Collecting terms and variants of terms fitting in a multilingual framework

Béatrice DAILLE

LS2N, University of Nantes, France

July 21st, 2017

TermVar

Workshop on Terminological Variation

LABORATOIRE DES SCIENCES DU NUMÉRIQUE DE NANTES

UNIVERSITÉ DE NANTES

Outline

Outline

Outline

Outline

Outline

Outline

Introduction

Automatic detection of term variant

Introduction

Automatic detection of term variant

- Symbolic approaches
- Empirical approaches

Introduction

Automatic detection of term variant

- Symbolic approaches
- Empirical approaches

Issues

- Variants have many forms
- Each form has very few occurrences

Many methods are required.

Assumption

Languages

French, English, German, Spanish and Russian.

these languages are subject to the same variation processes of variation and can be formally described.

The definition of variant

Definition

[Daille et al., 1996:201]
A variant of a term is an utterance which is semantically and conceptually related to an original term.

Terms

- simple or complex terms
- morphological compounds: native or neoclassical
- syntagmatic compounds

Organisation of variants

- **Denominative variants**: to respond to the properties of transparency and of minimality of the denominative core of the term;
- **Conceptual variants**: to anchor the term in the system of knowledge instantiated in the text;
- **Linguistic variants**: to link the term into the language system only.

Can be paired with Freixa's categories of variants [Freixa, 2006]

Denominative variants

Definition

Denominative variants reflect a synonymy relation.

exact synonyms or approximate synonyms
lexicalised forms
(MED) En: *histamine flare test* → *histamine test* [Collet, 1997]

paraphrases
(AGR) Fr: *protéine végétale* 'plant protein' → *protéine d'origine végétale* 'protein from plant'

