basic procedure

Given:

- ▶ A verb  $V$ ,
- ▶ A set of definitions  $D_V^1 \dots D_V^n$  associated to  $V$  by the TLFi,
- ▶ The set of synonyms  $Syn_V^1 \dots Syn_V^m$  associated to  $V$  by the synonym base.

For each synonym  $Syn_V^k$ :

Which are the definitions  $D_V^i$  for which  $Syn_V^k$  is synonymous to  $V$ ?

## Basic procedure, ctd.

1. Extract TLFi definitions, convert to index.  
Index = list of content words, lemmatised.
2. Associate indexes to each definition and each synonym.
  - ▶ A synonym's index  $\rightsquigarrow \cup$  of indexes of each of its definitions.
3. Measure the similarity of two indexes.  
Which definition  $D_V^i$  of  $V$  is most similar to the definitions of  $Syn_V^k$ ?
4. Associate synonyms and definitions. Each synonym is associated to those definitions  $D_V^i$  of  $V$  which are most similar to the synonym's index.

# Extracting the index.

TLFi definitions  $\rightsquigarrow$  index

1. We use XML tags to extract TLFi *definitions*.
2. Definitions without a gloss, synonym- or domain indicators are discarded.
3. Index = list of lemmatised content words contained in the gloss, the synonym- and the domain indicators.

# Examples

## TLFi entry of *projeter* (to project)

→ **PROJETER**, verbe trans.

A. — [Corresp. à *projection*]

1. Jeter loin en avant, avec force. *Il s'agit de projeter l'adversaire sur le tapis (de terre, dans un heurt un peu violent, il [le cheval] part gaiement (Jeux et sports, 19*

- 1. ... elle s'en alla, après m'avoir dévisagé, jugé, pesé, analysé de ce r
  - **eût projeté** sur les gens un de ces liquides épais dont se servent les pieu
- MAUPASS., *Contes et nouv.*, t.2, Inconnue, 1885, p.1000.

— *Au fig.* Pousser, inciter quelqu'un. V. *jaculation* ex. de Huysmans.

— *Empl. pronom.* Se lancer. *Il s'est projeté dans mes bras!* (CÉLINE, *Mort à crédit*,

2. Envoyer des rayons lumineux; représenter une image sur une surface. *Je vis m'élançai* (PROUST, *Swann*, 1913, p.35). *Jean Farou (...)* projetait sur les murs une

— **CIN., AUDIOVISUEL.** Passer dans un projecteur (v. ce mot B). *Proje cinématographiques de cellules agrandies de telle sorte que leur taille soit établissements furent obligés d'échanger leurs programmes et, après quelque tem*

• *Projeter quelque lumière* (au fig.). **Éclaircir.** *Synon. jeter quelque lumière* (v.

## Example: the verb *projeter*

Extracted definitions and their indexes:

**Definition :** Jeter loin en avant avec force. (To throw far ahead and with strength.)

**Index :** ⟨ jeter, loin, avant, force ⟩

**Definition :** *CIN. AUDIOVISUEL.* Passer dans un projecteur. (To show on a projector.)

**Index :** ⟨ cinéma, audiovisuel, passer, projecteur ⟩

**Definition :** Éclaircir. Synon. jeter quelque lumière. (To lighten, To throw some light.)

**Index :** ⟨ éclaircir, jeter, lumière ⟩



# Measuring the similarity of two indexes.

Experiments with 2 types of similarity measures:

1. Overlap of lemmas (or lemma sequences) between the two indexes.
2. First and second order vector similarity measures based on semantic space models.

Total of 6 similarity measures.

# Which similarity measure works best?

Building a reference.

- ▶ Gold standard as reference: 27 verbs, their definitions and for each definition the associated synonyms.
- ▶ Build triples  $\langle \text{Verb}, \text{Definition}, \text{Synonym} \rangle$ .
- ▶ To a triple  $\langle V, D_V, \text{Syn}_V \rangle$  we associate
  - the value 1 if  $\text{Syn}_V$  is considered synonymous to  $V$  with the sense given by  $D_V$ ,
  - the value 0 else.

Example:

$\langle \text{achever}, \text{Mettre la dernière main pour perfectionner}, \text{accomplir} \rangle \rightsquigarrow 1$   
 $(\langle \text{perfect}, \text{Finalize to improve}, \text{accomplish} \rangle \rightsquigarrow 1)$

- ▶ Triple  $\leftrightarrow$  value associations done by **system** and **annotators**
- ▶ Result comparisons done on the basis of standard evaluation measures: precision, recall and F-measure.

## Building the reference sample.

### 3 scales: genericity, frequency, polysemy

**genericity** position in EUROWORDNET's hierarchy,

**frequency** extracted from frequency list from 10 years "Le Monde"  
(newspaper) parsed by Syntex (D. Bourigault).

**polysemy** number of definitions given by the TLFi.

### 3 positions on these scales:

high, medium, low.

⇒ 27 verbs.

### 4 annotators :

- ▶ Fabienne Venant (U. Nancy 2), Mick Grzesitchak (CNRS/ATILF), Christiane Jadelot (CNRS/ATILF), Aurélie Merlot (U. Nancy 2/ATILF).
- ▶ Supervised by Evelyne Jacquey (CNRS/ATILF).

