Multiword Expressions and Collocations in NLP: A State of the Art Overview

Valia Kordoni Markus Egg
kordonie,markus.egg@anglistik.hu-berlin.de

Humboldt-Universität zu Berlin (Germany)

PARSEME COST Action IC1207 Inaugural Meeting Warsaw
16.-18.09.2013

Road Map I

Road Map I

MWEs: Overview

Multiword Expressions

- their syntactic or semantic properties cannot be derived from their parts [Sag et al., 2002a, Villavicencio, 2005]
- phrasal verbs (e.g., *get along*), noun compounds (e.g., *frying pan*), institutionalised phrases (e.g., *bread and butter*)
- fixed (*ad hoc*) vs flexible (*touch/find a nerve*) expressions
- opaque (*kick the bucket*) vs transparent (*eat up*) semantics
- MWEs must be listed in a lexicon [Evert, 2004]
- a combination of lexemes that must be treated as a unit at some level of linguistic processing. [Calzolari et al., 2002]

MWEs

Multiword expressions: *a first definition*

A **multiword expression** (MWE) is [Baldwin and Kim, 2010]

- decomposable into multiple simplex words
- lexically, syntactically, semantically, pragmatically and/or statistically idiosyncratic

Some examples

- *San Francisco, ad hoc, by and large, Where Eagles Dare, kick the bucket, part of speech, in step, the Oakland Raiders, telephone box, call (someone) up, take a walk, take (unfair) advantage of, pull strings, kindle excitement, fresh air, ...*

MWEs: Characteristics

Lexicosyntactic Idiomaticity

- by and large (???) $=$ by(P) and(conj) large(Adj)
- hit and run (V [trans]) $=$ hit (V [intrans]) and(conj) run (V [intrans])
- ad hoc (Adj) $=$ ad(?) hoc(?)

MWEs: Characteristics

Semantic Idiomaticity (Non-Identifiability)

- *kick the bucket* = die'
- *spill the beans* = reveal' (secret')
- *kindle excitement* = kindle' (excitement')
  - this includes institutionalisation/conventionalisation, as in *bread and butter*

Pragmatic Idiomaticity

- Situatedness: the expression is associated with a fixed usage context *good morning, all aboard*

MWEs: Characteristics

Statistical Idiomaticity

	unblemished	spotless	flawless	immaculate	impeccable
eye	−	−	−	−	+
gentleman	−	−	?	−	+
home	?	+	−	+	?
lawn	−	−	?	+	−
memory	−	−	+	−	?
quality	−	−	−	−	+
record	+	+	+	+	+
reputation	+	−	−	+	+
taste	−	−	−	−	+

Table : Adapted from [Cruse, 1986]

MWEs: Characteristics

MWE Markedness

MWE	Marked				
	Lex	Syn	Sem	Prag	Stat
ad hominem	✔	?	?	?	✔
at first	✗	✔	✗	✗	✗
first aid	✗	✗	✔	✗	?
salt and pepper	✗	✗	✗	✗	✔
good morning	✗	✗	✗	✔	✔
cat's cradle	✔	✔	✔	✗	?

MWEs: Characteristics

Other Indicators of MWE-hood ([Fillmore et al., 1988a], [Liberman and Sproat, 1992], [Nunberg et al., 1994])

- Figuration: the expression encodes some metaphor, metonymy, hyperbole, etc.
  - figurative expressions: *bull market, beat around the bush*
  - non-figurative expressions: *first off, to and fro*

MWEs: Characteristics

Other Indicators of MWE-hood ([Fillmore et al., 1988a], [Liberman and Sproat, 1992], [Nunberg et al., 1994])

- informality: the expression is associated with more informal or colloquial registers
- affect: the expression encodes a certain evaluation of affective stance toward the thing it denotes

MWEs: Characteristics

Other Indicators of MWE-hood ([Fillmore et al., 1988a], [Liberman and Sproat, 1992], [Nunberg et al., 1994])

- Prosody: the expression has a distinctive stress pattern which diverges from the norm
  - prosodically-marked MWE: *soft spot*
  - prosodically-unmarked MWE: *first aid, red herring*
  - prosodically-marked non-MWE: *dental operation*