Conceptual variants

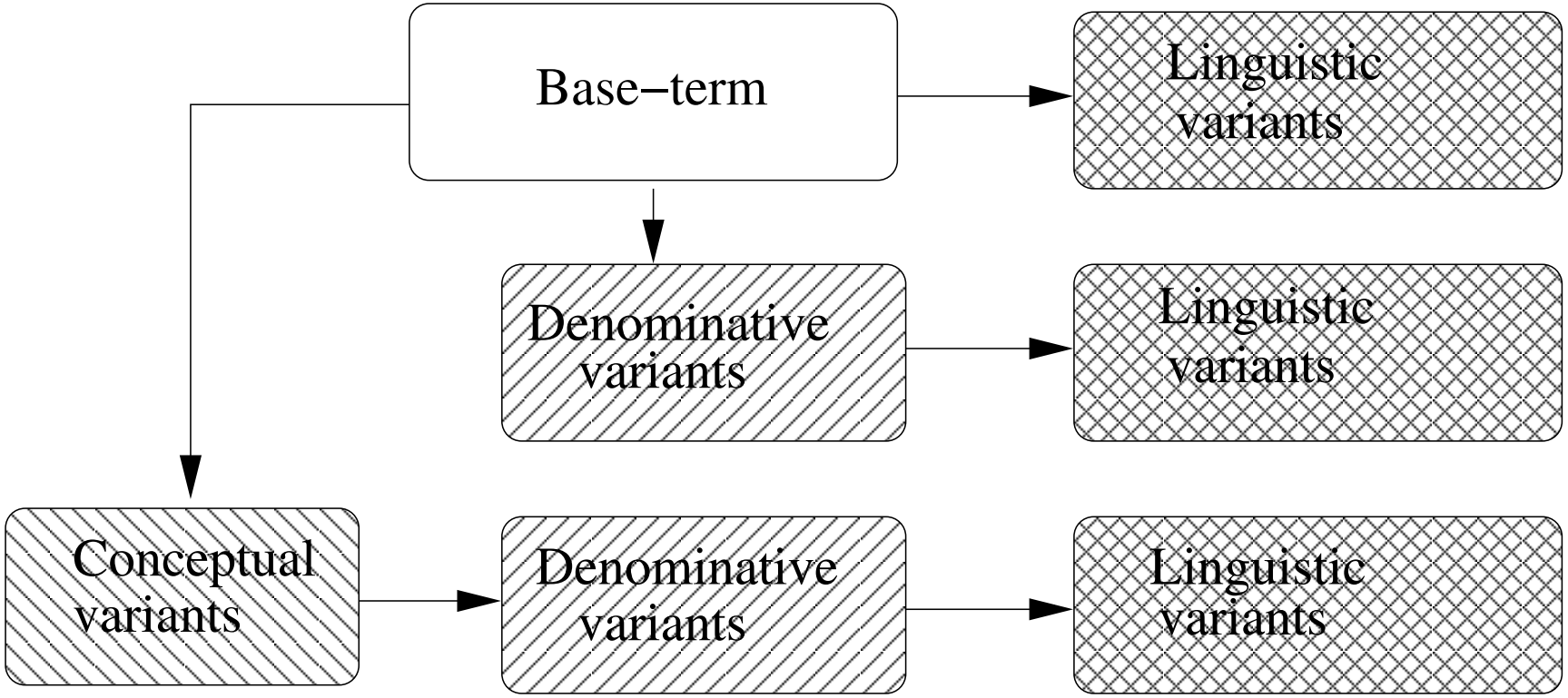
Definition

Conceptual variants reflect a conceptual or a semantic relation.

antonymy, taxonomy, meronymy and complex relations
lexicalisable forms
(MED) En: *blood cell* → *blood cell line*

lexicalised forms
(MED) En: *blood cell* → *cell* [Kister, 2000]

Relationships between categories of variants



Property of variants

1. a variant always involves at least one term;
2. a variant is obtained by applying at least one linguistic operation;
3. a term can produce several variants;
4. the number or utterances of the term in a text is slightly superior to the number of utterances of the variant.

Denominative variants

1. **synonymic substitution (lexical content)**
2. competing patterns
3. simplification (minimality criterion)
4. exemplification (transparency criterion)

Synonymic substitution (1)

Morpheme

several affixes in competition
(decision-making process) Fr: *-eur/-aire* for nouns
décid<u>eur</u> ↔ *décisionn<u>aire</u>* 'decision maker'
(decision-making process) Fr: *-el/-aire* for adjectives
processus décisionn<u>el</u> ↔ *processus décisionn<u>aire</u>* 'decision-making process'

Functional word

several prepositions in competion for N P N
(AGR) Fr: *chromatographie <u>en</u> colonne* ↔ *chromatographie <u>sur</u> colonne*
'column chromatography'

Synonymic lexical substitution (2)

Morphological compounds

several roots in competition
(Nervous system) Fr: *neurologie*
(Epidermis and dermis) Fr: *névrodermite*

several lexemes in competition
(EOL) De: *Rotorblattprofil* 'blade profile' $\leftrightarrow$ *Flügelprofile* 'wing profile'

Syntagmatic compounds

several lexemes in competition
(MED) Fr: *parenthèse thérapeutique* 'therapeutic range' $\leftrightarrow$ *fenêtre thérapeutique* 'therapeutic window'

Conceptual variants

1. **Expansion**
2. Anaphorical reduction

Expansion (1)

Derivation

(EOL) Fr: *éolien*/A 'eolian' → *proéolien*/A 'pro-eolian'

Predication

(DIA) En: *sentinel node* → *sentinel node biopsy*

Expansion: Modification (2)

Juxtaposition: initial position

$N_2+N_1 \rightarrow N+N_2+N_1$
(EOL) De: *Windpark* 'wind farm' → *Meerwindpark* 'marine wind farm'

$N_2\ N_1 \rightarrow A\ N_2\ N_1$
(SAT) En: *telecommunication satellite* → *geostationary telecommunication satellite*

Juxtaposition : post-position

$N_1\ A_2 \rightarrow N_1\ A_2\ A$
(SAT) Fr: *station terrienne* 'earth station' → *station terrienne brouilleuse* 'interfering earth station'

$N_1\ A_2 \rightarrow N_1\ A_2\ P\ A\ N$
(Transportation) Fr: *transport terrestre* 'ground transportation' → *transport terrestre à grande vitesse* 'high-speed ground transportation'

Expansion: Modification (3)

