Chapter 3
State of the Art in MWE Processing

In the previous chapter, we provided the historical and theoretical foundations for the study of multiword expressions. The set of definitions, characteristics and types described give an idea of the difficulty of the computational tasks involving MWEs. The goal of the present chapter is to draw an overview of the state of the art in computational methods for MWE treatment, focusing on acquisition. State-of-the-art techniques to deal with MWEs are the starting point of the methodology proposed in Chap. 5. Information contained in the present chapter allows better comparison and contextualisation of the present work in the computational linguistics panorama.

In Sect. 3.1, we start with a brief review of some practical *elementary notions*, defining concepts like n-grams, frequencies and association measures. These are the tools used by the techniques for automatic *MWE acquisition* described in Sect. 3.2. Although the largest part of research effort in the community has been devoted to acquisition, other tasks such as interpretation, disambiguation and representation are also relevant. Mainly in the last decade, work on these tasks started to emerge, and this is presented in Sect. 3.3.

3.1 Elementary Notions

In this section, we review standard NLP concepts useful in the present work. We focus on general notions that appear recurrently throughout the book, while more detailed explanations of concepts specific to a certain experiment are provided later, whenever they are required.[1]

[1]The goal of this section is not to provide a substantial introduction to empirical methods in computational linguistics. Instead, we remind and try to disambiguate as much as possible the definitions of concepts that are already familiar to the reader to some extent. If this is not the case, we recommend Jurafsky and Martin (2008) as a consolidated and wide introduction to NLP and

C. Ramisch, *Multiword Expressions Acquisition*, Theory and Applications
of Natural Language Processing, DOI 10.1007/978-3-319-09207-2_3

We define a *corpus* as a body of texts used in empirical language studies (Manning and Schütze 1999, p. 6). One usually wants for corpora to be representative of the target language, where the meaning of *representative* depends on the context (e.g., application, domain, genre, sublanguage). In our experiments, we use only written corpora like the collection of speech transcripts from the European parliament (Koehn 2005), the 100-million words British National Corpus (Burnard 2007) and the collection of news from the Brazilian *Folha de São Paulo* newspaper. Half a dozen of sentences in French are not a big enough corpus of general French language, as well as a million sentences of computer science articles in English are not representative if the target application will deal with botany texts or with texts in Portuguese.

A corpus may contain data in one language (monolingual) or in several languages (multilingual); when the sentences in one language are translations of sentences in another language, we consider it as a sentence-aligned *parallel corpus*. We will also use the term *general-purpose* corpus to refer to corpora that contain a wide variety of texts corresponding to most common language use over a given time span, while a *specialised* corpus contains texts of a specific knowledge domain or sub-language, like botany, computer science or sailing. We consider a *word token* to be an occurrence of a word in a corpus while a *word type* is a unique occurrence of a word as a lexeme in a dictionary, thesaurus or other lexical resource. The set of unique word types in a corpus constitutes its *vocabulary*.

In Sect. 3.1.1, we introduce linguistic notions such as part of speech and dependency syntax. We provide an overview of the statistical distribution of words in a corpus in Sect. 3.1.2. Section 3.1.3 is about n-grams, presenting the basics of n-gram probability estimation. In Sect. 3.1.4, we present lexical association measures frequently employed in the automatic acquisition of MWEs.

3.1.1 Linguistic Processing: Analysis

Linguistic analysis is the process of creating more abstract representations from raw text. It is generally seen as a set of steps, each of which must solve ambiguities inherent to language. More sophisticated systems may not solve ambiguities, but represent multiple solutions in the form of weighted lattices (Nasr et al. 2011). However, for concision purposes, we present here a simplified example in the form of a sequence of analysis steps which can be applied on corpora for MWE acquisition.

A corpus may be structured as a set of documents, each document being composed of several paragraphs, which in turn are sequences of sentences. While the higher level divisions are optional and task-dependent, most of the current NLP

Manning and Schütze (1999) for a more specific introduction to empirical methods. Our text is inspired by these two standard reference textbooks.

systems require the text to be split into sentences prior to processing. Splitting the sentences in running text can be accomplished through language-dependent regular expressions on anchor punctuation signs (such as periods and question marks) and lists of common exceptions like abbreviations (*Ph.D.*), acronyms (*Y.M.C.A.*) numbers (*1,399.99*), proper names (*Yahoo!*), filenames and web addresses (*www.google.com*). These can also be modelled through learning sequence tagging models from corpora (Mikheev 2002). Although apparently simple, sentence splitting is challenging in highly structured texts like scientific articles containing many tables, itemised lists and mathematical formulas. Ambiguities about possible splitting points must be dealt with by the system.

Further decomposition takes us from sentences to words. The definition of *word* is discussed in Sect. 2.2.1. In practice, for languages whose writing system uses spaces to separate words, one needs to split from adjacent words punctuation such as commas, periods, apostrophes, dashes and contractions (e.g., the English possessive marker *'s*). It may also be necessary to split contractions such as $du = de + le$ in French and $no = em + o$ in Portuguese.[2] Other morphological phenomena like prefixes, suffixes and agglutination can also be dealt with at this point. The process of word splitting is called *tokenisation*, and is generally accomplished using regular expressions or trained statistical sequence models. For example, consider the sentence:

Example 3.1. "Tomorrow, I'll be paying a visit to Mary's parents."

After tokenisation, it becomes[3]:

Example 3.2. "␣Tomorrow␣,␣I␣'ll␣be␣paying␣a␣visit␣to␣Mary␣'s␣parents␣.␣"

A word in the corpus occurs in its inflected form, also called the *surface form*. A surface form like *parents* is the plural of the base form *parent*, the verb form *paying* is the gerundive of *pay*, and so on. The morphology of languages models word formation (derivation, compounding) and modification (inflection). The latter often encodes information such as gender, number, tense, mood, voice, person, aspect and case of words. Moreover, distinctive capitalisation marking the beginning of a sentence, for example, needs to be normalised so that *Tomorrow* and *tomorrow* are considered as being the same word.[4] The base form from which an inflected word is derived is called *lemma*. The process of assigning lemmas to words is called *lemmatisation*.

[2]Contraction identification usually requires context-aware analysis. For instance, in French, the contraction $des = de + les$ is homonym to the partitive/indefinite article *des*.

[3]We use the character ␣ only to emphasise the spaces between words.

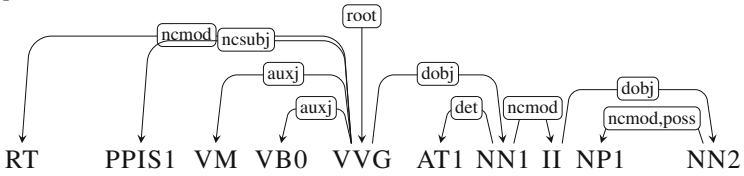
[4]However, it is not enough to lowercase the whole text as case information may be important, for instance, in domain-specific texts (chemical element *NaCl*), acronyms (*NASA*, *CIA*) and to distinguish named entities (*Bill Gates*, *March*) from common words (*pay the bill*, *open the gates*, *the soldiers march*).

Generally, lemmatisation is performed simultaneously or immediately after another process called *part-of-speech tagging*. The latter is the assignment of part-of-speech (POS) tags to each word. POS tags represent the grammatical function of words in a sentence, like nouns, verbs and adverbs. They are useful, for example, to distinguish closed-class words or *function words* like prepositions, pronouns and determiners, from open-class words or *content words* like verbs, nouns, adverbs and adjectives. The software performing POS tagging and, usually, lemmatisation, is the *POS tagger*. The set of all POS tags that can be assigned to words in a corpus or tool is the *tagset*. In some of our experiments, we use a POS tagger called TreeTagger (Schmid 1994), described in Appendix B. When the English version of the TreeTagger is applied to the sentence of Example 3.1, the system performs sentence splitting, tokenisation, lemmatisation and POS tagging, resulting in[5]:

Example 3.3. “ tomorrow$_{[NN]}$, I$_{[PP]}$ will$_{[MD]}$ be$_{[VB]}$ pay$_{[VBG]}$ a$_{[DT]}$ visit$_{[NN]}$ to$_{[TO]}$ Mary$_{[NNP]}$ ’s$_{[POS]}$ parent$_{[NNS]}$.”

The process of going from surface forms to more abstract representations like POS and lemmas is called *analysis*. In order to perform a deeper analysis, one can group POS-tagged words into chunks like noun phrases, and represent chunks as part of a *syntax tree*. There are several formalisms to represent syntactic structures in theory and in practice. We adopted *dependency syntax*. In dependency syntax, all nodes of the syntax tree are the words themselves and the arrows are the dependency relations, tagged with the corresponding relation type (e.g., direct object, subject, determination). A software capable of generating such trees from sentences is a *dependency parser*. For English, we use the RASP parser (Briscoe et al. 2006), described in Appendix B. RASP generates the following surface dependency tree[6] for the example sentence[7]:

Example 3.4.



[5]The tagset used by the TreeTagger in English is available at ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz and reproduced in Appendix D.4.

[6]Actually, RASP does not generate dependency relations directly, but it infers *grammatical relations* using equivalence rules applied to a traditional constituent parsing tree. Relations are mostly acyclic and exceptions can be dealt with on a case by case basis.

[7]Documentation about RASP’s tagset and grammatical relations is available at http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-662.pdf and in Appendix C of Jurafsky and Martin (2008). Moreover, the tags used by RASP for POS and syntax are reproduced in Appendices D.2 and D.3.

Notice that, for the dependency parser, some nodes like punctuation are ignored, as they are considered irrelevant for syntax.[8] Also, the tagset and the lemmas used by the RASP parser for morphosyntactic analysis are much more fine-grained than those of the TreeTagger. Finding equivalences and adapting the granularity of POS tags is a practical problem in NLP and often demands writing dedicated conversion scripts.

Many other parsers and formalisms exist for English and for other languages, and related work on MWE acquisition explores some of them, as we will discuss in Sect. 3.2.1. Nonetheless, we will limit our discussion to dependency parsing because it is the formalism used in our experiments. The advantage of the dependency formalism is that the resulting tree can be represented on a word basis, that is, for every word we assign two labels: the other word of the sentence on which it depends and the type of the relation. This has practical implications in the data structures used to represent parsed corpora, as we will discuss in Chap. 5. Moreover, more meaningful relations such as subject and object tend to appear closer to the root while auxiliaries and determiners appear as leafs, as shown in Example 3.4.

3.1.2 Word Frequency Distributions

In order to design statistical methods for dealing with corpora, one needs to understand how words and word counts behave in text. We will use as a toy example a fragment of 20,000 English sentences randomly chosen from the British National Corpus, henceforth BNC-frg. Table 3.1 summarises the number of tokens and types in the toy corpus. It can be considered as a sample of English language, and therefore the size N of the sample is the number of tokens, that is, around 414K tokens (surface forms). The vocabulary V is a set containing around 37.6K distinct word types.[9]

Let us define a function $c(w) : V \rightarrow \mathbb{N}$ which associates to each word type w in a vocabulary its number of occurrences in a corpus. When more than one corpus is considered simultaneously, the function is subscripted with the name of the corpus in which the token was counted, for instance, $c_{\text{BNC-frg}}(Mary) = 18$.