MWEs: Theoretical Linguistic Background

The study of MWEs

- is almost as old as linguistics itself
- in Generative Grammar, representing idioms poses a challenge, e.g., the idiom *first off* is an adverbial locution synonym to *firstly*
- in Construction Grammar, [Fillmore et al., 1988b]
  - *idiomatic* entries and their specific syntactic, semantic, and pragmatic characteristics are put into an appendix to the set of lexical units and syntactic rules of a language model
  - this makes idioms part of the core of the grammar, i.e., a full description of a language includes idioms and their properties

MWEs: Theoretical Linguistic Background

Psycholinguistics and cognitive linguistics have worked on learning

- verb-particle constructions [Villavicencio et al., 2012]
- noun compounds [Devereux and Costello, 2007]
- light verb constructions and
- multiword terms [Lavagnino and Park, 2010] based on corpora evidence and sophisticated cognitive models; these models try to validate computational models for MWE acquisition by checking their correlation with experiments that use similar models for human language acquisition [Joyce and Srdanović, 2008, Rapp, 2008]

MWEs: CL Background

In computational linguistics

- the study of MWEs arose from the availability of very large corpora and of computers capable of analysing them by the end of the 80's and beginning of the 90's
- the aim was to build systems for computer-assisted lexicography and terminography of multiword units [Choueka, 1988]
- [Smadja, 1993] proposed Xtract, a tool for collocation extraction based on some simple POS filters and on mean and standard deviation of word distance
- [Church and Hanks, 1990] suggested a more sophisticated association measure based on mutual information

MWEs: CL Background

In computational linguistics

- [Dagan and Church, 1994] propose the terminographic environment Termight, which uses this association score, performs bilingual extraction, and provides tools to easily classify candidate terms, find bilingual correspondences, define nested terms, and investigate occurrences through a concordancer
- [Justeson and Katz, 1995] propose a simple approach based on a small set of POS patterns and frequency thresholds

MWEs: CL Background

In computational linguistics

- [Dunning, 1993] proposed a 2-gram measure called *likelihood ratio*. It estimates directly how more likely a 2-gram is than expected by chance. In addition to being theoretically sound, Dunning's score is also easily interpretable. Nowadays, measures based on likelihood ratio (e.g., the log-likelihood score) are still largely employed in several MWE extraction contexts

MWEs: CL Background

In computational linguistics

- In the 2000's, the Stanford MWE project (`http://mwe.stanford.edu/`) revived interest of the NLP community in this topic.
    - Among its seminal papers is the "pain-in-the-neck" paper by [Sag et al., 2002b]. It provides an overview of MWE characteristics and types and presents some methods for dealing with them in the context of grammar engineering.
    - The Stanford MWE project is also at the origin of the MWE workshop series

MWEs: their importance for Linguistics and CL

And why is it that we care about MWEs?

- Because of the role of MWEs in:
  - Lexicography/dictionary making
  - Idiomaticity (coherent semantics)
  - Overgeneration
  - Undergeneration
  - Relevance in NLP and LT applications, including MT, IR, QA, ...

Road Map I

MWEs: Computational Methods

Overview

Adapted from [Anastasiou et al., 2009]

- **Acquisition**
  - **Extraction**
    How can we build a list of MWE types from corpora?
  - **Identification**
    How can we locate the tokens that correspond to MWEs *in context*?

MWEs: Computational Methods

Overview (contd.)

Adapted from [Anastasiou et al., 2009]

- **Classification**
  - **Interpretation**
    How can we discover the syntactic and semantic relations between the units that compose a MWE type?
  - **Disambiguation**
    How can we disambiguate the syntactic and semantic properties of a MWE token *in context*?

MWEs: Computational Methods

Overview (contd.)

Adapted from [Anastasiou et al., 2009]

- **Representation**
  How can we represent complex MWEs in computational lexicons?
- **Tasks and applications**
  How can we integrate MWEs in NLP tasks (parsing, WSD) and applications (IR, MT)?

