Language Acquisition of Multiword Expressions

from language technology to language learners

Aline Villavicencio

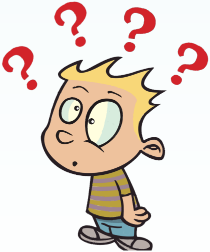
Institute of Informatics

Federal University of Rio Grande do Sul, Brazil

Saarbrücken, January, 2013

Multiword expressions (MWE)

1. What are they?
2. Why are they important?
3. What happens when we ignore them?



Multiword expressions (MWE)

Jumping the Shark

1. The moment when an established TV show changes in a significant manner in an attempt to stay fresh.

Multiword expressions (MWE)

Jumping the Shark

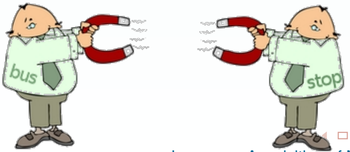
1. The moment when an established TV show changes in a significant manner in an attempt to stay fresh.

What are MWEs?

- *loan shark*
- *French kiss*
- *open mind*
- *vacuum cleaner*
- *voice mail*
- *high heel shoe*
- *make sense*
- *good morning*
- *take a shower*
- *upside down*
- . . .

- *es pan comido*
- *estiró la pata*
- *traer por la calle de la amargura*
- *dar gato por liebre*
- *alucinar en colores*
- *calcular a ojímetro*
- *dejar plantado*
- *meter la pata*
- . . .

- *quebrar um galho*
- *lavar roupa suja*
- *cara de pau*
- *amigo da onça*
- *aspirador de pó*
- *fazer sentido*
- *tomar banho*
- *dar-se conta*
- *nem te conto*
- *depois de amanhã*
- . . .



MWE: definition(s)

What is a word? What is a MWE? [Church, 2011]

- A unit whose exact meaning cannot be derived directly from the meaning of its parts [Choueka, 1988]
- Arbitrary and recurrent word combinations [Smadja, 1993]
- Idiosyncratic interpretations that cross word boundaries (or spaces) [Sag et al., 2002]

Multiword expression

A combination of words that must be treated as a unit at some level of linguistic processing.

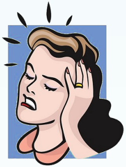
[Calzolari et al., 2002]

Characteristics I

1. **Arbitrariness and Institutionalisation**: *salt and pepper*, *?pepper and salt* [Smadja, 1993]
2. **Frequency**: 50% to 70% of the lexicon [Jackendoff, 1997, Krieger and Finatto, 2004, Ramisch, 2009]
3. **Limited lexical, syntactic and semantic variability**:*kick the bucket/?pail/?container* [Sag et al., 2002]

Why are MWEs important for NLP?

Because they are. . .



- Frequent [Sag et al., 2002]
- A marker of fluency
- Between lexicon and syntax [Calzolari et al., 2002]
- Hard to translate, parse, disambiguate, etc.
- An open problem in NLP [Schone and Jurafsky, 2001]

What happens if we ignore them?

We may get lost in translation:

From Greek to English

1. **Money laundering** represents between 2 and 5% ...
   - **The rinsing of dirty money** represents the 2 until 5%
2. as seen from the human **point of view**
   - as this is fixed by the human **optical corner**

What happens if we ignore them?

- MWEs are not as present in NLP applications as in languages
- Lexical resources construction is onerous

However

- Corpora are rich information sources
- MWE integration can improve the quality of NLP systems

Tasks [Anastasiou et al., 2009]

- **Acquisition**:

  [Silva and Lopes, 1999, Frantzi et al., 2000, Fazly et al., 2009, Seretan and Wehrli, 2009, Pecina, 2010, Kim and Baldwin, 2010]
- **Interpretation and disambiguation**:

  [Baldwin, 2006, Fazly et al., 2007, McCarthy et al., 2007, Nakov, 2008].
- **Representation**: [Laporte and Voyatzi, 2008, Grégoire, 2010, Graliński et al., 2010, Izumi et al., 2010, Schuler and Joshi, 2011]
- **Applications**:
  - Parsing: [Wehrli et al., 2010, Hogan et al., 2011]
  - IR: [Acosta et al., 2011, Xu et al., 2010]
  - WSD: [Finlayson and Kulkarni, 2011]
  - MT: [Ren et al., 2009, Pal et al., 2010, Carpuat and Diab, 2010]

Zoom on acquisition

1. Develop techniques for automatic acquisition of MWEs from corpora
2. Evaluate the usefulness of MWEs in NLP applications.
3. Investigate the application of MWE identification techniques for language acquisition studies.