Table 3.1 Statistics of BNC-frg—sample of 20,000 random sentences taken from the BNC

# of sentences	20,000
# of tokens	414,602
# of types	37,649

[8] This is a simplification, as described by Briscoe et al. (2006).

[9] The *type/token ration*, that is, the number of types with respect to the number of tokens in a text, has been used as a measure of the richness of the vocabulary. This measure depends on the corpus size (Baayen 2001). In BNC-frg, the type/token ratio is of 0.091.

Table 3.2 Counts of the 30 most frequent tokens in BNC-frg

r	$c(w)$	w	r	$c(w)$	w	r	$c(w)$	w
1	20,765	the	11	3,248	for	21	2,173	you
2	19,031	,	12	3,064	it	22	2,146	’
3	16,022	.	13	2,996	was	23	2,029	’
4	11,923	of	14	2,899	’s	24	1,899	by
5	9,830	to	15	2,816	I	25	1,800	are
6	9,771	and	16	2,550	on	26	1,782	at
7	7,346	a	17	2,535	be	27	1,727	have
8	6,758	in	18	2,405	with	28	1,668	not
9	4,351	that	19	2,356	as	29	1,532	from
10	4,029	is	20	2,255	The	30	1,496	he

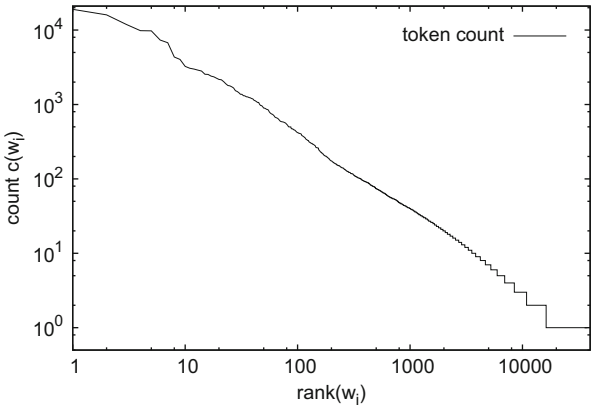


Fig. 3.1 Rank plot of the vocabulary of BNC-frg, with counts in descending order

The values of $c(\cdot)$ for the 30 most frequent tokens of BNC-frg are listed in Table 3.2. The most frequent words in any corpus are generally function words like prepositions, determiners, pronouns and punctuation signs. Notice that, as our corpus was not analysed using the tools described in Sect. 3.1.1, the words *the* and *The* are considered as two distinct tokens. Also, because of an encoding problem, there are two different apostrophe characters.

In Sect. 3.1.3, our goal is to estimate the probability of an arbitrary token or sequence of tokens. Therefore, it is useful to study the empirical distribution of the values of function $c(\cdot)$. Unlike the heights of humans or the grades of students in a class, the word counts in a corpus are not normally distributed around the mean. Instead, they are distributed according to a *power law*, also known as *Zipfian distribution*. Many other events in the world are distributed according to power laws (Newman 2005).

In order to illustrate the Zipfian distribution, we will use the rank plot of Fig. 3.1. A rank plot is a graphic where the word counts are sorted in descending order and assigned to their rank positions r, like those in the first column of Table 3.2. Formally, the rank r of a given word w can be defined as the value of a bijective

function $rank(w) : V \rightarrow [1..|V|]$ which assigns a distinct integer to each word respecting the constraint $(\forall w_1, w_2 \in V)[rank(w_1) \leq rank(w_2) \iff c(w_1) \geq c(w_2)]$. Notice that, in the presence of ties, the *rank* function is not uniquely defined. Any valid function respecting the aforementioned constraint could be used. Therefore, we assume that lexicographic order is used to assign the ranks of words with identical numbers of occurrences, uniquely defining the *rank* function.

The rank plot of Fig. 3.1 is in logarithmic scale, otherwise it would be impossible to visualise the counts. The main characteristic of power laws is that there is a very large number of rare events. In BNC-frg, for example, there are 21,423 word types occurring only once in the corpus,[10] that is, almost 57 % of the vocabulary. On the opposite end of the range, frequent words correspond to a tiny portion of the vocabulary. The graphic shows that the number of words decreases exponentially in the ranked vocabulary. In other words, Zipf's law states that the number of occurrences of a type in the corpus is inversely proportional to its position in the rank.

A derived property is that the size of the vocabulary increases logarithmically with the size of the corpus. This is exemplified by plotting the number of types in a corpus as its size, that is, the number of tokens, increases, on shown in Figure 3.2. While the number of tokens increases linearly, the number of types increases fast at the beginning and much slower afterwards. This happens because, as the sample increases, many common words tend to be repeated while new words become harder to find. One could assume that, if an infinitely large corpus (the sample) was available, the size of the vocabulary would stop growing at some point, converging to the total number of words used in that language (the population).

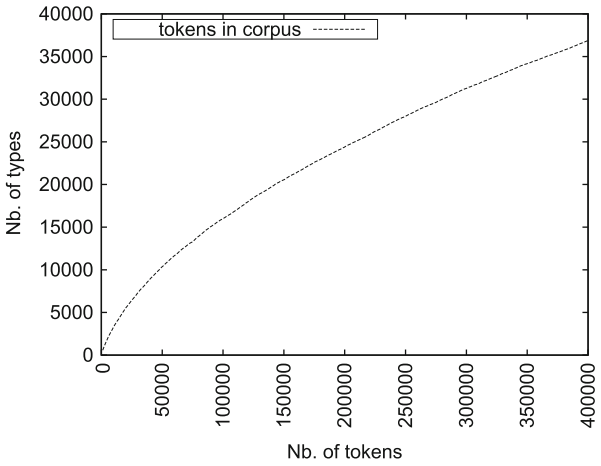


Fig. 3.2 Tokens in the corpus versus types in the vocabulary

[10] A word occurring once in the corpus is called a *hapax*, from the Greek *hapax legomena*.

The statistics of a corpus depend on its size, language, genre, domain and sublanguage. Nonetheless, the logarithmic relation between the number of word tokens and the number of different word types holds as well as Zipf's law. This means that, in general, around half of the words in a corpus occur only once. Distributions like these are called *large number of rare events* (LNRE). When the underlying model is a LNRE distribution, specific statistical tools able to deal with sparse data must be employed. Besides, one needs to be careful because standard assumptions for a sample drawn from a population normally distributed do not apply to corpora. Operations like parameter estimation, hypothesis testing and the like need to be adapted when working with LNRE distributions. This has an impact on the types of lexical association measures that can be used for unsupervised MWE acquisition (Sect. 3.1.4). For further details, one may refer to Baayen (2001).

3.1.3 *N*-Grams, Language Models and Suffix Arrays

When we consider word sequences, each token in the corpus is represented as w_i, where the subscript i stands for its position with respect to other tokens. For instance, a sequence of n consecutive tokens in the corpus can be represented as $w_1 w_2 \ldots w_{n-1} w_n$. Such contiguous sequences are called n-grams.[11] We use the abbreviated notation w_i^j to represent an n-gram formed by $j - i + 1$ words w_i through w_j. By extension, the function $c(\cdot)$ can be applied to n-grams and returns the number of times they occur in a corpus. For example, $c_{\text{BNC-frg}}(I\ will\ be) = 5$ because this 3-gram occurs 5 times in the corpus BNC-frg.

An *language model* (LM) is a probabilistic model that estimates to what extent a group of words belongs to a certain language. An n-gram LM is a set of probability density functions that estimate the probabilities of any n-gram in a language. For instance, an n-gram like *I will be paying* is more plausible in English than *will paying I be*, thus the model will assign it a higher probability. LMs are widely employed not only in MWE acquisition but in many other NLP applications like speech recognition, OCR, spell checking, MT. They are often used to choose among several possible outputs because they simulate grammatical and semantic preferences in sentences.

One way to estimate n-gram probabilities is to learn them from a sample of language: the training corpus. We can count all the n-grams in the training corpus and then return as a probability estimates the relative frequencies of the n-grams. For instance, according to Table 3.1, the BNC-frg corpus contains 414,602 tokens or unigrams. If we use BNC-frg as training corpus, the probability estimate p of the

[11] Discontiguous sequences are sometimes referred to as *flexigrams*, that is, n-grams with gaps.

unigram *Mary* is $p(Mary) \approx \frac{18}{414{,}602}$ and the probability estimate of the 3-gram *I will be* is $p(I\ will\ be) \approx \frac{c(I\ will\ be)}{N} = \frac{5}{414{,}602}$.

Although a good idea in theory, it is not feasible to store all the counts for each distinct n-gram of arbitrary length (1 to N) in a large corpus, as the number of n-grams grows quickly. For instance, BNC-frg contains 37,651 unigrams, 210,183 2-grams, 346,450 3-grams and so on. In order to solve this practical problem, we first apply the probability chain rule, that is, for an arbitrary n-gram:

$$p(w_1^n) = p(w_1) \times p(w_2|w_1) \times p(w_3|w_1^2) \ldots p(w_n|w_1^{n-1}) = p(w_1) \times \prod_{k=2}^{n} p(w_k|w_1^{k-1}) \tag{3.1}$$

We further simplify calculations by applying the Markov assumption in order to approximate the conditional probability of a token given a short history instead of using the whole preceding sequence. That is, given $m > 1$ as the fixed maximum size of n-gram that we can store, we ignore all words preceding w_{k-m+1}. The *order* m of the model typically ranges from 2 to 5 according to the target application. This simplification assumes that the presence of a word only depends on a short number of words to the left of it, completely ignoring the right context.

$$p(w_k|w_1^{k-1}) \approx p(w_k|w_{k-m+1}^{k-1}) \tag{3.2}$$

For instance, let us consider a model of order $m = 2$, built using the BNC-frg corpus as training data. Given this model, we want to estimate the probability of the 4-gram *I will be visiting*. Thus, $p(I\ will\ be\ visiting) = p(I) \times p(will|I) \times p(be|\ will) \times p(visiting|be) = \frac{c(I)}{N} \times \frac{c(I\ will)}{c(I)} \times \frac{c(will\ be)}{c(will)} \times \frac{c(be\ visiting)}{c(be)} = \frac{2{,}816}{414{,}602} \times \frac{34}{2{,}816} \times \frac{312}{1{,}093} \times \frac{1}{2{,}535} = 0.000000009$.