MWEs: Computational Methods

"Discovering" MWEs: Co-occurrences

- If *a word is characterized by the company it keeps* [Firth, 1957] then we can try to find MWEs using information about how often words co-occur together
- **Hypothesis**: the more frequently some words occur together, the more likely it is that they form a MWE

"Discovering" MWEs: Filtering with Association Measures

Statistical association measures (AMs)

- can give indication of strength of the association between words (or n-grams)
  - based on frequency of words individually and as a group

"Discovering" MWEs - Filtering with Association Measures

AMs for Ranking MWE Candidates

- **Hypothesis**: If the words are dependent then the candidate is a MWE
  1. Determine the probability given by the *Null Hypothesis* (that they are independent)
  2. Compare with the probability given by a statistical measure
     - t-test, Pearson's X^2, Pointwise Mutual Information, Mutual Information, ...
  3. If Null Hypothesis is rejected then they are dependent (MWEs)

"Discovering" MWEs: Alternative Measures: Entropy-based

Permutation Entropy

Hypothesis: MWEs prefer a certain word order (*give a demo* vs *a demo give*)

- If a candidate is result of random combination of words then word order in n-gram is not important: *of alcohol and*, *and of alcohol*, *alcohol and of*, etc
- Entropy: $S = -\frac{1}{\log N} \sum_{perm} P(abc) \log P(abc)$: $S \rightarrow 0$ (prevalent order) $\longrightarrow$ possible MWE

MWE	Pages	S
the burden of	36,600,000	0.366
but also in	27,100,000	0.038
to bring together	25,700,000	0.086
points of view	24,500,000	0.017
and the more	23,700,000	0.512
taking into account the	22,100,000	0.009

Evaluation of the Extraction of MWEs

Factors in MWE Extraction [Evert and Krenn, 2005]

- corpus size and type
- MWE type and language
- AMs

Comparison of AMs

- 84 measures among which some are rank-equivalent to one another [Pecina, 2008]
- comparison of their combination [Ramisch et al., 2008]

More on Evaluation of the Extraction of MWEs

For statistical approaches there are two important questions

- Q1 How reliable/generalizable are the results for a given corpus?
- Q2 How precise an association measure is to distinguish MWEs from noise?

Grammar Engineering and Parsing

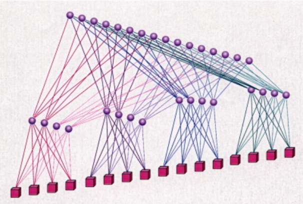
- Lexical coverage is a major barrier to broad-coverage linguistically deep processing
  - 40% parsing failures caused by missing lexical entries [Baldwin et al., 2004]
- MWEs are a significant part of the lexicon
  - Detect potential errors in parsing involving sequences of words
  - Identify MWE candidates
  - Generate new lexical entries based on corpus data

Extension of a hand-crafted linguistic resource with MWEs: English Resource Grammar [Flickinger, 2000]

- A large scale broad coverage precision HPSG grammar
- Lexicon coverage is a major problem
- MWEs comprise a large portion of the missing lexical entries

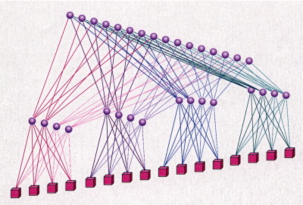
Lexical hierarchy and atomic lexical types

- The lexical information is encoded in atomic lexical types
- A lexicon is a $n : n$ mapping between lexemes and atomic lexical type



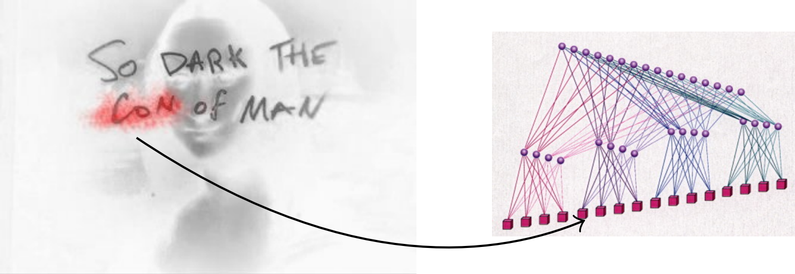
Lexical hierarchy and atomic lexical types

- The lexical information is encoded in atomic lexical types
- A lexicon is a $n : n$ mapping between lexemes and atomic lexical type



Lexical hierarchy and atomic lexical types

- The lexical information is encoded in atomic lexical types
- A lexicon is a $n : n$ mapping between lexemes and atomic lexical type