Zoom on acquisition

1. Develop techniques for automatic acquisition of MWEs from corpora
2. Evaluate the usefulness of MWEs in NLP applications.
3. Investigate the application of MWE identification techniques for language acquisition studies.

Zoom on acquisition

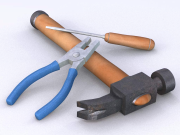
1. Develop techniques for automatic acquisition of MWEs from corpora
2. Evaluate the usefulness of MWEs in NLP applications.
3. Investigate the application of MWE identification techniques for language acquisition studies.

Outline

1. Multiword expressions (MWEs) in a Nutshell
2. A hard nut to crack
3. Lexicography
4. Machine Translation
5. VPCs in English Child Language
6. Conclusions and Future work

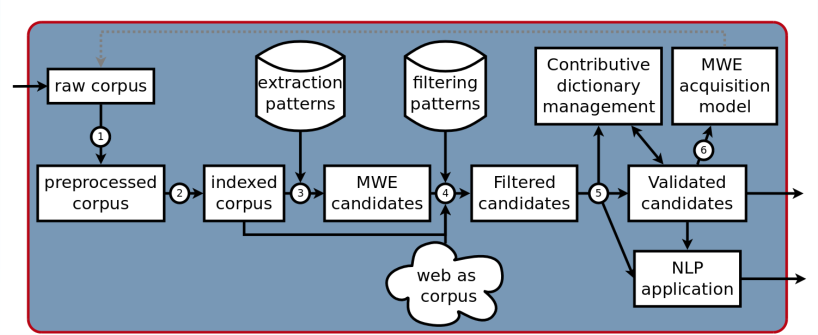
Tools for monolingual acquisition

- LocalMaxs – `hlt.di.fct.unl.pt/luis/multiwords/`
- Text::NSP – `search.cpan.org/dist/Text-NSP`
- UCS – `www.collocations.de/software.html`
- jMWE – `projects.csail.mit.edu/jmwe`
- Varro – `sourceforge.net/projects/varro/`
- Web services like Yahoo! terms
- Terminology extraction tools



A MWE processing framework

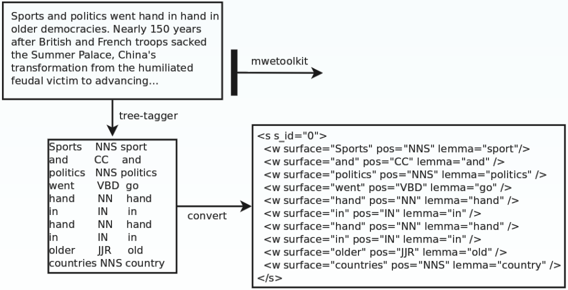
[Ramisch et al., 2010d, Ramisch et al., 2010b, Ramisch et al., 2012]



1. Preprocessing (external)

External tools for

1. Tokenisation, Lemmatisation, POS tagging, Dependency parsing



Sports and politics went hand in hand in older democracies. Nearly 150 years after British and French troops sacked the Summer Palace, China's transformation from the humiliated feudal victim to advancing...

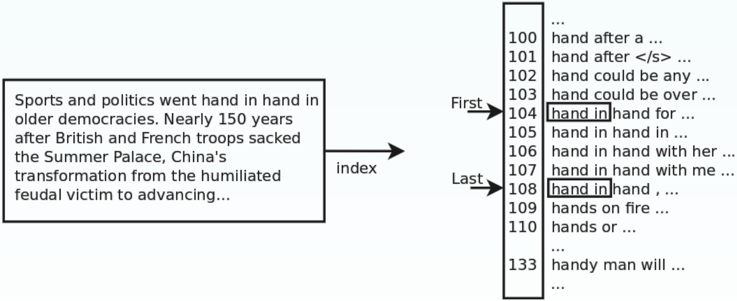
```
Sports    NNS politics
```

```
Sports    NNS sport
and       CC  and
politics  NNS politics
went      VBD go
hand      NN  hand
in        IN  in
hand      NN  hand
in        IN  in
older     JJR old
countries NNS country
```

```
<s s_id="0">
  <w surface="Sports" pos="NNS" lemma="sport"/>
  <w surface="and" pos="CC" lemma="and" />
  <w surface="politics" pos="NNS" lemma="politics" />
  <w surface="went" pos="VBD" lemma="go" />
  <w surface="hand" pos="NN" lemma="hand" />
  <w surface="in" pos="IN" lemma="in" />
  <w surface="hand" pos="NN" lemma="hand" />
  <w surface="in" pos="IN" lemma="in" />
  <w surface="older" pos="JJR" lemma="old" />
  <w surface="countries" pos="NNS" lemma="country" />
</s>
```