This model uses the principle of *maximum likelihood estimation* (MLE), that is, it assumes that the sample *is* the population. In other words, the chosen model parameters are those that maximise the likelihood of the observed sample. The problem with MLE is that it does not take into account n-grams that were not observed in the corpus as a side effect of sampling a very large event space. In other words, no matter how large a training corpus is, a large number of perfectly valid n-grams will surely be missing from the model, thus yielding zero probability for the whole product. In order to solve this problem, current LM tools implement sophisticated *smoothing* techniques. The idea of smoothing is to assign some probability mass to unseen events, discounting it from the probabilities of seen n-grams (Chen and Goodman 1999; Good 1953; Kneser and Ney 1995). Furthermore, it is also possible to use *backoff* in order to estimate the probabilities of larger unseen n-grams by combining the probabilities of smaller n-grams contained in them. Because such techniques are rarely employed in empirical MWE acquisition from corpora, we

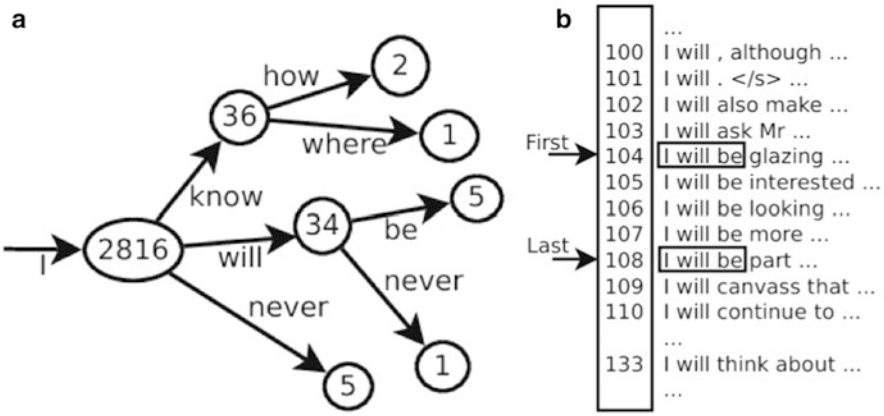


Fig. 3.3 (**a**) Example of suffix tree. (**b**) Example of suffix array

will not discuss their details here. One of the rare works concerning smoothing for MWE acquisition is that of Lapata (2002).

When dealing with very large corpora, it is crucial to have efficient access to n-gram counts in order to estimate their probabilities. The intuition behind quick access to n-gram counts in a corpus is to organise the n-grams in a data structure that allows fast search (that is, direct access or binary search). N-gram models with a fixed order m can be represented using structures based on suffix trees. A *suffix tree* is a representation in which each edge is labelled with a word and each node contains a count. Concatenating the words on the edges of a path from the root to a node n_i generates an n-gram whose count is stored in n_i. For example, in Fig. 3.3a, the path *I will be* leads to a node containing the value of $c(\text{I will be}) = 5$. In order to optimise the access to the child nodes, it is possible to build hash tables for constant access or ordered lists for binary search.

While suffix trees are appropriate for LMs with fixed order, counting arbitrarily long n-grams requires another kind of data structure. A *suffix array* is an efficient structure to represent n-grams of arbitrary size (Manber and Myers 1990; Yamamoto and Church 2001). The corpus is viewed as an array of N words w_1 to w_N. Each word w_i is the beginning of a corpus suffix of size $N - i + 1$, for instance, $w_{N-2}w_{N-1}w_N$ is a suffix of size 3. The trick is that the list containing all the N suffixes is sorted in lexicographic order. Therefore, one can perform binary search in order to locate the first and the last positions starting with the searched n-gram. For example, in Fig. 3.3b we represent part of a suffix array of BNC-frg. If we want to know how many times the n-gram *I will be* occurs in the corpus, we will perform two binary searches in $O(\log N)$ time to find the first index F and last index L in the array containing a suffix which starts with the searched n-gram. The number of occurrences of the n-gram is then simply $L - F + 1 = 108 - 104 + 1 = 5$. If now we need to obtain the count for *I will*, we repeat the procedure and find $133 - 100 + 1 = 34$.

Table 3.3 Top-15 most frequent n-grams in BNC-frg

r	$c(w_1w_2)$	w_1w_2	$c(w_1)$	$c(w_2)$	$E(w_1w_2)$	t-score
1	3,060	of the	11,923	20,765	597.2	44.5
2	1,788	in the	6,758	20,765	338.5	34.3
3	1,139	to the	9,830	20,765	492.3	19.2
4	772	on the	2,550	20,765	127.7	23.2
5	738	and the	9,771	20,765	489.4	9.2
6	733	to be	9,830	2,535	60.1	24.9
7	687	for the	3,248	20,765	162.7	20.0
8	526	at the	1,782	20,765	89.3	19.0
9	525	by the	1,899	20,765	95.1	18.8
10	500	that the	4,351	20,765	217.9	12.6
11	473	of a	11,923	7,346	211.3	12.0
12	457	from the	1,532	20,765	76.7	17.8
13	456	with the	2,405	20,765	120.5	15.7
14	369	it is	3,064	4,029	29.8	17.7
15	362	in a	6,758	7,346	119.7	12.7

In our implementation (see Sect. 5.1.2) each suffix is represented with an integer index pointing to the position in the corpus where it starts, thus optimising memory use. Thus, a suffix array uses a constant amount of memory with respect to N: if every word and every word position in the corpus is encoded as a 4-byte integer, a suffix array uses precisely $4 \times 2 \times N$ bytes, plus the size of the vocabulary, which is generally very small if compared to N.

3.1.4 Lexical Association Measures

The principle of corpus-based MWE acquisition is that words that form an expression will co-occur more often than if they were randomly combined by a coincidence of syntactic rules and semantic preferences. In this context, *lexical association measures* are applied to n-gram counts in order to estimate how much the occurrences of two or more words depend on each other.[12]

A simple method to acquire MWEs from corpora is to use ranked n-gram lists. For example, Table 3.3 lists the 15 most frequent n-grams of BNC-frg. Unfortunately, all of the returned items are uninteresting combinations of function words like determiners *the* and *a*, prepositions and auxiliary verbs. Moreover, the list only contains 2-grams and no 3-grams and larger n-grams. This is a consequence of the fact that the count of a larger n-gram will always be less than or equal to

[12]The term *association measure* is standard in MWE acquisition, but it would be more appropriate to talk about association scores instead, since not all the scores discussed here are proper measures.

the count of the n-grams that it contains, thus biasing the acquisition towards short n-grams.

We could solve these problems by separately acquiring n-grams of different lengths, using regular expression patterns to filter out sequences of function words contained in stopword lists or matching unwanted POS tags. This is actually performed in many real-world systems, specially for automatic terminology acquisition, with surprisingly good results (Justeson and Katz 1995; Ramisch 2009). However, if we are to acquire general MWEs (and not only multiword terms), we need a more sophisticated way to tell whether an n-gram is just a random co-occurrence of frequent words or whether it presents some statistical idiomaticity.

A common preprocessing step when dealing with n-gram counts is to eliminate all combinations that occur less than a fixed threshold. This is important because statistics tend not to be reliable in low frequency ranges. As the counts decrease, it is impossible to distinguish statistically significant events from coincidences due to sampling error. Unfortunately, there is no rule or algorithm for determining the value of such threshold except common sense and trial and error. For example, statistics calculated over hapax are surely unreliable while setting the threshold at 100 occurrences will probably result in too little data (if any).

3.1.4.1 Measures Based on Hypothesis Testing

Now, in order to investigate whether an n-gram is a MWE, let us assume that words are combined randomly. That is, the occurrence of words at given positions are independent events. This hypothesis does not hold, otherwise languages would have no grammar. Nonetheless, it provides a powerful way to test the association strength between words. By the definition of statistical independence, if the occurrence of a word w_2 does not depend on the occurrence of the preceding word w_1, then we expect that the joint probability of the 2-gram is the product of the probabilities of the individual events, that is:

$$p(w_1^2) = p(w_1) \times p(w_2) \tag{3.3}$$

For the sake of simplicity, let us use MLE estimators for the probabilities of the individual words through relative frequencies, that is $p(w_i) = \frac{c(w_i)}{N}$. Then, for an arbitrary n-gram w_1^n, the expected relative frequency would be the probability:

$$p(w_1^n) = \frac{c(w_1)}{N} \times \frac{c(w_2)}{N} \times \ldots \times \frac{c(w_n)}{N} = \frac{c(w_1) \times c(w_2) \times \ldots \times c(w_n)}{N^n} \tag{3.4}$$

We can scale this probability estimate by the approximate number of n-grams in the corpus ($N - n + 1 \approx N$) to obtain the expected count $E(w_1^n)$:

$$E(w_1^n) = N \times \frac{c(w_1) \times c(w_2) \times \ldots \times c(w_n)}{N^n} = \frac{c(w_1) \times c(w_2) \times \ldots \times c(w_n)}{N^{n-1}} \tag{3.5}$$

Column 6 of Table 3.3 shows the values of $E(w_1^n)$ for the top-30 most frequent 2-grams in BNC-frg. Combinations of frequent words are expected to occur frequently while combinations involving rarer words are expected to occur less. One way to test whether the difference between the expected count $E(\cdot)$ and the observed count $c(\cdot)$ is statistically significant is to use a hypothesis test. Our null hypothesis H_0 is that an n-gram will occur as many times as we expect it to occur, and our alternative hypothesis H_1 is that the observed number of occurrences $c(w_1^n)$ is actually greater than the expected number of occurrences:

$$H_0 : c(w_1^n) = E(w_1^n)$$
$$H_1 : c(w_1^n) > E(w_1^n)$$

Our goal is to reject this null hypothesis using a one-tailed test. If we can reject H_0 at some significance level, this means that the number of occurrences of the n-gram is above what we expected, and we have good chances of finding a MWE. In theory, we should perform an exact test, using the binomial or Poisson test statistics,[13] that model the discrete distribution of n-gram counts (Evert 2004). For instance, a binomomial test statistic uses the binomial distribution to estimate the probability of observing $c(w_1^n)$ or more occurrences of an n-gram given its expected probability $\frac{E(w_1^n)}{N}$:

$$p(X > c(w_1^n)) = \sum_{k=c(w_1^n)}^{N} \binom{N}{k} \left(\frac{E(w_1^n)}{N}\right)^k \left(1 - \frac{E(w_1^n)}{N}\right)^{N-k} \tag{3.6}$$

In practice, however, this test statistic is computationally costly. The number of terms in the sum, $N - c(w_1^n) + 1$ is prohibitive in most cases because N is much larger than $c(w_1^n)$. Moreover, for large values of N, the binomial distribution can be approximated by a normal distribution. As a consequence, it is possible use a z test statistic instead of the exact binomial test statistic: $p(X > c(w_1^n)) = \frac{c(w_1^n) - E(w_1^n)}{E(w_1^n)}$.

A very common test statistic employed in MWE acquisition is *Student's t test* statistic, a heuristic variation of the z test statistic in which the variance of the sample is estimated through its observed count $c(w_1^n)$ rather than from the expected count $E(w_1^n)$. This approximation holds if we consider the corpus as a sequence of randomly generated n-grams and a Bernoulli trial that assigns 1 to the occurrence of w_1^n and 0 otherwise. Then, the probability p of generating 1 is the mean of the sample, $\bar{x} = p = \frac{c(w_1^n)}{N}$. For small values of p, $s^2 = p \times (1 - p) \approx p$, thus the variance of the sample s^2 is equivalent to the mean $\bar{x}$. Finally, the estimated theoretical mean μ is the normalised estimated count $\frac{E(w_1^n)}{N}$, thus yielding the following formulation for the t test statistic:

[13] The *test statistic* is a random variable with a known distribution, from which we can obtain the p-value. If the p-value is below a certain significance level, we can reject the null hypothesis.