Maximum Entropy Model-based Lexical Type Predictor

- A statistical classifier that predicts for each occurrence of an unknown word or a missing lexical entry
- Input: features from the context
- Output: atomic lexical types

$$p(t, c) = \frac{exp(\sum_i \theta_i f_i(t, c))}{\sum_{t' \in T} exp(\sum_i \theta_i f_i(t', c))}$$

"Words-with-spaces" vs. compositional approaches

Words-with-spaces approach [Zhang et al., 2006]

- Assign lexical types for the entire MWE
- Grammar coverage significantly improves
- Loss in generality for productive MWEs

Compositional approach

- Assign new lexical entries for the head word to treat the MWE as compositional
- Hopefully the grammar coverage improves without drop in accuracy

"Words-with-spaces" vs. compositional approaches

Words-with-spaces approach [Zhang et al., 2006]

- Assign lexical types for the entire MWE
- Grammar coverage significantly improves
- Loss in generality for productive MWEs

Compositional approach

- Assign new lexical entries for the head word to treat the MWE as compositional
- Hopefully the grammar coverage improves without drop in accuracy

Experiment

- Rank all the MWE candidates according to the three statistical measures: MI, χ^2, PE, and select the top 30 MWE with highest average ranking
- Extract sub-corpus from BNC_f which contains at least one of the MWE for evaluation (674 sentences)
- Use heuristics to extract head words (20 head words)
- Run lexical acquisition for head words on the sub-corpus (21 new entries)

Grammar Coverage

	item #	parsed #	avg. analysis #	coverage %
ERG	674	48	335.08	7.1%
ERG + MWE	674	153	285.01	22.7%

- The coverage improvement is largely compatible with the results of "words-with-spaces" approach reported in [Zhang et al., 2006] (about 15%)
- Great reduction in lexical entries added

Grammar Accuracy

- 153 parsed sentences are analyzed by hand
- 124 (81.0%) of them receive at least one correct/acceptable analysis (comparable to the accuracy reported by [Baldwin et al., 2004])
- Parse selection model finds best analysis in top-5 for 66% of the cases, and top-10 for 75%

Outlook

- Hand-crafted precision grammars usually face coverage/robustness challenges when applied to unseen data with unknown words/MWEs, unknown constructions, etc., all over the place
- [Baldwin et al., 2004] reported parsing coverage of **18**% on unseen BNC data parsed with the ERG, with the majority of parsing failures related to missing lexical entries
- The Lexical Type Prediction model presented as an example above is used to handle unknown words (simplex and MWE) on-the-fly
- With the use of this model the ERG achieves around **84**% parsing coverage on unseen WSJ data

Outlook

Other “Deep” Parsing Systems

- LFG
  - XLE 79.6% F-Score [Kaplan et al., 2004]
- CCG
  - C&C 81.86% F-Score [Clark and Curran, 2007]
- HPSG
  - Enju 82.64% F-Score [Sagae et al., 2008]
- The aforementioned systems are evaluated on 700 sentences selected from WSJ data (PARC 700), using Grammatical Relations (GR)

Road Map I

1. MWEs: Theoretical Background & Motivation
   - Definitions
   - Characteristics
   - Linguistic and CL Theories
2. MWEs: Computational Methods
   - “Discovering” MWEs
   - NLP Tasks and Applications
3. Resources, tasks and applications
   - Tools
   - Resources
   - Tasks and applications
   - Evaluation
4. Future challenges and open problems

Tools for acquisition

Text::NSP

- *N*-gram statistics in text files
- Set of Perl scripts for counting and calculating AMs
- Mostly 2-grams, some measures for 3- and 4-grams
- Customization: sub-*n*-gram counts and non-tokens

`http://search.cpan.org/dist/Text-NSP`

[Pedersen et al., 2011, Banerjee and Pedersen, 2003]

Tools for acquisition

UCS

- Large set of sophisticated AMs
- Input: list of bigrams and their counts (proper extraction must be performed externally, e.g. with NSP)
- Perl and R scripts, includes advanced statistical tools for evaluation

http://www.collocations.de/software.html

[Evert, 2004]

Tools for acquisition

LocalMaxs

- Extracts MWEs based on the local maxima of the distribution of a customisable AM
- Relaxed and strict versions
- Non-contiguous variation
- Scalability

http://hlt.di.fct.unl.pt/luis/multiwords/

[Silva and Lopes, 1999, da Silva et al., 1999]