2. Corpus Indexing

- Suffix array

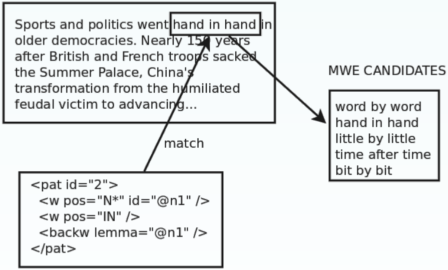


Sports and politics went hand in hand in older democracies. Nearly 150 years after British and French troops sacked the Summer Palace, China's transformation from the humiliated feudal victim to advancing...

		...
	100	hand after a ...
	101	hand after </s> ...
	102	hand could be any ...
	103	hand could be over ...
First	104	hand in hand for ...
	105	hand in hand in ...
	106	hand in hand with her ...
	107	hand in hand with me ...
Last	108	hand in hand , ...
	109	hands on fire ...
	110	hands or ...
		...
	133	handy man will ...
		...

3. Candidate extraction

- Linguistic Patterns



Sports and politics went hand in hand in older democracies. Nearly 150 years after British and French troops sacked the Summer Palace, China's transformation from the humiliated feudal victim to advancing...

```
<pat id="2">
  <w pos="N*" id="@n1" />
  <w pos="IN" />
  <backw lemma="@n1" />
</pat>
```

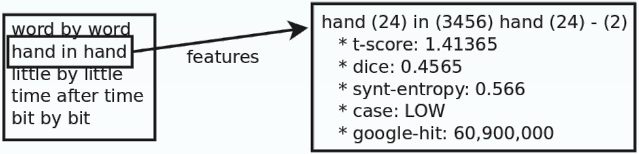
word by word
hand in hand
little by little
time after time
bit by bit

4. Candidate filtering

Features:

- Association measures, Variation entropy

[Ramisch et al., 2008]



word by word
hand in hand
little by little
time after time
bit by bit

features

hand (24) in (3456) hand (24) - (2)
* t-score: 1.41365
* dice: 0.4565
* synt-entropy: 0.566
* case: LOW
* google-hit: 60,900,000

Some association measures:

$$\text{t-score} = \frac{c(w_1^n) - E(w_1^n)}{\sqrt{c(w_1^n)}} \qquad \text{pmi} = \log_2 \frac{c(w_1^n)}{E(w_1^n)}$$

$$\text{dice} = \frac{n \times c(w_1^n)}{\sum_{i=1}^{n} c(w_i)} \qquad \text{ll} = \sum_{w_i w_j} \log \left[\frac{c(w_i w_j)}{E(w_i w_j)} \right]^{c(w_i w_j)}$$

5. Validation

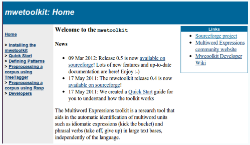
- Intrinsic using dictionaries, experts' or native speakers' judgements
- Extrinsic within NLP application

6. Machine Learning

- Export to WEKA machine learning toolkit
- Learn classifiers
- Apply to new data

The `mwetoolkit`

`mwetoolkit.sf.net`



mwetoolkit: Home

Home

> Installing the mwetoolkit
> Quick Start
> Defining Patterns
> Preprocessing a corpus using TreeTagger
> Preprocessing a corpus using Rasp
> Developers

Welcome to the `mwetoolkit`

News

- 09 Mar 2012: Release 0.5 is now available on sourceforge! Lots of new features and up-to-date documentation are here! Enjoy :-)
- 17 May 2011: The mwetoolkit release 0.4 is now available on sourceforge!
- 17 May 2011: We created a Quick Start guide for you to understand how the toolkit works

The Multiword Expressions toolkit is a research tool that aids in the automatic identification of multiword units such as idiomatic expressions (kick the bucket) and phrasal verbs (take off, give up) in large text bases, independently of the language.