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{\frac{c(w_1^n)}{N} - \frac{E(w_1^n)}{N}}{\sqrt{\frac{c(w_1^n)}{N^2}}} = \frac{c(w_1^n) - E(w_1^n)}{\sqrt{c(w_1^n)}} \tag{3.7}$$

As we have seen in Sect. 3.1.2, word counts do not follow a normal distribution, but they can be modelled using a power law distribution, and the same applies to n-grams. As a consequence, from a theoretical perspective, the use of Student's t test statistic here does not make sense as it assumes that the $c(w_1^n)$ follows a normal distribution. Nonetheless, most of the time in MWE acquisition, our goal is to rank candidate n-grams according to their association strength. Thus, the value of the t test statistic is not used to calculate the p-value, but is used directly as a ranking criterion. This ranking criterion is called the `t-score`, and it is interpreted as follows: a large value means strong word association and thus a potential MWE, a small value means that n-gram is more likely to be a random word combination, thus uninteresting for MWE acquisition. Notice that, for the examples in Table 3.3, the statistic is larger when the combination is composed of rarer words.

3.1.4.2 Measures Based on Pointwise Comparisons

The `t-score` is an example of lexical *association measure* (AM), that is, a numerical score that estimates the degree of dependence or association strength between the number of occurrences of the n-gram and the number of occurrences of the individual words that compose it. Similarly to n-gram counts, when more than one corpus is involved, we will subscribe the name of the association measure with the name of the corpus from which the counts used to calculate it were obtained, like in $\texttt{t-score}_{\text{BNC-frg}}$. In addition to the `t-score`, there are many other proposed measures in the literature, not necessarily based on hypothesis testing. Church and Hanks (1990), for instance, suggest to use *pointwise mutual information* (`pmi`), a notion coming from information theory which estimates the predictability of a word given the preceding words. In other terms, `pmi` quantifies the discrepancy between the observed count and the expected count:

$$\texttt{pmi} = \log_2 \frac{c(w_1^n)}{E(w_1^n)} \tag{3.8}$$

This AM has its maximal value when the two words only occur together, that is $c(w_1) = \ldots = c(w_n) = c(w_1^n)$. One disadvantage of `pmi` is that there is no lower bound for its value, since $\log 0$ is undefined. Another disadvantage is that this score requires hard frequency thresholds, as it overestimates the importance of rare n-grams. For instance, the maximal value can be obtained for any n-gram with $c(w_1) = \ldots = c(w_n) = c(w_1^n) = 1$.

Another commonly employed AM is Dice's coefficient, a classical score used in information retrieval to calculate the similarity between two sets. The original

version of Dice's coefficient divides the size of the intersection between n sets (scaled by n) by the sum of their individual sizes. In the case of n-grams, we adapt this measure to compare the n-gram count (intersection between all words) with the sum of the counts of the individual words contained in it:

$$\texttt{dice} = \frac{n \times c(w_1^n)}{\sum_{i=1}^{n} c(w_i)} \tag{3.9}$$

These scores are applicable to arbitrary-length n-grams, but they only take into account the observed number of occurrences $c(w_1^n)$. In spite of their simplicity, they are quite popular in MWE acquisition and in other NLP tasks.

3.1.4.3 Measures Based on Contingency Tables

Nonetheless, more robust and theoretically sound AMs exist for the special case of 2-grams. These measures are based on *contingency tables*, that is, a representation like the one showed in Table 3.4, in which we consider the occurrence of two words as two random variables. We denote as $\neg w_i$ the occurrence of any word different from w_i. Notice that all the cell values are derived from the count of the 2-gram $c(w_1 w_2)$, the individual word counts $c(w_1)$, $c(w_2)$ and the total number of tokens in the corpus N. The values in the last row represent the sum of the values of the inner cells, and analogously for the last column. These are often called *marginal counts* because they are written in the margins of the contingency table. The value of the cell in the last row and column corresponds the number of elements in the sample N, and is equivalent to the sum of the marginal counts in both directions.

For every cell in the contingency table, it is possible to calculate the equivalent expected value if the occurrences of the two words were independent events, as follows:

$$\forall w_i \in \{w_1, \neg w_1\}, \forall w_j \in \{w_2, \neg w_2\}, E(w_i w_j) = \frac{c(w_i) \times c(w_j)}{N} \tag{3.10}$$

Table 3.4 Contingency table for two random variables: the occurrence of the first word w_1 and the occurrence of the second word w_2, $\neg w_i$ denotes the occurrence of any word except w_i

	w_2	$\neg w_2$	
w_1	$c(w_1 w_2)$	$c(w_1 \neg w_2)$ $= c(w_1) - c(w_1 w_2)$	$c(w_1)$
$\neg w_1$	$c(\neg w_1 w_2)$ $= c(w_2) - c(w_1 w_2)$	$c(\neg w_1 \neg w_2)$ $= N - c(w_1) - c(w_2) + c(w_1 w_2)$	$c(\neg w_1)$ $= N - c(w_1)$
	$c(w_2)$	$c(\neg w_2)$ $= N - c(w_2)$	N

We can employ the X^2 test statistic in order to estimate whether the difference between observed and expected contingency tables is statistically significant, as we did for the count $c(w_1^n)$. The X^2 test statistic is a scaled mean squared error measure between observed and expected cell values. That is, for all values of $w_i \in \{w_1, \neg w_1\}$ and $w_j \in \{w_2, \neg w_2\}$,

$$X^2 = \sum_{w_i, w_j} \frac{[c(w_i w_j) - E(w_i w_j)]^2}{E(w_i w_j)} \tag{3.11}$$

The X^2 test statistic for two random variables has an asymptotic χ^2 distribution with one degree of freedom. Thus, it is possible to obtain the p-value which, if sufficiently small, indicates a significant difference between the tables. However, as for the t test, usually the test statistic is considered by itself as a ranking criterion.

A very popular AM based on contingency tables is the *log-likelihood ratio* (`ll`). It was proposed for the first time for MWE acquisition by Dunning (1993). This measure is preferable over X^2 because, for small samples with LNRE distributions, it provides more accurate association estimators, as demonstrated through numerical simulation by Dunning (1993). The simplified version of the `ll` AM is:

$$\texttt{ll} = 2 \times \sum_{w_i, w_j} c(w_i w_j) \times \log \frac{c(w_i w_j)}{E(w_i w_j)} \tag{3.12}$$

This measure has the advantage that, in addition to being theoretically sound, numerically simple and robust to low frequencies, it has a simple interpretation. Its value equals the number of times the 2-gram is more likely under the hypothesis that the words are not independent than the individual counts would suggest. While on the one hand `ll` is robust and theoretically sound, on the other hand it is only applicable to the case where $n = 2$. Extensions to larger n-grams, although possible, are far from being intuitive (see the documentation of the NSP package, described in Sect. 3.2.3.1, for an example).

3.1.4.4 Other Measures

There are numerous AMs available for MWE acquisition. Pecina (2008) presents a table containing 84 measures among which some are rank-equivalent to each other. Hoang et al. (2009) propose and evaluate the adaptation of traditional AMs for word pairs in which one word is very frequent and the other is rather rare, like it is the case for English phrasal verbs formed by rare verbs (e.g., *nail*) with frequent prepositions and adverbs (e.g., *down*).

Table 3.5 shows the top-15 n-grams acquired from BNC-frg as ranked by some of the AMs presented here. A threshold of at least 3 occurrences was set to reduce noise. The first measure, the `t-score`, seems to retrieve rather long specialised MWEs like proper names (*Unix System Laboratories Inc*) and terminological

Table 3.5 Top-15 *n*-grams (2–5) extracted from BNC-frg and ranked according to AMs

`t-score`	`pmi`	`dice`	`ll`
Net earnings per share amounted	of the	CHANCERY DIVISION	of the
reported first quarter net profit	in the	homoclinic orbits	in the
Microsoft Corp 's Windows NT	, and	Los Angeles	, but
(7) mm Hg	to be	Yours sincerely	to be
or fume or other impurity	, but	Greenhouse Effect	I 'm
earnings per share amounted to	on the	Hong Kong	have been
7) mm Hg in	for the	gon na	do n't
dust or fume or other	. '	inter alia	, and
has reported first quarter net	to the	Khmer Rouge	will be
[CHANCERY DIVISION]	at the	Inland Revenue	the the
Inc has reported first quarter	by the	Sri Lanka	per cent
; [1991] 2	from the	Cruz Operation	, ,
N C V O	it is	per cent	has been
Unix System Laboratories Inc	will be	Molecular biology	on the
you 're gon na get	it's	Winston Churchill	the .

phraseology (*reported first quarter net profit*). The list illustrates one of the problems with *n*-gram based methods: the extraction of nested expressions, that is, a shorter expression like *first quarter* contained in a larger one like *first quarter net profit*. Delimiting the borders of a MWE is a current challenge in acquisition tools and methods.

The `dice` coefficient, on the other hand, retrieves shorter *n*-grams among which we find many MWE types like proper names (*Sri Lanka*, *Winston Churchill*), noun compounds (*Greenhouse Effect*, *molecular biology*), formulaic sequences (*Yours sincerely*) and fixed expressions (*per cent*, *inter alia*). Both `t-score` and `dice` tend to retrieve rarer sequences, which only occur three to four times in the corpus.

The other two measures seem to fail in extracting any interesting MWE, as they give much weight to frequent combinations of function words. The `ll` score retrieves some cases of rare double commas or double *the* determiners. Most of the applications of `pmi` and `ll` in the literature are targeted, as these AMs are used to classify possible collocates for a given fixed word, and not to blindly acquire unknown MWEs from a corpus (Dunning 1993; Church and Hanks 1990). The unfortunate reality in AMs for MWE acquisition is that sometimes the most theoretically sound measures perform worse than intuitive heuristics.

This example is an illustration of how AMs work and shows that their results are complementary, suggesting that their combination should be envisaged for broad coverage acquisition. Although there is some published work on fair comparisons among AMs (Pearce 2002; Evert and Krenn 2005; Wermter and Hahn 2006; Schone and Jurafsky 2001), this falls out of the scope of our work and is not the goal of our example. Moreover, the measures have different weaknesses: some overestimate the importance of rare *n*-grams while others are not capable of dealing with frequent

items. Thus, different count thresholds should be applied for each AM, specially for such a small corpus as the BNC-frg. In addition, further cleaning of function words and punctuation is an easy step that should be performed in any case.

Besides association measures, there are also other types of statistical measures that can be used as evidence for MWE discovery in corpora. Pecina (2005), for example, discovered that context measures that consider the adjacent words of the n-grams are more adequate to acquire idiomatic expressions. In terminology acquisition, contrastive measures like C-NC and csMWE are employed as a way of verifying the pertinence of the n-gram to the target domain (Frantzi et al. 2000; Bonin et al. 2010a). These measures estimate the difference between the occurrences of an n-gram in a specialised corpus and in a generic corpus (also called the contrastive corpus). They are recommended to identify terms that occur frequently in a specialised corpus but not in general language (see Sect. 5.1.2.4)

For further material on AMs, please refer to Evert (2004), Seretan (2011), and Pecina (2008). A summary of common association measures can also be found on Stefan Evert's website http://www.collocations.de/.

3.2 Methods for Automatic MWE Acquisition

The tasks involved in the computational treatment of MWEs have been structured by the organisers of the 2009 MWE workshop (Anastasiou et al. 2009) as follows.