Tools for acquisition

Varro

- Find regularities in treebanks
- Rank regular subtrees by description length

`http://sourceforge.net/projects/varro/`

[Martens, 2010, Martens and Vandeghinste, 2010]

Tools for acquisition

mwetoolkit

- Multi-level patterns for candidate generation
- Several filtering methods
- Focused on genericity and flexibility

`http://mwetoolkit.sourceforge.net`

[Ramisch et al., 2010a, Ramisch et al., 2010b]

Tools for acquisition

Embedded

- FIPS parser [Seretan and Wehrli, 2009, Seretan and Wehrli, 2011]
- Stanford parser [Green et al., 2011]
- Phrasal verbs in RASP
- Most parsers include (minimal) MWE processing

Tools for acquisition

Related tools

- Complex corpus searches: CQP [Christ, 1994] and Manatee [Rychlý and Smrz, 2004]
- Terminology extraction
  - TermoStat
    http://olst.ling.umontreal.ca/~drouinp/termostat_web/
  - AntConc
    http://www.antlab.sci.waseda.ac.jp/software.html
  - TerMine
    http://www.nactem.ac.uk/software/termine/
- Named entity recognition

Tools for acquisition

Which one to chose? [Ramisch et al., 2012]

	LocMax	mwetk	NSP	UCS
Cand. extr.	+	+	+	−
N-grams $n > 2$	+	+	+	−
Non-adjacent	−	+	+	
Ling. filter	−	+	−	−
Robust measures	−	−	+	+
Large corpora	Partly	+	+	−
Language independent	+	Partly	+	+
Token identification	−	−	−	−
Availability	Free	Free	Free	Free

Resources

Why do we need MWE acquisition?

- MWEs are very frequent in human languages [Jackendoff, 1997]
- Computational resources (corpora, grammars, lexicons) do not reflect this

Resources

Corpora

- At least 17% of Europarl sentences contain a phrasal verb
- 70% of terms in Genia are multiwords
- Flat annotation of noun compounds in treebanks (PTB, French treebank, etc)

Resources

Wordnet

	Non-MWE	MWE
Nouns	57535	60292
Verbs	8729	2829
Adverbs	3796	714
Adjectives	21012	496

- Other languages?
- Missing MWE types (e.g. support verb constructions)?
- New expressions?

Resources

Wordnet

	Non-MWE	MWE
Nouns	57535	60292
Verbs	8729	2829
Adverbs	3796	714
Adjectives	21012	496

- Other languages?
- Missing MWE types (e.g. support verb constructions)?
- New expressions?

Tasks and applications

- Parsing
- Information retrieval
- Word sense disambiguation
- Machine translation
- Educational testing
- Sentiment analysis

Tasks and applications

Machine translation (rule-based)

- Morphological and syntactic analysis in ITS-2 [Wehrli, 1998, Wehrli et al., 2010]
- MWE-specific rules in semantic transfer system Jaen [Haugereid and Bond, 2011]
- French-japanese terms [Morin and Daille, 2010]
- Web as corpus for disambiguating traslation [Grefenstette, 1999]
- Japanese compounds through compositional translation + SVM ranker [Tanaka and Baldwin, 2003, Baldwin and Tanaka, 2004]

Tasks and applications

Machine translation (statistical)

- Phrases in Moses [Koehn et al., 2007]
- Static and dynamic strategies for English MWEs from Wordnet [Carpuat and Diab, 2010]
- Monolingual paraphrases for increasing training data [Nakov, 2008]
- Pre- and post-processing for German compounds [Stymne, 2011, Stymne, 2009]
- Named entities and compound verbs tokenisation [Pal et al., 2010]
- Corpus and phrase-table artificial extensions [Ren et al., 2009]

Evaluation context

1. What are the acquisition goals (that is, the target applications) of the resulting MWEs?
2. What is the nature of the evaluation measures that we intend to use?
3. What is the cost of the resources (dictionaries, reference lists, human experts) required for the desired evaluation?
4. How ambiguous are the target MWE types?