Insertion

$N_1\ A_2 \rightarrow N_1\ A/E{+}A_2$
(EOL) Es: *acoplamiento dinámico* → *acoplamiento* <u>*aero*</u>*dinámico*

$N_1\ A_1 \rightarrow N_1\ A\ A_1$
(EOL) Fr: *parc marin* 'marine park' → *parc* <u>*naturel*</u> *marin* 'natural marine park'

$N_1\ A_2\ P_3\ N_4 \rightarrow N_1\ A_2\ \underline{A}\ P_3\ N_4$
(SAT) Fr: *service de radiodiffusion par satellite* 'fixed-satellite service' → *services* <u>*communautaires*</u> *de radiodiffusion par satellite* 'domestic fixed-satellite service'

Syntagmatic analysis

Structural rules

Term: $X_1\ X_2 \rightarrow$ Variant: $X_1\ X\ X_2$

Linguistic preprocessing

- part-of-speech tagging
- lemmatisation
- splitting

Grammar of variants

	Term	Conceptual variant	Example
Insertion			
M	N/E_1+N_2	$N/E_1+N\ N_2$	bioengineering ↔ bioresource engineering
S	$A_1\ N_1$	$A_1\ A\ A\ N_1$	predominant acid ↔ predominant volatile fatty acid
Expansion			
M	N/N_1+N_2	$N/N+N_1+N_2$	Windpark ↔ Meerwindpark
S	$N_2\ N_1$	$A\ N_2\ N_1$	amino acid ↔ bacterial amino acid

Features of the grammar of variants

		De	En	Es	Fr	Ru
CP	M	3	3	2	2	4
	S	2	5	4	5	1
AR	**M**	**6**	**5**	**2**	**4**	**2**
	S	**8**	**17**	**17**	**13**	**5**
CT	M	0	2	0	2	0
	S	11	9	15	11	4
Total	M	9	10	4	8	6
	S	21	31	36	29	10
Nb rules		30	41	40	37	16

Examples

Corpus FoodTech En (3 millions tokens built from ISTEX Database)

Rule	Nb	Example
M-I-AN-N|A|R	3867	high-pressure $\Longrightarrow$ high hydrostatic pressure
M-I-EN-N|A	2821	biotechnological process $\Longrightarrow$ bioprocess
M-I-NN-N	643	foodservice $\Longrightarrow$ food information service
M-I2-(A|N)N-E	197	egg yolk protein $\Longrightarrow$ egg yolk lipoprotein

Distributional analysis

two words are in a semantic relation if they share the same lexical contexts
Many studies based on distributional paradigm
[Hindle, 1990, Grefenstette, 1994, Lin, 1998, Hagiwara, 2008, Ferret, 2010]

Context modelling

- size of the context
- contextual items
- association measures

Comparison of contexts

sililarity measures

Denominative variants and distributional analysis

A clear distinction between synonyms and other semantically related words is not obvious [Lin et al., 2003, van der Plas and Tiedemann, 2006]
Semantic proximity:

- classical lexical relationships : synonymy, antonymy, hyperonymy, co-hyponymy, etc.
- non classical semantic relationships such as action/agent

Example of results of distributional approach for Fr, En and Es on [EOL][1]

Rank	Fr: *ferme*	En: *construction*	Es: *torre*
1	*éolien*	*operation*	*aerogeneradores*
2	*horn*	*project*	*aerogenerador*
3	*aller*	*impact*	*pala*
4	*fréquence*	*road*	*rotor*
5	*puissance*	*require*	*estar*
6	*rev*	*aera*	***turbina***
7	*réseau*	*activity*	*altura*
8	***parc***	*plan*	*celosia*
...			
62		***tower***	

[1] corpus available at
www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html

Compositional method

assumption of semantic composition to generate the synonyms of a complex term

First work [Hamon and Nazarenko, 2001]