Links

- Sourceforge project
- Multiword Expressions community website
- Mweoolkit Developer Wiki

- Target users: computational linguists
- Modular, customisable system
- Independent of language, n-gram length , adjacency, formalism, preprocessing tool

Outline

1. Multiword expressions (MWEs) in a Nutshell
2. A hard nut to crack
3. Lexicography
4. Machine Translation
5. VPCs in English Child Language
6. Conclusions and Future work

For creating lexical resources

- The `mwetoolkit` can be used for identifying and suggesting MWE entries

Creating MWE resources

- English MWE lexicon extension for parsing

  [Zhang et al., 2006, Villavicencio et al., 2007]
- Compositionality detection of English VPCs

  [Ramisch et al., 2008]
- Greek nominal expressions lexicon

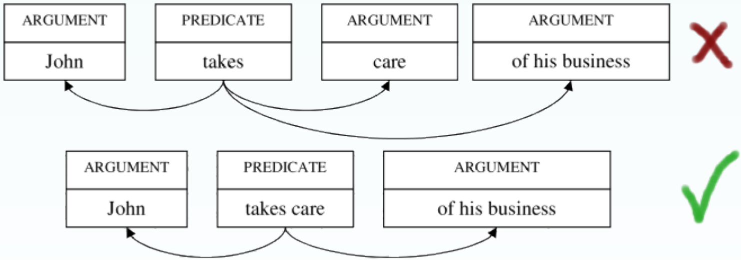
  [Linardaki et al., 2010]
- Portuguese Light Verb lexicon [Duran et al., 2011]

Portuguese Light Verb lexicon

[Duran et al., 2011]

Portuguese Light Verb lexicon

Light Verb + Noun: *take care, take shower, take walk, tomar cuidado, tomar banho, dar caminhada*



Problem: coverage of light verbs in lexical resources

Portuguese Light Verb lexicon

Corpus

PLN-BR-Full: 29M words, news, POS tagged

Patterns:

1. V + N + P: *abrir mão de* (*give up*, lit. *open hand of*)
2. V + P + N: *deixar de lado* (*ignore*, lit. *leave at side*)
3. V + DT + N + P: *virar as costas para* (*ignore*, lit. *turn the back to*)
4. V + DT + ADV: *dar o fora* (*get out*, lit. *give the out*)
5. V + ADV: *ir atrás* (*follow*, lit. *go behind*)
6. V + P + ADV: *dar para trás* (*give up*, lit. *give to back*)
7. V + ADJ: *dar duro* (*work hard*, lit. *give hard*)

Portuguese Light Verb lexicon I

pattern	acquired	analysed	− idiom.	+ idiom.
V + N + P	69,264	2,140	327	8
V + P + N	74,086	1,238	77	8
V + DT + N + P	178,956	3,187	131	4
V + DT + ADV	1,537	32	0	0
V + ADV	51,552	3,626	19	41
V + P + ADV	5,916	182	0	2
V + ADJ	25,703	2,140	145	11
Total	407,014	12,545	699	74

Portuguese Light Verb lexicon

Traditional (*take, make, do*), and more unusual (*provide*) light verbs

- *dar tratamento = tratar*

  give treatment= treat

- *dar medo = amedrontar*

  give fear = frighten

- *tornar responsável = responsabilizar*

  hold responsible = responsibilise

- *prestar atenção = atentar?*

  pay attention = attend?

Outline

1. Multiword expressions (MWEs) in a Nutshell
2. A hard nut to crack
3. Lexicography
4. Machine Translation
5. VPCs in English Child Language
6. Conclusions and Future work

MWEs and machine translation (MT)

- MWEs introduce cross-lingual asymmetries
- Pilot study of their impact on MT quality
- Introduction in MT systems $\Longrightarrow$ +quality

Source: English verb-particle constructions (VPCs) *(give up, take off)*

Target: Portuguese verbs *(desistir, decolar)*

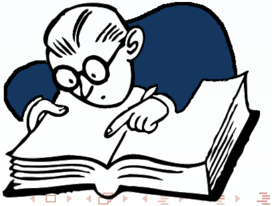
Verb-particle constructions (VPCs) in English

Semantic variability:

- *give back*
- *give up*
- *look up*

Syntactic variability:

- *She gave up*
- *She gave it up*
- *She gave up smoking*



Experimental context

- Baseline: Moses with WMT 2011 parameters on fragment of Europarl v6
- 660-sentences test set

Integration strategy 1/3: TOK

Concatenate verb and particle to treat them as a unit

*Europe will **give** it **up***

⇓

*Europe will **give_up** it*

Integration strategy 2/3: VPC?