Evaluation context

Acquisition goals

- **Intrinsic**: Evaluate the MWEs per se, using human annotation or gold standard dictionaries.
- **Extrinsic**: Evaluate an application output which includes MWE acquisition.

Acquisition context

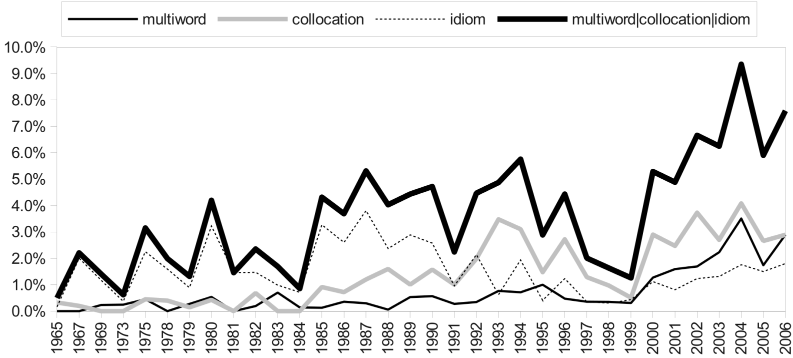
Generalisation of evaluation results depends on parameters of acquisition context:

- Characteristics of target MWEs
  - Type
  - Language
  - Domain
- Characteristics of corpora
  - Size
  - Nature
  - Level of analysis
- Existing resources

Road Map I

1. MWEs: Theoretical Background & Motivation
   - Definitions
   - Characteristics
   - Linguistic and CL Theories
2. MWEs: Computational Methods
   - “Discovering” MWEs
   - NLP Tasks and Applications
3. Resources, tasks and applications
   - Tools
   - Resources
   - Tasks and applications
   - Evaluation
4. Future challenges and open problems

MWE community



MWE community

Trending topics

- Semantics
- Multilingualism
- Applications
- Evaluation
- Machine learning

MWE community

Current and future activities

- MWE workshop series (9th edition at NAACL HLT 2013: http://aclweb.org/anthology//W/W13/W13-10.pdf; 10th edition planned in conjunction to EACL 2014)
- ACM TSLP Special Issue on MWEs in 2 parts (http://dl.acm.org/citation.cfm?id=2483691&picked=prox)
- SIGLEX-MWE Section (http://multiword.sourceforge.net/)
- ICT COST Action IC1207: Parsing and multiword expressions - Towards linguistic precision and computational efficiency in natural language processing (PARSEME; http://www.cost.eu/domains_actions/ict/Actions/IC1207)

MWE community

Future challenges

- Identification is not a solved problem
- Integration and representation in applications
- Robust methods for new MWEs in web texts

Further reading

Please refer to complete list of references :-)

Further reading

Thank you!

For Further Reading I

(1993).
Comp. Ling., 19(1).

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.
70 p.

Baldwin, T., Bender, E. M., Flickinger, D., Kim, A., and Oepen, S. (2004).
Road-testing the English Resource Grammar over the British National Corpus.
In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Baldwin, T. and Kim, S. N. (2010).
Multiword expressions.
In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.

Baldwin, T. and Tanaka, T. (2004).
Translation by machine of complex nominals: Getting it right.
In Tanaka, T., Villavicencio, A., Bond, F., and Korhonen, A., editors, *Proc. of the ACL Workshop on MWEs: Integrating Processing (MWE 2004)*, pages 24–31, Barcelona, Spain. ACL.

For Further Reading II

Banerjee, S. and Pedersen, T. (2003).
The design, implementation, and use of the Ngram Statistic Package.
In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City, Mexico.

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 Fluhr, C. and Walker, D. E., editors, *Proceedings of the 2nd International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications - RIA 1988)*, pages 609–624, Cambridge, MA, USA. CID.

Christ, O. (1994).
A modular and flexible architecture for an integrated corpus query system.
In *COMPLEX 1994*, pages 23–32, Budapest, Hungary.

For Further Reading III

Church, K. and Hanks, P. (1990).
Word association norms mutual information, and lexicography.
Comp. Ling., 16(1):22–29.

Clark, S. and Curran, J. (2007).
Formalism-Independent Parser Evaluation with CCG and DepBank.
In *Proceedings of ACL2007*.