Synonymy of complex terms is compositional

$$R_1 \ : \ T_1 = T_2 \wedge syn(E_1, E_2) \supset syn(CCT_1, CCT_2)$$
$$R_2 \ : \ E_1 = E_2 \wedge syn(T_1, T_2) \supset syn(CCT_1, CCT_2)$$
$$R_3 \ : \ syn(T_1, T_2) \wedge syn(E_1, E_2) \supset syn(CCT_1, CCT_2)$$

with $CCT_1 = (T_1, E_1)$; $CCT_2 = (T_2, E_2)$ are complex terms and $syn(CCT_1, CCT_2)$ a synonym relation between the candidate terms CT_1 and CT_2

–> R_1 means that the heads are identical and the expansions are synonymous

Denominative variants of MWT

Assumption

A synonym of a MWT can be obtained by extracting synonyms and/or semantically related words of each component individually thanks to distributional methods

Synomymic MWT

Example

- **Synomyms**: (EOL) En: *energy output/energy production* where *ouput* and *production* are synonyms (Termium)
- **Hyperonyms**: (EOL) Es: *implantación de las máquinas/implantación de aerogeneradores* where *máquina* is an hyperonym of *aerogenerador* (GDT)
- **Undefined**: (EOL) Fr: *arbre lent/arbre primaire* with no relation between *lent* and *primaire* (Terminalf)

Semi-Compositional Method

- Distributional Method
  - Provides synonyms of each lexical element of the MWT
- Extend the rules from [Hamon and Nazarenko, 2001] to the semantic level and generalize them to MWT of any length
  - $R_1^G : T_1 = T_2 \wedge sem(E_1, E_2) \supset sem(CCT_1, CCT_2)$
  - $R_2^G : E_1 = E_2 \wedge sem(T_1, T_2) \supset sem(CCT_1, CCT_2)$

Examples of denominative variant discovered in [EOL] and [CAN]

French

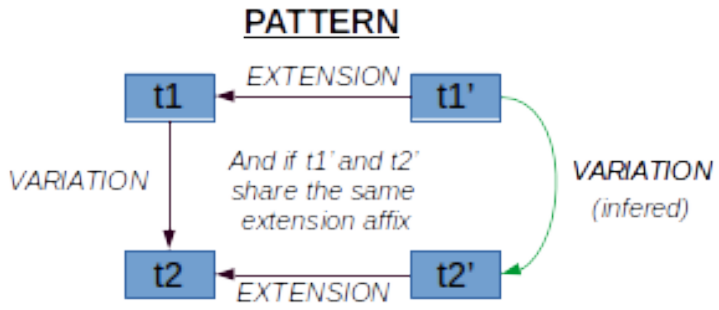
N *A* → N *A'*	vents modérés	vents moyens
A N → *A'* N	générateur synchrone	alternateur synchrone
N P *N* → N P *N'*	coût de l'électricité	coût de l'énergie
N P N → *N'* P N	prix de l'électricité	coût de l'électricité

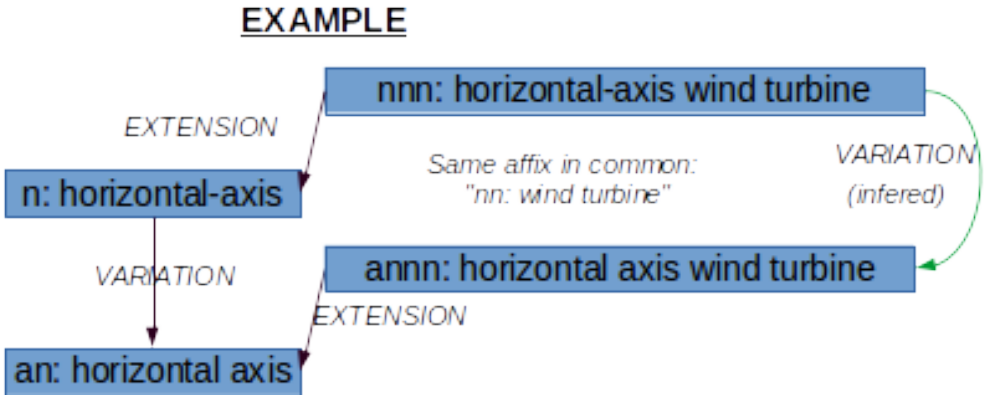
English

A *N* → A *N'*	initial surgery	initial operation
A N → *A'* N	dynamic stall	static stall
N P *N* → N P *N'*	type of surgery	type of operation
N P N → *N'* P N	center of the blade	midpoint of the blade
N *N* → N *N'*	mastectomy swimsuit	mastectomy swimwear
N N → *N'* N	flow field	exchange field