Extra binary feature in translation model that flags VPCs

Source s	**Target** t	$p(t\|s)$	$lex(t\|s)$	$p(s\|t)$	$lex(s\|t)$	***VPC?***
a backward step .	de uma regressão .	1	0.0280	0.5	0.0025	***0***
a backward step .	uma regressão .	1	0.0280	0.5	0.0278	***0***
a backward step	de uma regressão	1	0.0287	0.5	0.0026	***0***
a backward step	uma regressão	1	0.0287	0.5	0.0288	***0***
. . .						
give up	desistimos	1	0.0187	0.5	0.0266	***1***
has ***given up*** the	desistiu da	1	0.0227	0.8	0.0654	***1***
has never ***given up***	nunca desistiu	1	0.0287	0.1	0.0022	***1***

Integration strategy 3/3: BILEX

Add bilingual lexicon of VPCs

Manual evaluation

- Scoring scheme:
  - 3 - good
  - 2 - acceptable
  - 1 - bad
  - 0 - untranslated

Translation quality

	% 3	% 2	% 1	% 0	Score
Baseline	59.88	9.58	***30.54***	0.00	***383***
TOK	47.31	6.59	17.37	***28.74***	288
VPC?	59.88	10.78	***29.34***	0.00	***385***
BILEX	64.07	8.38	***27.54***	0.00	***395***

3 - good, 2 - acceptable, 1 - bad, 0 - untranslated

Outline

1. Multiword expressions (MWEs) in a Nutshell
2. A hard nut to crack
3. Lexicography
4. Machine Translation
5. VPCs in English Child Language
6. Conclusions and Future work

VPCs in English Child Language

[Villavicencio et al., 2012a]

Why Verb-Particle Constructions (VPCs)?

- Profiling of VPCs in English and their usage in child-produced and child-directed sentences
- Ground work for computational models of VPC learning

Corpus

English CHILDES [MacWhinney, 1995]

- child-produced and child-directed speech
- annotated with POS-tags, parses, verb semantic classes and psycholinguistic information [Villavicencio et al., 2012b]

VPCs in CHILDES

Sentences	Children Set	Adults Set
Parsed	482,137	988,101
with VPCs	38,326	82,796
% with VPCs	7.95	8.38

Children's Age in months	VPC Sentences
0-24	2,799
24-48	26,152
48-72	8,038
72-96	1,337
>96	514

VPCs in CHILDES

Rank	Chidren VPC	Adult VPC	Child Rank
1	put on	come on	7
2	go in	put on	1
3	get out	go on	9
4	take off	get out	3
5	fall down	take off	4
6	put in	put in	6
7	come on	sit down	8
8	sit down	go in	2
9	go on	come out	10
10	come out	pick up	18

Outline

1. Multiword expressions (MWEs) in a Nutshell
2. A hard nut to crack
3. Lexicography
4. Machine Translation
5. VPCs in English Child Language
6. **Conclusions and Future work**

Summary

- Develop techniques for automatic acquisition of MWEs from corpora
- Evaluate the usefulness of MWEs in language technology applications.
- Investigate the application of MWE identification techniques for language acquisition studies.

Future work

- Clustering methods
- Further investigate use of entropy
- Explore cross lingual (a)symmetries
- Classification (interpretation and disambiguation)

Acknowledgements

- *This research is a collaboration between UFRGS (Brazil), U. of Grenoble (France), U. Saarland (Germany) and MIT (USA)*
- *It is in great part described in Carlos Ramisch's PhD thesis and most of the slides are his.*
- *It is partly funded by CNPq Projects 551964/2011-1, 202007/2010-3, 305256/2008-4 and 309569/2009-5 and CAPES/COFECUB 707/11*

Selected publications I

- Ramisch, C., Villavicencio, A., and Boitet, C. (2010d). Web-based and combined language models: a case study on noun compound identification. In Huang, C.-R. and Jurafsky, D., editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 1041–1049, Beijing, China. The Coling 2010 Organizing Committee
- Ramisch, C., Villavicencio, A., and Boitet, C. (2010b). Multiword expressions in the wild? the mwetoolkit comes in handy. In Liu, Y. and Liu, T., editors, *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China. The Coling 2010 Organizing Committee
- Ramisch, C., de Medeiros Caseli, H., Villavicencio, A., Machado, A., and Finatto, M. J. (2010a). A hybrid approach for multiword expression identification. In *Proc. of the 9th PROPOR (PROPOR 2010)*, volume 6001 of *LNCS (LNAI)*, pages 65–74, Porto Alegre, RS, Brazil. Springer
- Villavicencio, A., Ramisch, C., Machado, A., de Medeiros Caseli, H., and Finatto, M. J. (2010). Identificação de expressões multipalavra em domínios específicos. *Linguamática*, 2(1):15–33
- Ramisch, C., Villavicencio, A., and Boitet, C. (2010c). mwetoolkit: a framework for multiword expression identification. In *Proc. of the Seventh LREC (LREC 2010)*, pages 662–669, Malta. ELRA
- de Medeiros Caseli, H., Ramisch, C., das Graças Volpe Nunes, M., and Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. In [jou, 2010], pages 59–77
- Araujo, V. D., Ramisch, C., and Villavicencio, A. (2011). Fast and flexible MWE candidate generation with the mwetoolkit. In [Kordoni et al., 2011], pages 134–136
- Duran, M. S., Ramisch, C., Aluísio, S. M., and Villavicencio, A. (2011). Identifying and analyzing Brazilian Portuguese complex predicates. In [Kordoni et al., 2011], pages 74–82
- Duran, M. S. and Ramisch, C. (2011). How do you feel? investigating lexical-syntactic patterns in sentiment expression. In *Proceedings of Corpus Linguistics 2011: Discourse and Corpus Linguistics Conference*, Birmingham, UK