Cruse, A. (1986).
Lexical Semantics.
Cambridge University Press, Cambridge, UK.

da Silva, J. F., Dias, G., Guilloré, S., and Lopes, J. G. P. (1999).
Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units.
In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, EPIA 1999, pages 113–132, London, UK. Springer.

Dagan, I. and Church, K. (1994).
Termight: Identifying and translating technical terminology.
In *Proc. of the 4th ANLP Conf. (ANLP 1994)*, pages 34–40, Stuttgart, Germany. ACL.

Devereux, B. and Costello, F. (2007).
Learning to interpret novel noun-noun compounds: evidence from a category learning experiment.
In Buttery, P., Villavicencio, A., and Korhonen, A., editors, *Proc. of the ACL 2007 Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 89–96, Prague, Czech Republic. ACL.

For Further Reading IV

Dunning, T. (1993).
Accurate methods for the statistics of surprise and coincidence.
In [jou, 1993], pages 61–74.

Evert, S. (2004).
The Statistics of Word Cooccurrences: Word Pairs and Collocations.
PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany.
353 p.

Evert, S. and Krenn, B. (2005).
Using small random samples for the manual evaluation of statistical association measures.
Comp. Speech & Lang. Special issue on MWEs, 19(4):450–466.

Fillmore, C., Kay, P., and O'Connor, M. (1988a).
Regularity and idiomaticity in grammatical constructions.
Language, 64:501–38.

Fillmore, C. J., Kay, P., and O'Connor, M. C. (1988b).
Regularity and idiomaticity in grammatical constructions: The case of let alone.
Language, 64:501–538.

Firth, J. R. (1957).
Papers in Linguistics 1934-1951.
Oxford UP, Oxford, UK.
233 p.

For Further Reading V

Flickinger, D. (2000).
On building a more efficient grammar by exploiting types.
Natural Language Engineering, 6(1):15–28.

Green, S., de Marneffe, M.-C., Bauer, J., and Manning, C. D. (2011).
Multiword expression identification with tree substitution grammars: A parsing tour de force with French.
In Barzilay, R. and Johnson, M., editors, *Proc. of the 2011 EMNLP (EMNLP 2011)*, pages 725–735, Edinburgh, Scotland, UK. ACL.

Grefenstette, G. (1999).
The World Wide Web as a resource for example-based machine translation tasks.
In *Proc. of the Twenty-First Translating and the Computer*, London, UK. ASLIB.

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

Haugereid, P. and Bond, F. (2011).
Extracting transfer rules for multiword expressions from parallel corpora.
In [Kordoni et al., 2011], pages 92–100.

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

For Further Reading VI

Joyce, T. and Srdanović, I. (2008).
Comparing lexical relationships observed within Japanese collocation data and Japanese word association norms.
In [Zock and Huang, 2008], pages 1–8.

Justeson, J. S. and Katz, S. M. (1995).
Technical terminology: some linguistic properties and an algorithm for identification in text.
Nat. Lang. Eng., 1(1):9–27.

Kaplan, R., Riezler, S., King, T. H., Maxwell, J., and Vasserman, A. (2004).
Speec and accuracy in shallow and deep stochastic processing.
In *Proceedings of HLT-NAACL'04*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007).
Moses: open source toolkit for statistical machine translation.
In *Proc. of the 45th ACL (ACL 2007)*, pages 177–180, Prague, Czech Republic. ACL.

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.

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.

For Further Reading VII

Lavagnino, E. and Park, J. (2010).
Conceptual structure of automatically extracted multi-word terms from domain specific corpora: a case study for Italian.
In Zock, M. and Rapp, R., editors, *Proc. of the 2nd COGALEX workshop (COGALEX 2010)*, pages 48–55, Beijing, China. The Coling 2010 Organizing Committee.

Liberman, M. and Sproat, R. (1992).
The stress and structure of modified noun phrases in English.
Lexical Matters – CSLI Lecture Notes, 24:99–108.

Martens, S. (2010).
Varro: An algorithm and toolkit for regular structure discovery in treebanks.
In Huang, C.-R. and Jurafsky, D., editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 810–818, Beijing, China. The Coling 2010 Organizing Committee.

Martens, S. and Vandeghinste, V. (2010).
An efficient, generic approach to extracting multi-word expressions from dependency trees.
In [Laporte et al., 2010], pages 84–87.