- **Identification (or acquisition).** Given a text as input, try to locate the interesting multiword units in it.
- **Interpretation.** Given a multiword unit out of context, try to discover its internal structure both in terms of syntactic and semantic relations.
- **Disambiguation.** Given a multiword unit in its context, try to classify it with respect to a closed set of categories. Typically, one tries to distinguish literal from idiomatic uses, but other disambiguation tasks are possible, for instance, distinguishing general-purpose from specialised uses and performing multiword sense disambiguation.
- **Application.** Given a lexicon of MWEs, try to integrate it in another application such as parsing, information retrieval or MT.

Interpretation and disambiguation are similar as both can be modelled as classification tasks. However, they are distinct as the former concerns MWE types whereas the latter deals with MWE tokens as they occur in text. In addition to the four topics above, we consider an additional task which lies between disambiguation and application, representation:

- **Representation.** Given a lexicon containing MWEs (automatically or manually acquired), try to optimise their representation in a given formalism considering their properties and the target application.

As pointed out in the call for papers of the MWE 2009 workshop[14]:

> The above topics largely overlap. For example, identification can require disambiguating between literal and idiomatic uses since MWEs are typically required to be non-compositional by definition. Similarly, interpreting three-word noun compounds like *morning flight ticket* and *plastic water bottle* requires disambiguation between a left and a right syntactic structure, while interpreting two-word compounds like *English teacher* requires disambiguating between (a) 'teacher who teaches English' and (b) 'teacher coming from England (who could teach any subject, e.g., math)'.

As a large part of the research developed and presented in this book focuses on the first task, the present section is entirely dedicated to MWE acquisition. We start with a summary of related work on monolingual acquisition in Sect. 3.2.1, and on multilingual acquisition in Sect. 3.2.2. Then, we present a more practical description of tools that perform automatic acquisition, distinguishing between those freely available developed by academic researchers and those which were developed and commercialised by companies.

3.2.1 Monolingual Methods

In this section, we discuss some relevant papers on monolingual MWE acquisition methods. The references discussed here are complemented by the work that has been developed for other MWE tasks (Sect. 3.3).

One of the goals of monolingual MWE acquisition techniques is to help and speed up the creation of lexical resources such as printed or machine-readable dictionaries and thesauri containing multiword entries. We distinguish two types of acquisition tasks:

- *MWE extraction*. The input is a text and the expected output is a list of MWE candidates found in the text. The evaluation can be done on a type basis, as if each expression was an entry of a lexicon, independently of the input corpus. In extraction, it is usual to consider two separate steps: (a) *candidate extraction* and (b) *candidate filtering and/or ranking*. We consider that an *MWE candidate* is a sequence of words which has some of the characteristics described in Sect. 2.3 as measured by some objective measure, but that was not yet validated by a manual or automatic evaluation process.
- In *MWE identification*. The input is a text and the expected output is a mark-up indicating the places where MWEs occur. This may include the use of an existing dictionary or the discovery of new MWEs. What makes MWE identification more difficult than simple regular-expression matching is non-adjacency, morphological inflection and ambiguity of some MWEs that can be used both as compositional and idiomatic sequences (e.g., *look up* as consult

[14] http://multiword.sourceforge.net/mwe2009

a dictionary or as staring towards a higher position). In MWE identification, a token-based evaluation is required, taking into account the context in which the expression occurs.

Candidate extraction methods are usually based on some kind of pattern matching, where the patterns range from simple n-grams to structured sequences of part-of-speech tags and syntactic relations. The level of linguistic information employed in candidate extraction depends on various factors such as the language, the type and syntactic variability of the target MWEs and the available analysis tools.

The use of surface forms alone is rare, as generally at least minimal patterns based on stopwords or POS are employed (Gurrutxaga and Alegria 2011). However, there might be cases where flat n-gram extraction is required, for instance, when the target MWEs are generic keyphrases for document description and indexation (Silva and Lopes 2010). The sliding window method consists of considering as MWE candidates pairs that co-occur in a window of at most w words, thus retrieving discontiguous n-grams (Smadja 1993). The extraction of candidates using sliding windows can pose a challenge in terms of computational performance. Indeed, optimised data structures and algorithms must be used because the number of possible combinations, even for relatively small sizes of n, explodes with the size of the corpus (Gil and Dias 2003).

Part of speech sequences are one of the major approaches in candidate extraction because (i) many languages have available push-button POS taggers and (ii) this approach provides good results when the target constructions are relatively rigid in terms of word order, like fixed phrases and nominal MWEs. POS sequences have been used originally in multiword terminology acquisition (Justeson and Katz 1995; Daille 2003), but have also been applied to the extraction of other MWE types, specially noun compounds (Vincze et al. 2011). Even when dealing with more variable constructions such as verbal expressions, POS tag patterns can be used in the absence of syntactic information (Baldwin 2005; Duran et al. 2011). POS patterns can be defined based on various criteria, from linguistic intuition and expert knowledge (Bonin et al. 2010b) to systematic empirical observation of a sample (Duran et al. 2011). Sequences of POS can also be automatically learnt from annotated corpora, using the same methodology as for words, that is, by maximising some AM on the extracted POS n-grams (Dias 2003).

When a parser is available, patterns based on syntactic relations can be used for candidate extraction and/or identification. For example, one may retrieve all candidates that are formed by a noun which is the direct object of a verb (*take/V* $\leftarrow_{\text{DOBJ}}$ *time/N*). According to the accuracy of the parser, simple syntactic patterns can be more precise than POS sequences, specially in the extraction of non-fixed MWEs like "true" collocations (Seretan 2008; Seretan and Wehrli 2009; Seretan 2011). Lexicons containing more or less refined MWE representations can be used to identify MWEs during parsing, with the advantage of disambiguating instances in context (Bejček et al. 2013; Constant et al. 2013). Tree substitution grammars can also be used in order to learn syntactic MWE models from annotated corpora,

as in the French version of the Stanford parser (Green et al. 2011). Regardless of the syntactic information and labels, structural regularities in parsing trees can also be used to retrieve MWE candidates using a minimal description length algorithm (Martens and Vandeghinste 2010). Often, these identification techniques require an annotated corpus with MWE markup. While some MWE types may be represented in existing treebanks, in most cases an additional annotation layer must be added (Laporte et al. 2008; Uresova et al. 2013).

In addition to analysed corpora, other monolingual and multilingual resources can be used for MWE acquisition. For instance, by comparing the titles of Wikipedia pages using cross-language links, it is possible to detect multiword titles whose translation in one of the other languages is a single word (Attia et al. 2010). Another way to use the web as a source of information for MWE acquisition is to generate candidates according to generic combination rules and further validate them using web search engine hit counts (Villavicencio et al. 2005). This is explored in our experiments in Sect. 6.2.3.2. The current trend is the integration of several complementary information sources (including linguistic analysis, statistics, the web) in order to maximise the recall of the extraction (de Medeiros Caseli et al. 2009; Attia et al. 2010).

More complex candidate extraction methods, not based on pattern matching, have also been proposed. The LocalMaxs algorithm, for instance, performs extraction based on the maximisation of an AM applied to adjacent word pairs. Thus, it naturally handles nested expressions, extracting maximal sequences that recursively include adjacent words while the overall AM score increases (Silva and Lopes 1999). Similarly, a tightness measure is used in a Chinese IR system for the automatic identification, concatenation and optimised querying of strongly associated word sequences (Xu et al. 2010). Duan et al. (2006) propose a string matching algorithm, inspired by computational biology, to extract sequences that occur recurrently throughout the corpus. Sentences are viewed as DNA sequences and a dynamic programming algorithm is used to match corresponding parts for each sentence pair in the corpus, taking into account gaps that represent variable parts of the expression. These techniques generally do not distinguish candidate extraction from filtering, performing both simultaneously.

As for candidate filtering in MWE extraction, some straightforward procedures are the use of stopword lists and of count thresholds to remove candidates for which statistical information is insufficient. Lexical association measures like those described in Sect. 3.1.4 are also widely employed to rank the candidates and keep only those whose association score is above a certain threshold (Evert and Krenn 2005; Pecina 2005). When several AMs are available, they can be combined using machine learning, possibly considering additional information coming from auxiliary resources. Thus, it is necessary to annotate part of the MWE lists to obtain an annotated extraction dataset (Grégoire et al. 2008). Then, a supervised learning method can be used to build a classifier modelling the optimal weights of all the AMs and extra features (Ramisch et al. 2008; Pecina 2008).

There is a strong predominance of methods based on 2-grams (or more generally on word pairs, not necessarily adjacent) in current techniques for monolingual

MWE acquisition. This is justified because (i) the majority of the interesting and challenging MWEs are formed by two words and (ii) "experiments with longer expressions would require processing of much larger amount of data and [there is a] limited scalability of some methods to [handle] high order n-grams" (Pecina 2005). While this seems like a reasonable justification to keep the methodology simple, it does not correspond to the reality of NLP applications, where many MWEs longer than two words also require proper treatment (Nakov and Hearst 2005; Kim and Nakov 2011).

Monolingual methods have been developed in several languages and are sometimes language independent. The advantage of language-independent methods is that they do not depend on the availability of a specific resource (POS tagger, parser) and can thus be applied to virtually any language, including poorly resourced ones. On the other hand, the use of linguistic information generally improves the precision and the coverage of the acquisition. Finding an adequate trade-off between language independence and quality when designing a method for monolingual acquisition is a challenging problem. However, as MWEs seem to be a universal phenomenon, being present in all human languages, it is important to build methods and evaluate them in multilingual contexts (Seretan and Wehrli 2006).

3.2.2 Bi- and Multilingual Methods

Even though many of the methods described in the previous section can be applied to arbitrary corpora, independently of the language, they are still considered as monolingual methods because the result is a list of MWEs or a marked up text with no cross-lingual correspondences. The extraction of bilingual MWEs is a task in which the resulting list of expressions is bilingual, that is, if a candidate is returned in one language, it contains translation links which relate it to its correspondent candidate(s) in the other language(s). Hence, bi- and multilingual MWE acquisition is different from language-independent MWE acquisition. Existing techniques for bilingual MWE acquisition are frequently based on parallel corpora.

Automatic word alignment can provide lists of MWE candidates by themselves, as described in de Medeiros Caseli et al. (2010). They aligned a Portuguese–English corpus in both directions using GIZA++, and then joined the alignments using the grow-diag-final heuristic. Word sequences of two or more words on the source side aligned to sequences of one or more words on the target side were filtered using several stopword patterns and the resulting candidates were considered as MWEs. The comparison with a simple monolingual n-gram method showed that alignment-based extraction is much more precise, but has very limited recall. This technique has been further validated for the acquisition of Portuguese multiword terms in technical and scientific domains (Ramisch et al. 2010).

Tsvetkov and Wintner (2010) extended this method with two main improvements: (a) they consider all non-bijective 1:1 alignments as candidate sequences and (b) they validate their candidates using an association measure, PMI^n, calculated

on a large monolingual corpus. Furthermore, they combine alignment information with other linguistic sources (Tsvetkov and Wintner 2011). They evaluate their method by integrating the automatically extracted bilingual lexicon in a statistical MT system, obtaining small but significant improvements in translation quality.