Selected publications II

- Mangeot, M. and Ramisch, C. (2012). A serious lexical game for building a Portuguese lexical-semantic network. In *Proceedings of the ACL 2012 3rd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources and their Applications to NLP*, Jeju, Republic of Korea. Association for Computational Linguistics
- Granada, R., Lopes, L., Ramisch, C., Trojahn, C., Vieira, R., and Villavicencio, A. (2012). A comparable corpus based on aligned multilingual ontologies. In *Proceedings of the ACL 2012 First Workshop on Multilingual Modeling (MM 2012)*, Jeju, Republic of Korea. Association for Computational Linguistics
- Ramisch, C. (2012). Une plate-forme générique et ouverte pour le traitement des expressions polylexicales. In Molina Mejia, J. M. and Schwab, D., editors, *Actes de 14e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2012)*, Grenoble, France
- Villavicencio, A., Idiart, M., Ramisch, C., Araujo, V. D., Yankama, B., and Berwick, R. (2012a). Get out but don't fall down: verb-particle constructions in child language. In Berwick, R., Korhonen, A., Poibeau, T., and Villavicencio, A., editors, *Proc. of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss*, pages 43–50, Avignon, France. ACL

Language Acquisition of Multiword Expressions

from language technology to language learners

Aline Villavicencio

Institute of Informatics

Federal University of Rio Grande do Sul, Brazil

Saarbrücken, January, 2013

References I

(2010).
Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing, 44(1-2).

Acosta, O., Villavicencio, A., and Moreira, V. (2011).
Identification and treatment of multiword expressions applied to information retrieval.
In [Kordoni et al., 2011], pages 101–109.

Anastasiou, D., Hashimoto, C., Nakov, P., and Kim, S. N., editors (2009).
Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009), Suntec, Singapore. ACL.

Araujo, V. D., Ramisch, C., and Villavicencio, A. (2011).
Fast and flexible MWE candidate generation with the mwetoolkit.
In [Kordoni et al., 2011], pages 134–136.

Baldwin, T. (2006).
Compositionality and multiword expressions: Six of one, half a dozen of the other?
In [Moirón et al., 2006], page 1.

References II

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., Macleod, C., and Zampolli, A. (2002).
Towards best practice for multiword expressions in computational lexicons.
In *Proc. of the Third LREC (LREC 2002)*, pages 1934–1940, Las Palmas, Canary Islands, Spain. ELRA.

Carpuat, M. and Diab, M. (2010).
Task-based evaluation of multiword expressions: a pilot study in statistical machine translation.
In *Proc. of HLT: The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 242–245, Los Angeles, California. ACL.

Choueka, Y. (1988).
Looking for needles in a haystack or locating interesting collocational expressions in large textual databases.
In *RIAO'88*, pages 609–624.

Church, K. (2011).
How many multiword expressions do people know?
In [Kordoni et al., 2011], pages 137–144.

References III

de Medeiros Caseli, H., Ramisch, C., das Graças Volpe Nunes, M., and Villavicencio, A. (2010).
Alignment-based extraction of multiword expressions.
In [jou, 2010], pages 59–77.

Duran, M. S. and Ramisch, C. (2011).
How do you feel? investigating lexical-syntactic patterns in sentiment expression.
In *Proceedings of Corpus Linguistics 2011: Discourse and Corpus Linguistics Conference*, Birmingham, UK.

Duran, M. S., Ramisch, C., Aluísio, S. M., and Villavicencio, A. (2011).
Identifying and analyzing Brazilian Portuguese complex predicates.
In [Kordoni et al., 2011], pages 74–82.

Eisner, J., editor (2007).
Proc. of the 2007 Joint Conference on EMNLP and Computational NLL (EMNLP-CoNLL 2007), Prague, Czech Republic. ACL.

Fazly, A., Cook, P., and Stevenson, S. (2009).
Unsupervised type and token identification of idiomatic expressions.
Comp. Ling., 35(1):61–103.