Infered variants

Patterns in sub-graphs





Infered variants (2)

Mixing distributional and syntagmatic analysis

Morphological analysis

chemotherapy → hormone therapy
individual chemotherapy drugs → individual hormone therapy drugs

Syntagmatic analysis

(EOL) Fr: générateur synchrone → alternateur synchrone
(EOL) Fr: générateur synchrone à rotor → alternateur synchrone à rotor

TermSuite

https://github.com/termsuite
Documentation
http://termsuite.github.io/

TermSuite Getting Started Documentation Publications About

Documentation TerminologyExtractorCLI

- Getting Started
- Command Line API
  - PreprocessorCLI
  - TerminologyExtractorCLI
  - AlignerCLI
- Java API
- Graphical User Interface
- Theory and Architecture
- Resources
- Links

- TerminologyExtractorCLI
  - Usage
  - Description
  - Mandatory options
    - `--from-text-corpus`, `--from-prepared-corpus`
    - `--tsv`, `--tbx`, `--json`
  - Other options
    - `--capped-size` INT
    - `--context-assoc-rate` INT or FLOAT
    - `--context-coocc-th` INT or FLOAT
    - `--context-scope` INT
    - `--contextualize` *(no arg)*
    - `--disable-derivative-splitting` *(no arg)*
    - `--disable-gathering` *(no arg)*
    - `--disable-merging` *(no arg)*
    - `--disable-morphology` *(no arg)*
    - `--disable-native-splitting` *(no arg)*
    - `--disable-post-processing` *(no arg)*
    - `--disable-prefix-splitting` *(no arg)*
    - `--enable-semantic-gathering` *(no arg)*
    - `--encoding`, `-e` ENC
    - `--from-prepared-corpus` DIR

Output of TermSuite with TSV format

Extract from [EOL]

	type	pilot	freq	spec	semScore	isDico	isDistrib
1	T	rotor	848	4,82			
2	T	wind turbine	1855	4,56			
2	V[h]+	wind power-plant	2	1,90	0,97	0	1
2	V[h]+	wind channel	2	2,20	0,97	0	1
2	V[h]+	Wind turbines-a	2	2,20	0,97	0	1
2	V[h]+	wind farm	488	3,20	0,89	0	1
2	V[s]	wind turbine rotor	31	3,38			
2	V[s]+	vertical-axis wind turbine	6	2,37			
2	V[s]	WIND TURBINE APPLICATIONS	86	3,83			
2	V[s]	wind turbine blades	48	3,57			
2	V[h]+	Enfield-Andreau turbine	3	2,37	0,54	0	1
2	V[s]+	wind turbine concepts	37	3,46			
2	V[s]+	wind turbine generator	27	3,02			
2	V[s]+	Domestic Wind Turbines	29	3,35			
2	V[s]+	small wind turbines	33	3,41			
2	V[s]	MW wind turbine	10	2,89			
4	T	wind power	278	4,34			
4	V[s]	wind turbine power	10	2,89			
4	V[s]+	wind power stations	24	3,27			
4	V[s]	Wind Power Project	19	3,17			
4	V[s]	wind power development	14	3,04			
4	V[s]+	wind power generation	11	2,63			
4	V[s]+	wind power capacity	6	2,67			
4	V[s]+	wind power penetration	4	2,50			
4	V[s]	Wind Power Installation	2	2,20			
5	T	airfoil	236	4,26			
6	T	voltage	214	4,22			
25	T	TURBINE SOUND	80	3,80			
25	V[s]+	WIND TURBINE SOUND	71	3,74			
25	V[h]+	turbine noise	52	3,61	0,65	0	1

TermSuite users

TermSuite was intiated during the research project TTC (*/*FP7/2007-2013*/*) under Grant Agreement no. 248005 from 2010 to 2012 and has been supported by ISTEX, French Excellence Initiative of Scientific and Technical Information, from 2015 to 2017