Bai et al. (2009) present an algorithm capable of mining translations for a given MWE in a parallel aligned corpus. Then, the different translations are ranked according to standard association measures in order to choose the appropriate one. They integrated this extraction method into the statistical MT system Moses for the English–Chinese language pair, obtaining improved translations when compared to a baseline.

The automatic discovery of non-compositional compounds from parallel data has been explored by Melamed (1997). Considering a statistical translation model, he introduced a feature based on mutual information and proposed an iterative algorithm that retrieves an increasing number of compounds. These can in turn be used to improve the quality of the statistical translation system itself.

Conversely, it has been shown that MWEs can improve the quality of automatic word alignment. The English-Hindi language pair presents large word order variation, and it has been shown that MWE-based features that model compositionality can help reduce alignment error rate (Venkatapathy and Joshi 2006). When compared with baseline GIZA++, a system enriched with MWE features obtains significantly lower error rates, from 68.92 to 50.45 %.

The acquisition of bilingual verbal expressions requires not only the availability of parallel corpora, but also of syntactic analysis of both languages. Zarrieß and Kuhn (2009) used syntactically analysed corpora and GIZA++ alignments to extract verb-object pairs from a German–English parallel corpus. They considered a candidate as a true MWE if (i) a verb on the source side was aligned to a verb on the target side, (ii) the noun heading the object of the verb on the source side was tagged as a noun on the target side and (iii) there was a syntactic object relation on the target side between the target verb and the target noun. Their method retrieves 82.1 % of correct translations, and almost 60 % of translations which can be considered as MWEs.

Bouamor et al. (2012) also perform bilingual MWE acquisition using word alignments, but instead of using them as a starting point, they use them a posteriori, after monolingual extraction has been performed on source and target corpora. Therefore, acquiring a bilingual MWE lexicon is seen as an alignment problem between expressions acquired separately in each language.

Instead of relying on large parallel word-aligned corpora, which are not always available for a given language pair, it is possible to use comparable corpora as a source for acquisition. Daille et al. (2004) performed multiword term extraction independently in French and in English using comparable corpora in the environmental domain. Then, using the distances between the context vectors of the acquired terms, they obtained cross-lingual equivalences that were evaluated against a bilingual terminological dictionary. The dictionary reference translation occurred among the top-20 retrieved translations in 47–89 % of the translations, depending on the translation relation type (single word vs multiword).

The acquisition of bilingual MWEs has been explored more often in the context of machine translation. In Sect. 3.3.4.4, we provide an overview of attempts to integrate MWEs into different MT applications. This is further developed in the experiments described in Sect. 7.2.

3.2.3 Existing Tools

The maturity of a research field depends not only on theoretical models and experimental results, but also on concrete tools and available software on the basis of which it is possible to reproduce results, build extensions and perform systematic evaluations. Thus, tools for the automatic acquisition of MWEs are very important for the evolution of this research field. Here, we distinguish two types of tools: those which are freely available for the community (Sect. 3.2.3.1) and those that are either commercial or available in restricted contexts (Sect. 3.2.3.2).

3.2.3.1 Freely Available Tools

To date, the existing research tools follow the main trends in the area, using linguistic analysis and statistical information as clues for finding MWEs in texts. Here, we present a list of freely available tools that can be used mostly for monolingual MWE acquisition.

1. **LocalMaxs**: http://hlt.di.fct.unl.pt/luis/multiwords/
   The "Multiwords" scripts are the reference implementation[15] of the LocalMaxs algorithm. It extracts MWEs by generating all possible n-grams from a sentence and then further filtering them based on the local maxima of a customisable AM's distribution (Silva and Lopes 1999). On the one hand this approach is based purely on word counts and is completely language independent. On the other hand, it is not possible to directly integrate linguistic information in order to target a specific type of construction or to remove noisy ungrammatical candidates.[16] The tool includes a strict version, which prioritises high precision, and a relaxed version, which focuses on high recall. A separate tool is provided to deal with big corpora. A variation of the original algorithm, SENTA, has been proposed to deal with non-contiguous expressions (da Silva et al. 1999). However, it is computationally costly because it is based on the calculation of all possible n-grams in a sentence, which explodes when going from contiguous to

[15] Recommended by the author of the algorithm in personal communication.

[16] Although this can be simulated by concatenating words and POS tags together in order to form a token.

non-contiguous n-grams. Furthermore, there is no freely available implementation.

2. **Text::NSP**: http://search.cpan.org/dist/Text-NSP
   The N-gram Statistics Package (NSP) is a standard tool for the statistical analysis of n-grams in text files developed and maintained since 2003 (Pedersen et al. 2011; Banerjee and Pedersen 2003). It provides Perl scripts for counting n-grams in a text file and calculating AMs, where an n-gram is either a sequence of n contiguous words or n words occurring in a window of $w \geq n$ words in a sentence. While most of the measures are only applicable to 2-grams, some of them are also extended to 3- and 4-grams, notably the log-likelihood measure. The set of available AMs includes robust and theoretically sound measures such as Fischer's exact test. The input to the NSP tool is a corpus and a parameter value fixing the size of the target n-grams. The output is a list of types extracted from the corpus along with the counts, which can further be used to calculate the AMs. Although there is no direct support to linguistic information such as POS, it is possible to simulate them to some extent using the same workaround as for LocalMaxs.[16] The tool allows complex expressions in order to express what counts should be calculated in terms of the sub-n-grams contained in a given n-gram.
3. **UCS**: http://www.collocations.de/software.html
   The `UCS` toolkit provides a large set of sophisticated AMs, in addition to other mathematical procedures like dispersion test, frequency distribution models and evaluation methods. It was developed in Perl and uses the R statistics package. UCS focuses on high accuracy calculations for 2-gram AMs, but, unlike the other approaches, it does not properly perform MWE acquisition. Instead of a corpus, it receives a list of candidates and their respective counts, relying on external tools for corpus preprocessing and candidate extraction. Then, it calculates the measures and ranks the candidates. Therefore, the question about contiguous n-grams or support of linguistic filters is not relevant for UCS.
4. **jMWE**: projects.csail.mit.edu/jmwe
   The jMWE tool (Kulkarni and Finlayson 2011) is aimed at dictionary-based in-context MWE token *identification* in running text, which makes it quite different from *extraction* tools. It is available in the form of a Java library, and expects a corpus as input, possibly annotated with lemmas and parts of speech. In addition, it requires an initial dictionary of valid known MWEs. The system then looks for instances (occurrences) in the corpus of the MWEs included in its internal dictionary. It does not perform any automatic discovery of new expressions, thus the quality of the output heavily depends on the availability of MWE dictionaries. While jMWE is not language independent, it can be configured and straightforwardly adapted to other languages for which a suitable dictionary is available. The system allows quite powerful instance search, similar to multilevel regular expressions. It is possible to deal with non-contiguous expressions and to apply successive filters on the output. jMWE also provides heuristics for disambiguating nested compounds. On the other hand, it is not

possible to express constraints based on syntax, nor to apply AMs in order to remove words that co-occur by chance.

5. **Varro**: http://sourceforge.net/projects/varro/
   This tool is not specifically aimed at MWE acquisition, but rather at finding regularities in treebanks (Martens 2010). It implements an optimised version of the *Apriori* algorithm with many adaptations that allow for the efficient and compact representation of tree structures. Statistical scores based on description length gain have been proposed to rank regular subtrees returned by the tool, thus helping in the acquisition of MWEs (Martens and Vandeghinste 2010). In contrast with the preceding tools, the Varro toolkit is not based on word sequences, but it requires syntactically analysed corpora as input. It is thus well suited for the extraction of flexible expressions such as idioms, formulaic phrases, "true" collocations and verbal expressions.

There are also numerous freely available web services and downloadable tools for automatic term extraction. These tools are generally language dependent, having versions for major European languages like English, Spanish, French and Italian. Although multiword terms are included in our definition of MWE, these tools are not appropriate for general-purpose extraction of expressions in everyday language. Examples of such tools are TermoStat,[17] AntConc[18] and TerMine.[19] The Wikipedia page on terminology extraction[20] lists many other freely available tools.

The methodological framework introduced in the present work has also been implemented in a freely available tool, the `mwetoolkit`.[21] This tool is described in detail in Chap. 5.

3.2.3.2 Commercial Tools

There are numerous commercialised systems for automatic terminology extraction from specialised texts. As a great part of terminology is multiword, this kind of software performs MWE acquisition at some point. Déjean et al. (2002), for example, describe a method developed at Xerox that uses morphosyntactic patterns for monolingual term recognition. Afterwards, they perform automatic alignment and extract English–German terminology, reaching an F-measure of around 80 %. This kind of technique has certainly been integrated into their Xerox Terminology Suite (XTS). This software is not commercialised any more, since it has been

[17] http://olst.ling.umontreal.ca/~drouinp/termostat_web/

[18] http://www.antlab.sci.waseda.ac.jp/software.html

[19] http://www.nactem.ac.uk/software/termine/

[20] http://en.wikipedia.org/wiki/Terminology_extraction

[21] http://mwetoolkit.sourceforge.net

acquired by the text mining company Temis.[22] Nowadays, it has become part of the Luxid® information extraction package.[23]

Another large company which developed a tool for terminology extraction is Yahoo!. Their term extraction service is freely available for research and personal purposes, limited to 5,000 queries per day per IP address.[24] However, this service is limited to short English texts and is probably based on term dictionaries and gazetteers.

The Fips parser, developed at the University of Geneva, has been used for collocation extraction in several languages (Seretan and Wehrli 2009). Even though it is academic research, the collocation extraction tool FipsCo, based on Fips, is not freely available. The tool is able to extract collocations from monolingual corpora in English, French, Spanish and Italian, and there is a version for Greek (Michou and Seretan 2009). The tool has been used in MT experiments, suggesting that it is able to extract bilingual collocations from word-aligned parallel corpora. Although the system itself is not free, its visualisation tool, FipsCoView,[25] is freely available as a web interface (Seretan and Wehrli 2011).

Translation memory software may use MWEs as basic segments to retrieve. Indeed, MWEs are somehow in-between sentences and words. On the one hand, the retrieval of simple words in a hypothetical translation memory would be of little use. The number of possible translations for a word out of its context is potentially large and additional information is required to choose among the options. Therefore, it would lack precision. On the other hand, the retrieval of whole sentences would be highly precise, but an extremely large translation memory would be required in order to obtain reasonable recall. If the memory of previously translated segments is small, only from time to time (and with some luck) a sentence will be retrieved. Many sentences containing part of the translation would be useful, but will be ignored by a sentence-based exact match system.

One example of system performing bilingual MWE extraction is Similis,[26] previously commercialised by Lingua et Machina and now freely available. According to the official website, "Similis [...] includes a linguistic analysis engine, uses chunk technology to break down segments into intelligent terminological groups (chunks), and automatically generates specific glossaries." The technique implemented in the system is an evolution of the one described in Planas and Furuse (2000). In this article, the authors describe a clever technique for retrieving similar segments in the source language and their correspondences in the target language. Their approach applies a dynamic programming algorithm on a multi-layered structure where sentences are represented as a sequence of surface forms, lemmas and parts

[22] http://www.temis.com/

[23] http://www.temis.com/index.php?id=201&selt=1

[24] http://developer.yahoo.com/search/content/V1/termExtraction.html

[25] http://129.194.38.128:81/FipsCoView

[26] http://similis.org/

of speech. The combination of the matchings in these three layers allows for a good balance between precision and recall for the retrieval of bilingual segments.