References IV

Fazly, A., Stevenson, S., and North, R. (2007).
Automatically learning semantic knowledge about multiword predicates.
Lang. Res. & Eval., 41(1):61–89.

Finlayson, M. and Kulkarni, N. (2011).
Detecting multi-word expressions improves word sense disambiguation.
In [Kordoni et al., 2011], pages 20–24.

Frantzi, K., Ananiadou, S., and Mima, H. (2000).
Automatic recognition of multiword terms: the C-value/NC-value method.
Int. J. on Digital Libraries, 3(2):115–130.

Graliński, F., Savary, A., Czerepowicka, M., and Makowiecki, F. (2010).
Computational lexicography of multi-word units: How efficient can it be?
In [Laporte et al., 2010], pages 1–9.

Granada, R., Lopes, L., Ramisch, C., Trojahn, C., Vieira, R., and Villavicencio, A. (2012).
A comparable corpus based on aligned multilingual ontologies.
In *Proceedings of the ACL 2012 First Workshop on Multilingual Modeling (MM 2012)*, Jeju, Republic of Korea. Association for Computational Linguistics.

References V

Grégoire, N. (2010).
DuELME: a Dutch electronic lexicon of multiword expressions.
In [jou, 2010], pages 23–39.

Grégoire, N., Evert, S., and Krenn, B., editors (2008).
Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008), Marrakech, Morocco.

Hogan, D., Foster, J., and van Genabith, J. (2011).
Decreasing lexical data sparsity in statistical syntactic parsing - experiments with named entities.
In [Kordoni et al., 2011], pages 14–19.

Izumi, T., Imamura, K., Kikui, G., and Sato, S. (2010).
Standardizing complex functional expressions in Japanese predicates: Applying theoretically-based paraphrasing rules.
In [Laporte et al., 2010], pages 63–71.

Jackendoff, R. (1997).
Twistin' the night away.
Language, 73:534–559.

References VI

Kim, S. N. and Baldwin, T. (2010).
How to pick out token instances of English verb-particle constructions.
In [jou, 2010], pages 97–113.

Kordoni, V., Ramisch, C., and Villavicencio, A., editors (2011).
Proc. of the ACL Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011), Portland, OR, USA. ACL.

Krieger, M. and Finatto, M. J. B. (2004).
Introdução à Terminologia: teoria & prática.
Editora Contexto, São Paulo, SP, Brazil.
223 p.

Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors (2010).
Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010), Beijing, China. ACL.

Laporte, É. and Voyatzi, S. (2008).
An electronic dictionary of French multiword adverbs.
In [Grégoire et al., 2008], pages 31–34.

References VII

Linardaki, E., Ramisch, C., Villavicencio, A., and Fotopoulou, A. (2010).
Towards the construction of language resources for Greek multiword expressions: Extraction and evaluation.
In Piperidis, S., Slavcheva, M., and Vertan, C., editors, *Proc. of the LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*, pages 31–40, Valetta, Malta. May.

MacWhinney, B. (1995).
The CHILDES project: tools for analyzing talk.
Hillsdale, NJ: Lawrence Erlbaum Associates, second edition.

Mangeot, M. and Ramisch, C. (2012).
A serious lexical game for building a Portuguese lexical-semantic network.
In *Proceedings of the ACL 2012 3rd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources and their Applications to NLP*, Jeju, Republic of Korea. Association for Computational Linguistics.

McCarthy, D., Venkatapathy, S., and Joshi, A. (2007).
Detecting compositionality of verb-object combinations using selectional preferences.
In [Eisner, 2007], pages 369–379.

References VIII

Moirón, B. V., Villavicencio, A., McCarthy, D., Evert, S., and Stevenson, S., editors (2006).
Proc. of the COLING/ACL Workshop on MWEs: Identifying and Exploiting Underlying Properties (MWE 2006), Sidney, Australia. ACL.

Nakov, P. (2008).
Paraphrasing verbs for noun compound interpretation.
In [Grégoire et al., 2008], pages 46–49.

Pal, S., Naskar, S. K., Pecina, P., Bandyopadhyay, S., and Way, A. (2010).
Handling named entities and compound verbs in phrase-based statistical machine translation.
In [Laporte et al., 2010], pages 45–53.

Pecina, P. (2010).
Lexical association measures and collocation extraction.
In [jou, 2010], pages 137–158.

References IX

Ramisch, C. (2009).
Multiword terminology extraction for domain-specific documents.
Master's thesis, École Nationale Supérieure d'Informatique et de Mathématiques Appliquées, Grenoble, France.
79 p.