## Inter-annotator agreement.

Syn. Dic.	Triples	Annotators	Agr.
Robert	2422	P1	80.39
		P2	85.59
		P3	75.55
		Q	59.537
Larousse	2573	P1	79.90
		P2	87.213
		P3	80.10
		Q	63.894
Du Chazaud	4893	P1	81.484
		P2	<b>87.96</b>
		P3	73.88
		Q	<b>57.319</b>
All dic.	7047	P1	81.481
		P2	87.072
		P3	76.5
		Q	63.374

# Outline

- 1 Introduction and motivation
  - Example
- 2 Approach
- 3 Resources
- 4 Method
  - Extracting indexes.
  - Similarity of two indexes.
  - Reference sample.
- 5 Results
  - Evaluation measures
  - Reflexive usage.
- 6 Conclusion
- 7 Outlook

# Evaluation measures

**Recall** Number of reference triples identified by the system / Total number of reference triples.

**Precision** Number of reference triples identified by the system / Total number of triples identified by the system.

**F-measure** Harmonic mean of recall and precision.

# Recall, Precision and F-measure

Sim. measure	R	P	F
baseline	0.497	0.315	0.385
Simple overlap	0.725	0.508	0.598
Extended overlap	0.723	0.508	0.597
Ext. overlap normalised	<b>0.729</b>	<b>0.513</b>	<b>0.602</b>
1 <sup>st</sup> order vectors	0.727	0.510	0.560
2 <sup>nd</sup> order vectors, w/o tf.idf	0.715	0.503	0.590
2 <sup>nd</sup> order, vectors with tf.idf	0.717	0.505	0.592

**Table:** Precision, recall and F-measure for the various similarity measures. Baseline: random association of synonyms and definitions.

# Reflexive usages

- ▶ The same TLFi entry may comprise several distinct usages, in particular reflexive and non-reflexive usages – *s'abandonner* vs. *abandonner* (*abandon oneself* vs. *abandon*).
- ▶ These distinct usages often have distinct synonyms:

## Example

*abandonner* abdiquer, abjurer, abolir, accorder, aliéner, capituler, cesser, concéder, confier, céder, disparaître, donner, délaisser, . . .

*s'abandonner* céder, faillir, fléchir, mollir, parler, s'adonner, s'effondrer, s'enfoncer, s'ouvrir, s'écouter, satisfaire, se donner, . . .

- ▶  $\rightsquigarrow$  separate, in a single TLFi entry, definitions of reflexive from non-reflexive usages.  
 $\rightsquigarrow$  only compare synonyms of (non-)reflexive usages with definitions of (non-)reflexive usages.



# Results with reflexive vs. non-reflexive distinction

Measure	W/o. refl. vs. non-refl. dist.			With refl. vs. non-refl. dist.		
	R	P	F	R	P	F
baseline	0.497	0.315	0.385	0.440	0.433	0.437
Over. 1	0.725	0.508	0.598	0.697	0.685	0.691
Over. 2	0.723	0.508	0.597	0.697	0.685	0.691
Over. 3	<b>0.729</b>	<b>0.513</b>	<b>0.602</b>	<b>0.711</b>	0.670	<b>0.706</b>
Vect. 1	0.727	0.510	0.560	0.704	<b>0.693</b>	0.698
Vect. 2	0.715	0.503	0.590	0.698	0.686	0.692
Vect. 3	0.717	0.505	0.592	0.701	0.689	0.695

**Table:** Precision, recall and F-measure for various similarity measures with and w/o reflexive vs. non-reflexive distinction.

A more fine-grained linguistic preprocessing improves the results.

# Outline

- 1 Introduction and motivation
  - Example
- 2 Approach
- 3 Resources
- 4 Method
  - Extracting indexes.
  - Similarity of two indexes.
  - Reference sample.
- 5 Results
  - Evaluation measures
  - Reflexive usage.
- 6 Conclusion
- 7 Outlook

# Conclusions

- ▶ We presented an unsupervised method to associate synonyms and definitions with reasonably high F-score (0.706 with an inter-annotator agreement of 0.87 as upper bound).
- ▶ The method does not introduce noise because it does not group together words that are not synonyms.
- ▶ The linguistic preprocessing has an important impact on the quality of the results.
- ▶ The method does not depend on the synonym dictionary, hence several dictionaries may be merged via attachments to TLFi definitions.
- ▶ It is applicable to other languages, but results will vary depending on general purpose dictionary used.

# Outline

- 1 Introduction and motivation
  - Example
- 2 Approach
- 3 Resources
- 4 Method
  - Extracting indexes.
  - Similarity of two indexes.
  - Reference sample.
- 5 Results
  - Evaluation measures
  - Reflexive usage.
- 6 Conclusion
- 7 Outlook

# Outlook

- ▶ Extending coverage by integrating further external resources (Wiktionary, Wikipedia, EUROWORDNET).
- ▶ Link obtained synonym subsets in a WORDNET like structure:
  1. use ontology fusion methods for merging with the French EUROWORDNET or WOLF (Sagot and Fišer 2008).
  2. combine with a translation approach to associate <verb, definition> pairs to WORDNET synsets.

Thank you!

# Bibliography

B. Sagot and D. Fišer.

*Building a Free French WordNet from Multilingual Resources.*

Proc. of Ontolex, 2008.

Jean-Luc Manguin, Jacques François, Rembert Eufe, Ludwig Fesenmeier, Corinne Ozouf and Morgane Sénéchal.

*Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux.*

Les Cahiers du CRISCO, 2004.