INIST: book indexation
Meteojob: alignment cv and job advertisements
Health oriented Innovation Lab (US): clinical NLP
University of the Basque country: specific bilingual dictionaries for MT
Prometil: detect errors in your requirements

Conclusion and Future works

Discovering terms and variants in a multilingual framework using both syntagmatic and distributional analysis

Conclusion and Future works

Discovering terms and variants in a multilingual framework using both syntagmatic and distributional analysis

Perspectives

- Verbal variants
- Distributional analysis

Conclusion and Future works

Discovering terms and variants in a multilingual framework using both syntagmatic and distributional analysis

Perspectives

- Verbal variants
- Distributional analysis
  - paraphrases

Conclusion and Future works

Discovering terms and variants in a multilingual framework using both syntagmatic and distributional analysis

Perspectives

- Verbal variants
- Distributional analysis
  - paraphrases
  - word embeddings

Conclusion and Future works

Discovering terms and variants in a multilingual framework using both syntagmatic and distributional analysis

Perspectives

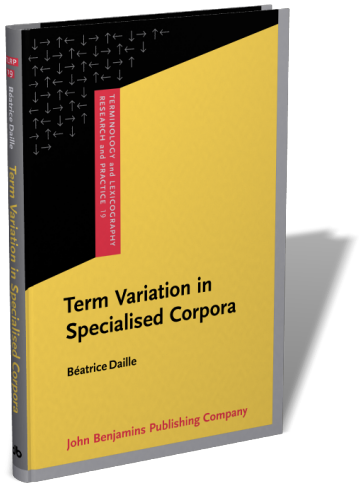
- Verbal variants
- Distributional analysis
  - paraphrases
  - word embeddings
- Multilingual variants

References

Damien Cram and Béatrice Daille. *Terminology Extraction with Term Variant Detection*. Proceedings of ACL-2016 System Demonstrations, 2016.

Term variation in specialised corpora characterisation, automatic discovery and applications.
B. Daille

Collection
Terminology and Lexicography Research and Practice
John Benjamins (printing)



Thank you for your attention

Thank you for your attention

Collet, T. (1997).
La réduction des unités terminologiques complexes de type syntagmatique.
Meta : journal des traducteurs / Meta: Translators' Journal, 42(1):193–206.

Daille, B., Habert, B., Jacquemin, C., and Royauté, J. (1996).
Empirical observation of term variations and principles for their description.
Terminology, 3(2):197–257.

Ferret, O. (2010).
Testing semantic similarity measures for extracting synonyms from a corpus.
In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *LREC*.

Freixa, J. (2006).
Causes of denominative variation - a typology proposal.
Terminology, 12(1):51–77.

Grefenstette, G. (1994).
Explorations in Automatic Thesaurus Discovery.
Kluwer Academic Publisher, Boston, MA, USA.

Hagiwara, M. (2008).
A supervised learning approach to automatic synonym identification based on distributional features.
In *Proceedings of the ACL-08: HLT Student Research Workshop*, pages 1–6, Columbus, Ohio. Association for Computational Linguistics.

Hamon, T. and Nazarenko, A. (2001).
Detection of synonymy links between terms: experiment and results.
In *Recent Advances in Computational Terminology*, pages 185–208. John Benjamins.

Hindle, D. (1990).
Noun classification from predicate-argument structures.
In *ACL*, pages 268–275.

Kister, L. (2000).

Is it possible to predetermine a referent included in a french n de n structure?
In Botley, S. and McEnnery, A., editors, *Corpus-Based and Computation Approches to Discourse Anaphora*. John Benjamins, Amsterdam.

Lin, D. (1998).
Automatic retrieval and clustering of similar words.
In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lin, D., Zhao, S., Qin, L., and Zhou, M. (2003).
Identifying synonyms among distributionally similar words.
In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, Acapulco, Mexico.

van der Plas, L. and Tiedemann, J. (2006).
Finding synonyms using automatic word alignment and measures of distributional similarity.

In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics ACL'06*, Sydney, Australia.