Ramisch, C. (2012).
Une plate-forme générique et ouverte pour le traitement des expressions polylexicales.
In Molina Mejia, J. M. and Schwab, D., editors, *Actes de 14e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2012)*, Grenoble, France.

Ramisch, C., Araujo, V. D., and Villavicencio, A. (2012).
A broad evaluation of techniques for automatic acquisition of multiword expressions.
In *Proc. of the ACL 2012 SRW*, pages 1–6, Jeju, Republic of Korea. ACL.

References X

Ramisch, C., de Medeiros Caseli, H., Villavicencio, A., Machado, A., and Finatto, M. J. (2010a).
A hybrid approach for multiword expression identification.
In *Proc. of the 9th PROPOR (PROPOR 2010)*, volume 6001 of *LNCS (LNAI)*, pages 65–74, Porto Alegre, RS, Brazil. Springer.

Ramisch, C., Schreiner, P., Idiart, M., and Villavicencio, A. (2008).
An evaluation of methods for the extraction of multiword expressions.
In [Grégoire et al., 2008], pages 50–53.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010b).
Multiword expressions in the wild? the mwetoolkit comes in handy.
In Liu, Y. and Liu, T., editors, *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China. The Coling 2010 Organizing Committee.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010c).
mwetoolkit: a framework for multiword expression identification.
In *Proc. of the Seventh LREC (LREC 2010)*, pages 662–669, Malta. ELRA.

References XI

Ramisch, C., Villavicencio, A., and Boitet, C. (2010d).
Web-based and combined language models: a case study on noun compound identification.
In Huang, C.-R. and Jurafsky, D., editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 1041–1049, Beijing, China. The Coling 2010 Organizing Committee.

Ren, Z., Lü, Y., Cao, J., Liu, Q., and Huang, Y. (2009).
Improving statistical machine translation using domain bilingual multiword expressions.
In [Anastasiou et al., 2009], pages 47–54.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002).
Multiword expressions: A pain in the neck for NLP.
In *Proc. of the 3rd CICLing (CICLing-2002)*, volume 2276/2010 of *LNCS*, pages 1–15, Mexico City, Mexico. Springer.

Schone, P. and Jurafsky, D. (2001).
Is knowledge-free induction of multiword unit dictionary headwords a solved problem?
In Lee, L. and Harman, D., editors, *Proc. of the 2001 EMNLP (EMNLP 2001)*, pages 100–108, Pittsburgh, PA USA. ACL.

References XII

Schuler, W. and Joshi, A. (2011).
Tree-rewriting models of multi-word expressions.
In [Kordoni et al., 2011], pages 25–30.

Seretan, V. and Wehrli, E. (2009).
Multilingual collocation extraction with a syntactic parser.
Lang. Res. & Eval. Special Issue on Multilingual Language Resources and Interoperability, 43(1):71–85.

Silva, J. and Lopes, G. (1999).
A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora.
In *Proceedings of the Sixth Meeting on Mathematics of Language (MOL6)*, pages 369–381, Orlando, FL, USA.

Smadja, F. A. (1993).
Retrieving collocations from text: Xtract.
Comp. Ling., 19(1):143–177.

References XIII

- Villavicencio, A., Idiart, M., Ramisch, C., Araujo, V. D., Yankama, B., and Berwick, R. (2012a).
  Get out but don't fall down: verb-particle constructions in child language.
  In Berwick, R., Korhonen, A., Poibeau, T., and Villavicencio, A., editors, *Proc. of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss*, pages 43–50, Avignon, France. ACL.

- Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., and Ramisch, C. (2007).
  Validation and evaluation of automatically acquired multiword expressions for grammar engineering.
  In [Eisner, 2007], pages 1034–1043.

- Villavicencio, A., Ramisch, C., Machado, A., de Medeiros Caseli, H., and Finatto, M. J. (2010).
  Identificação de expressões multipalavra em domínios específicos.
  Linguamática, 2(1):15–33.

- Villavicencio, A., Yankama, B., Berwick, R., and Idiart, M. (2012b).
  A large scale annotated child language construction database.
  In *Proceedings of the 8th LREC*, Istanbul, Turkey.

References XIV

Wehrli, E., Seretan, V., and Nerima, L. (2010).
Sentence analysis and collocation identification.
In [Laporte et al., 2010], pages 27–35.

Xu, Y., Goebel, R., Ringlstetter, C., and Kondrak, G. (2010).
Application of the tightness continuum measure to Chinese information retrieval.
In [Laporte et al., 2010], pages 54–62.

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2006).
Automated multiword expression prediction for grammar engineering.
In [Moirón et al., 2006], pages 36–44.