3.3 Other Tasks Related to MWE Processing

Given that MWE acquisition is our main concern, the whole Sect. 3.2 is dedicated to a detailed review of the state of the art. Here, we overview the state of the art in the other tasks involved in MWE treatment, according to the classification of MWE tasks, namely interpretation (Sect. 3.3.1), disambiguation (Sect. 3.3.2), representation (Sect. 3.3.3) and application (Sect. 3.3.4).

3.3.1 Interpretation

The interpretation and disambiguation of several types of MWEs are the focus of a large body of literature, specially in the computational semantics community. Both can be modelled as classification tasks, so that machine learning algorithms are often employed. Therefore, it is possible to distinguish supervised from unsupervised approaches. In the former, a large effort is usually dedicated to the annotation of a data set that is subsequently used to build classifiers. In the latter, the class attribution is made based on thresholds or rules directly applied to data features. Like for most solutions based on machine learning, supervised methods outperform unsupervised methods. However, unsupervised methods may sometimes perform as well as supervised methods when they are applied on very large corpora like, for instance, web-based corpora (Keller and Lapata 2003).

MWE interpretation can be applied on expressions whose meaning does not change too much according to their occurrence contexts, like compound nouns and some specific types of phrasal verbs and support verb constructions. However, it is not suitable to interpret ambiguous expressions such as phrasal verbs (*look up a word* vs *look up to the sky*) and idioms (*my grandfather kicked the bucket* vs *the cleaning lady accidentally kicked the bucket*). These are explored in MWE disambiguation tasks (see Sect. 3.3.2). Noun compounds (*traffic light*, *nuclear transcription factor*), on the other hand, are rarely ambiguous and their interpretation has been an active research area. We distinguish two types of noun compound interpretation: syntactic and semantic.

The *syntactic interpretation* has been explored by Nicholson and Baldwin (2006), who distinguish three syntactic relations in noun–noun compounds: subject (*product replacement*), direct object (*stress avoidance*) and prepositional object (*side show* → *show on the side*). For compounds in which the second noun is a

nominalisation,[27] they used the inflections of the corresponding verb to generate paraphrases that were looked up in Google. The paraphrases and additional features were fed into a nearest-neighbour classifier, but the results failed to improve over the state of the art.

Three-word or longer noun compounds like *liver cell line* and *liver cell antibody* require syntactic interpretation of the constituent hierarchy. That is, one needs to distinguish left bracketing like in *(liver cell) antibody* from right bracketing like in *liver (cell line)*. Therefore, Nakov and Hearst (2005) compare two models, based on adjacency and on dependency. They use a set of heuristics to generate surface-level paraphrases and then use search engine counts to estimate model probabilities. They obtain sizeable improvements over state of the art on a set of biomedical compounds.

One of the most challenging interpretation problems is the *semantic interpretation* of the relations involved in noun compounds. The goal is to assign to each noun compound one (or several) tags that describe the semantic relation between the two nouns. Nakov and Hearst (2008) try to solve this task using a methodology similar to the one they employed for syntactic interpretation. First, they generate a large number of paraphrases involving verbs related to the semantic classes (e.g., *causes, implies, generates* for relation *CAUSE*) and the relative *that*. Then, they retrieve web counts for the paraphrases and assign the classes with maximal probability according to the corresponding paraphrases. Their method is completely unsupervised. The resource developed in their work, containing noun compounds and corresponding features, is freely available on the MWE community website (Nakov 2008b). Kim and Nakov (2011) revisited the problem, this time using a combination of data bootstrapping and web counts. The main difference is that they generated paraphrases not based on surface forms but on parse trees, thus obtaining more accurate results. A related method is proposed and evaluated by Kim and Baldwin (2013).

Paraphrases can be used not only as means but also as ends. That is, they may be the *actual* representation of semantic classes instead of a set of (somehow arbitrary) abstract tags. The representation of semantic classes for noun–noun relations is discussed in depth by Girju et al. (2005), who compare Lauer's eight prepositional tags with a proposed classification using 35 abstract tags. Moreover, they annotate a corpus sample using both schemes and investigate the correspondences between them. In addition, paraphrases can be used, for instance, in order to artificially generate new data for training MT systems (Nakov 2008a).

Lapata (2002) focuses on the interpretation of noun compounds involving nominalisations. She reformulates noun compound interpretation as a disambiguation task, re-creating missing evidence from corpus. She extracts the counts of the nouns and of the related verb from the BNC, and then uses them as features in a supervised machine learning tool that automatically learns association rules. She

[27] A noun derived from a verb, like *replacement* is a nominalisation of the verb *replace*.

also discusses and evaluates several smoothing techniques that help obtaining more realistic counts. Keller and Lapata (2003) used this task as one of their case studies in order to investigate the use of web counts in NLP disambiguation tasks.

Latent semantic analysis has also been employed for the semantic classification of noun–noun compounds (Baldwin et al. 2003). In order to distinguish compositional from idiomatic constructions, the authors compare the context vectors of the compound with the context vectors of the individual nouns composing it. This approach can be generalised and has also been applied and evaluated on other types of MWEs. A similar technique is proposed by Séaghdha and Copestake (2013). However, they use string kernels to build more or less lexicalised semantic representations for the words in the compound. Then, they use standard composition techniques in order to infer the combined semantics of the compound.

A comprehensive and detailed revision of the semantic interpretation of noun compounds can be found in Nakov's Ph.D. thesis (Nakov 2007) and in his later survey article (Nakov 2013). Several techniques used in this task are described in the proceedings of SemEval 2010 and 2013, which feature shared tasks on this topic (Hendrickx et al. 2010; Butnariu et al. 2010). The Special Issue on noun compounds of the Natural Language Engineering journal (Szpakowicz et al. 2013) presents some advances in this area and includes an article that describes in detail the best Semeval system for noun compound interpretation (Nulty and Costello 2010, 2013).

Besides noun compounds, other MWE types also require interpretation. English phrasal verbs are ambiguous and can be used both idiomatically (*look up a word*) and literally (*look up to the sky*). However, if we consider only the most usual sense, it is possible to perform type-based interpretation. Cook and Stevenson (2006) use support vector machines to classify the meaning of the particle *up* in English phrasal verbs. According to the verb, it can have a sense of vertical, completion, goal or reflexive. These are simplified using a 2-way and a 3-way classification. The features used are standard syntactic slots of the verb, particle characteristics such as distance from the verb, and word co-occurrences.

Considering a larger range of constructions, Bannard (2005) investigates the extent to which the components of a phrasal verb contribute their meanings to the interpretation of the whole. He models compositionality through an entailment task, for instance, *split up* $\Longrightarrow$ *split*? In a comparison between `pmi`, `t-score` and a newly proposed measure based on context cosine similarity, the latter correlates better with human judgements.

A similar work is that of McCarthy et al. (2003). They propose several measures involving an automatically acquired distributional thesaurus in order to estimate the idiomaticity of phrasal verbs. Their annotated data set uses a numeric scale from 0 (totally opaque) to 10 (fully compositional). They show that the best association measure, mutual information, is less correlated to human judgements than a proposed measure which calculates the number of neighbours with the same particle as the phrasal verb minus the equivalent number of simple neighbours. A similar annotation scale has been proposed by Roller et al. (2013), who provide a data set with compositionality assessments for German expressions. They show that

it is important to perform outlier identification and removal to obtain reliable data. In general, interpretation results based on human annotation of compositionality must always be carefully analysed, since this is a hard task even for humans.

Venkatapathy and Joshi (2006) explore the type compositionality of verb–noun pairs. They describe the creation of an annotated data set with compositionality judgements ranging from 1 to 6. Then, they present seven distinct features to estimate compositionality which are further combined using a support vector machine. They evaluate the features separately and show that the Spearman correlation between the classifier results and human judgements is around 0.448, which is better than all individual features.

Using a variation of the same data, McCarthy et al. (2007) investigate the use of selectional preferences in this task. They propose three different algorithms to obtain this information from parsed corpora: two based on Wordnet and one based on an automatically constructed thesaurus. They show that the best performance is obtained by combining selectional preferences and a subset of Venkatapathy and Joshi's features through a support vector machine.

3.3.2 Disambiguation

Recall that the disambiguation of MWEs is analogous to their interpretation, except that they are considered together with the context in which they appear (sentences). Nicholson and Baldwin (2008) present a data set for noun–noun compound disambiguation where a large set of sentences has been manually annotated with syntactic and semantic information about the compounds contained in it. Girju et al. (2005) investigate methods for their disambiguation. They perform a separate analysis of two- and three-noun compounds, annotating their semantics according to two tagging schemes in a training set of around 3K sentences. In addition to a detailed analysis of the coverage and correspondences between the tagging schemes, they apply several supervised learning techniques. Like for the syntactic disambiguation of three-word compounds, they also employ classifiers. They achieves an accuracy of 83.10 % by using as features (a) the top three WordNet synsets for each noun, (b) derivationally related forms and (c) a flag telling whether the noun is a nominalisation.

Whereas, for MWE interpretation, the majority of publications concerns noun compounds, when it comes to disambiguation a large number of MWE types has been studied. However, English still predominates. One of the rare works on a language different from English concerns the interpretation of German preposition–noun–verb triples (Fritzinger et al. 2010). Constructions like *in Gang kommen* have both a literal interpretation as *to reach the hallway (in den Gang kommen)*, but also idiomatic interpretations as *to be set in motion (in Gang kommen)* and *to get organised (in die Gänge kommen)*. They manually analysed a large set of such constructions retrieved by a parser, classifying them as either literal, compositional

or unknown.[28] Then, they investigated the correlation between these classes and morphosyntactic characteristics such as determiners, plural and passivisation.

Light/support verbs in Japanese have also been studied in the past. They include sequences like *donari-ageru* (*shout*) and *odosi-ageru* (*threaten*), that is, formed by two lexical units where the verb is usually highly polysemous like *ageru* (*raise*). Uchiyama et al. (2005) propose two disambiguation methods: a statistical approach using a sense inventory, context and a support vector machine; and a rule-based method where the rules were manually defined based on syntax and on the semantics of the first verb. The rule-based method (94.6 %) outperforms the statistical method (82.6 %) in terms of accuracy, but the latter obtains a surprisingly high performance given its simplicity.

The interpretation of expressions of the type verb–noun has also been explored in English. Cook et al. (2007) explore the idiomaticity of verb–noun pairs, where the noun is the direct object of the verb and may have an idiomatic (*make a face*) or literal (*make a cake*) interpretation. Their basic hypothesis is that idiomatic uses are syntactically more rigid. Thus, they describe a fully unsupervised approach which considers syntactic and context information in order to calculate the similarity with the canonical form of the idiom. In their evaluation, they report results comparable to a supervised approach. The data set used in their experiments is freely available (Cook et al. 2008).