Morin, E. and Daille, B. (2010).
Compositionality and lexical alignment of multi-word terms.
Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing, 44(1-2):79–95.

For Further Reading VIII

Nakov, P. (2008).
Improved statistical machine translation using monolingual paraphrases.
In Ghallab, M., Spyropoulos, C. D., Fakotakis, N., and Avouris, N. M., editors, *Proc. of the 18th ECAI (ECAI 2008)*, volume 178 of *Frontiers in Artificial Intelligence and Applications*, pages 338–342, Patras, Greece. IOS Press.

Nunberg, G., Sag, I., and Wasow, T. (1994).
Idioms.
Language, 70:491–538.

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. (2008).
Reference data for Czech collocation extraction.
In [Grégoire et al., 2008], pages 11–14.

Pedersen, T., Banerjee, S., McInnes, B., Kohli, S., Joshi, M., and Liu, Y. (2011).
The *n*-gram statistics package (text::NSP) : A flexible tool for identifying *n*-grams, collocations, and word associations.
In [Kordoni et al., 2011], pages 131–133.

For Further Reading IX

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.

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. (2010a).
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. (2010b).
mwetoolkit: a framework for multiword expression identification.
In *Proc. of the Seventh LREC (LREC 2010)*, pages 662–669, Valetta, Malta. ELRA.

Rapp, R. (2008).
The computation of associative responses to multiword stimuli.
In [Zock and Huang, 2008], pages 102–109.

For Further Reading X

Ren, Z., Lü, Y., Cao, J., Liu, Q., and Huang, Y. (2009).
Improving statistical machine translation using domain bilingual multiword expressions.
In Anastasiou, D., Hashimoto, C., Nakov, P., and Kim, S. N., editors, *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, pages 47–54, Suntec, Singapore. ACL.

Rychlý, P. and Smrz, P. (2004).
Manatee, bonito and word sketches for Czech.
In *Proceedings of the Second International Conference on Corpus Linguisitcs*, pages 124–131, Saint-Petersburg, Russia.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002a).
Multiword expressions: A pain in the neck for NLP.
In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002b).
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.

Sagae, K., Miyao, Y., Matsuzaki, T., and Tsujii, J. (2008).
Challenges in Mapping of Syntactic Representations for Framework-Independent Parser Evaluation.
In *Proceedings of Workshop on Automated Syntatic Annotations for Interoperable Language Resources at the First International Conference on Global Interoperability for Language Resources (ICGL'08)*, Hong Kong.

For Further Reading XI

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.

Seretan, V. and Wehrli, E. (2011).
Fipscoview: On-line visualisation of collocations extracted from multilingual parallel corpora.
In [Kordoni et al., 2011], pages 125–127.

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.
In [jou, 1993], pages 143–177.

Stymne, S. (2009).
A comparison of merging strategies for translation of German compounds.
In *Proc. of the Student Research Workshop at EACL 2009*, pages 61–69.

Stymne, S. (2011).
Pre- and postprocessing for statistical machine translation into Germanic languages.
In *Proc. of the ACL 2011 SRW*, pages 12–17, Portland, OR, USA. ACL.

For Further Reading XII

Tanaka, T. and Baldwin, T. (2003).
Noun-noun compound machine translation A feasibility study on shallow processing.
In Bond, F., Korhonen, A., McCarthy, D., and Villavicencio, A., editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 17–24, Sapporo, Japan. ACL.

Villavicencio, A. (2005).
The availability of verb-particle constructions in lexical resources: How much is enough?
Journal of Computer Speech and Language Processing, 19.

Villavicencio, A., Idiart, M., Ramisch, C., Araujo, V. D., Yankama, B., and Berwick, R. (2012).
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.

Wehrli, E. (1998).
Translating idioms.
In *Proc. of the 36th ACL and 17th COLING, Volume 2*, pages 1388–1392, Montreal, Quebec, Canada. ACL.

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

For Further Reading XIII

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2006).
Automated multiword expression prediction for grammar engineering.
In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia. Association for Computational Linguistics.

Zock, M. and Huang, C.-R., editors (2008).
Proc. of the COLING 2008 COGALEX workshop (COGALEX 2008), Manchester, UK. The Coling 2008 Organizing Committee.