Fazly and Stevenson (2007) propose a more fine-grained classification for light verb–noun constructions. They use a supervised learning strategy based on decision trees in order to perform a 4-way semantic disambiguation. In their scheme, a light verb may be used with its literal meaning (*make a cake*), with its abstract meaning (*make a living*), in light-verb constructions (*make a decision*) or idiomatically (*make a face*). These classes are a mix of syntactic and semantic characteristics and could arguably be improved using more systematic criteria. Even though they perform type-based annotation of their data sets, this work can be considered as disambiguation because the noun is the context used to disambiguate the semantics of a closed set of polysemous light verbs. Considering a random baseline with 25 % accuracy, they obtain an overall accuracy of 58.3 %. F-measure varies according to the classe: abstract constructions are harder to classify (46 %) than light verb constructions (68 %).

3.3.3 Representation

The lexical representation of MWEs was one of the main goals of the MWE project at Stanford, and has for a long time haunted lexicographers in the compilation of lexical resources. Most NLP applications contain at least a small amount of

[28]The context unit used for annotation was the sentence. However, due to anaphora, sometimes it was impossible to know the intended meaning without looking at neighbour sentences.

MWE entries, specially closed-class expressions. The Stanford parser, for instance, contains a list of several dozens of 2-word and 3-word conjunctions. However, when it comes to open-class expressions, this coverage is too limited and ways to efficiently represent MWEs in computational lexicons are required. Sag et al. (2002) proposed two approaches: words-with-spaces and compositional. However, between these extremes of the compositionality spectrum, there are some other possibilities, sometimes explored in related work. Laporte and Voyatzi (2008), for instance, describe a dictionary containing 6,800 French adverbial expressions like *de nos jours* (*nowadays*). A set of 15 flat sequences of parts of speech is used to describe the morphosyntactic pattern of each entry using the lexicon–grammar format.

Graliński et al. (2010) present a qualitative and quantitative comparison between two structured representations for Polish MWEs: Multiflex and POLENG. While the former is designed to be generic and language independent, the latter has a more implicit structure aimed at specific applications. The authors focus on nominal compounds and analyse the power of each formalism to incorporate morphological inflection rules such as case, gender and number agreement. They also measure the time taken by one expert and two novice lexicographers to encode new MWEs. Multiflex does not allow the description of non-contiguous units nor units containing slots and it takes much longer for lexicographers to learn and use it. POLENG offers a complementary approach, allowing a faster description of MWEs including non-contiguous ones.

A more corpus-based representation has been proposed for the representation of entries in the Dutch electronic lexicon of MWEs (Grégoire 2007, 2010). She uses an equivalence class method that groups similar expressions according to their syntactic characteristics. In addition to numbers of occurrences and examples, each entry contains a link to a pattern that describes the syntactic behaviour of the expression. This description is quite practical, as the lexicon is aimed for NLP systems such as the Alpino parser and Rosetta MT system.

Izumi et al. (2010) suggest a rule-based method to normalise Japanese functional expressions in order to optimise their representation. They consider two separate problems: the insertion of omitted parts and the removal of satellite parts that do not contribute much to the meaning of the sentence. In a comparison with manually generated paraphrases, they obtain a precision of 77 %. If such normalised representation are adopted in the lexicon, the same paraphrasing rules can be applied to running text in order to align it with expressions contained in the lexicon.

The use of tree-rewriting grammars for describing MWEs is proposed by Schuler and Joshi (2011). They provide arguments and formal proof that this formalism is adequate to represent non-fixed expressions such as *raise X to the Yth power*. The generalisation of their approach to other types of expressions, however, remains to be demonstrated.

Finally, concerning the hierarchical structure among MWEs, SanJuan et al. (2005) explore three strategies (lexical inclusion, Wordnet similarity and clustering) to organise a set of multiword terms manually extracted from the Genia corpus. This

kind of representation can be very useful to include extracted expressions in more sophisticated concept nets and ontologies.

When it comes to bilingual and multilingual dictionaries, the problem becomes more complex since it is necessary to represent not only the internal structure of the entries but also cross-language links at global and local levels. To the best of our knowledge, there is little research concerning this problem. Section 3.3.4.4 contains a discussion on the representation of MWEs in MT systems.

In short, due to the modest amount of research in this area and to the complexity of the problem, a model for the efficient lexical representation of MWEs in general remains an open problem.

3.3.4 Applications

A list of potential NLP applications where MWEs are relevant was introduced in Sect. 1.1.2. Here, we provide a summary of these target applications for which concrete results have been obtained. Many results presented here concern pilot studies and techniques as simple as joining contiguous MWE components with an underscore character as a preprocessing step. From all the MWE tasks discussed in this section, application is by far the one with the least amount of published results.

3.3.4.1 Syntactic Analysis

A small set of fixed MWEs like conjunctions are represented in most existing parsers, chunkers and POS taggers. However, the further insertion of additional multiword entries can improve the coverage of the analysis, as more complex MWEs like noun compounds and verbal expressions are valuable information for syntactic disambiguation.

Concerning POS tagging, Constant and Sigogne (2011) present an evaluation on French. They assign special tags to words corresponding to the beginning and to the ending of multiword units. Using a model based on conditional random fields, they learn the MWE-aware tagger from a corpus in which the training data was automatically annotated with entries coming from several lexica containing compounds and proper nouns. This technique obtains 97.7 % accuracy, improving considerably over standard taggers like TreeTagger and TnT. They also compare this strategy with the integration of the same lexicons into a parser trained using a probabilistic context-free grammar with latent annotations (PCFG-LA). Their results show that MWE recognition and parsing quality improve with both approaches (Constant et al. 2013).

Korkontzelos and Manandhar (2010) obtain impressing improvements by enriching a baseline shallow parser with MWEs. Their method simply consists of joining contiguous nominal expressions with an underscore prior to parsing. This makes the system treat them as unknown tokens and assign them a parse based on the context.

They analyse a set of 118 2-word MWEs from WordNet, classifying them by POS sequences and by compositionality. They conclude that, in all cases, the accuracy of the parses was improved, specially for non-compositional adjective–noun pairs, for which the substantial improvements ranged from 15.32 to 19.67 %.

As for deep parsing, Zhang and Kordoni (2006) extended the lexicon of an English HPSG parser with 373 MWE entries represented as words-with-spaces. They obtained a significant increase in the coverage of the grammar, from 4.3 to 18.7 %. Using a compositional representation, Villavicencio et al. (2007) added 21 new MWEs to the same parser, obtaining an increase in the grammar coverage from 7.1 to 22.7 %, without degrading accuracy. However, MWEs do not always improve the performance of the parser, as shown by Hogan et al. (2011). They try to include a set of named entities in their parsing system, replacing them by placeholders. However, they did not obtain significant improvements over the baseline, even when tuning count thresholds.

The use of lexical association measures to model collocational behaviour in dependency parsing has been investigated by Mirroshandel et al. (2012). They enrich the dependency relations in the parsing model, learnt from the treebank, with pairs collected from a very large parsed corpus. Their results show significant improvements in parsing accuracy for specific syntactic relations.

As far as we know, the English and Italian parser Fips is one of the few systems dealing with variable MWEs (Wehrli et al. 2010). Its approach is more sophisticated than words-with-spaces, as it dynamically identifies expressions at the same time as it constructs the parse tree. This technique performs better than post-processing the trees after they are produced. The authors demonstrate that MWEs are not a "pain in the neck" but actually a valuable information to reduce syntactic ambiguity.

3.3.4.2 Word Sense Disambiguation

Given an occurrence of a polysemous word, word sense disambiguation consists of picking up a single sense among those listed in an inventory. For example, the verb *fire* can mean *make somebody lose his/her job*, *pull the trigger of a gun*, or *make something burn*. The sentence in which the verb occurs will determine which of these senses is intended. Although context information is used, MWEs are generally ignored in WSD tasks. As a consequence, not only the correct sense will be ignored but also wrong senses will be inferred for the individual words. For example, in Wordnet, none of the senses of *voice* and of *mail* indicates that *voice mail* means *system that records messages on a telephone when nobody answers*.

As exemplified by Finlayson and Kulkarni (2011), while the word *voice* has 11 senses and the word *mail* has 5, the expression *voice mail* only has 1. They show that, in Wordnet 1.6, the average polysemy of MWEs is of 1.07 synsets, versus 1.53 for simple words. To the best of our knowledge, their work is the first to report a considerable improvement on word sense disambiguation performance due to the detection of MWEs. Despite its simplicity, their method reaches an improvement of 5 F-measure points given lower and upper bounds of 3.3 and 6.1.

Bungum et al. (2013) perform collocation identification prior to translation in order to improve the quality of cross-lingual WSD. They argue that, while the coverage of the method depends on the quality of the collocation dictionaries, it helps improving precision. Thus, the use of even simple MWE resources like monolingual lists can help in a complex task like cross-lingual WSD.

3.3.4.3 Information Retrieval (IR)

Let us consider a simplified IR system, modelling documents as bags of words and not keeping track of co-occurrences. For instance, if a document contains the term *rock star*, it will probably be retrieved as an answer to queries on geology (*rock*) and astronomy (*star*). If this MWE was represented as a unit in the index of the system, the precision of the retrieved documents could increase. Most current IR systems allow more sophisticated queries to be expressed through quotes and wildcards. However, representing only relevant MWEs instead of all possible n-grams in the documents could speed up the searches.

Joining the words of MWEs before indexation is a simple idea that was put in practice by Acosta et al. (2011). They tested the impact of a large set of automatically and manually acquired MWE dictionaries on standard IR test sets from the CLEF campaign. Their results show that there is a gain in mean average precision when MWEs are tokenised as single words.

Choosing the appropriate granularity for units to be indexed can be complicated in languages like Chinese, which do not use spaces to separate words. In this case, a prior phase of segmentation generally takes place before traditional IR indexation. Xu et al. (2010) propose a new measure for the tightness of 4-character sequences, as well as three procedures for word segmentation based on this measure. Then, they compare a standard segmentation tool with their methods in an IR system. They show that two of their segmentation strategies improve mean average precision on a test set.

A related task is topic modelling, a popular approach to joint clustering of documents and terms. The standard document representation in this task is a bag of words. However, as presented by Baldwin (2011), it is possible to consolidate the microlevel document representation with the help of MWEs. He argues that recent experimental results demonstrate that linguistically-richer document representations can enhance topic modelling.

3.3.4.4 Machine Translation

In current MT systems, various practical solutions have been implemented. The expert MT system ITS-2 handles MWEs at two levels (Wehrli 1998). Contiguous compounds are dealt with during lexical analysis and treated as single words in subsequent steps. Idiomatic, non-fixed units are treated by the syntax analysis module, requiring a much more sophisticated description. Once they are correctly identified,

however, their transfer is executed in the same way as regular structures. The system implements a more sophisticated approach for non-fixed MWE identification in the syntactic analysis module (Wehrli et al. 2010). When evaluated on a data set of English/Italian→French translations, this strategy improved the quality of 10–16 % of a test set of 100 collocations (manual evaluation).

Haugereid and Bond (2011) enriched the Jaen Japanese–English MT system with MWE rules. Jaen is a semantic transfer MT system based on the HPSG parsers JACY and ERG. The authors use GIZA++ and Anymalign in order to generate phrase tables from parallel corpora, from which they automatically extract the new transfer rules. These rules are then filtered and, when added to the system, improve translation coverage from 19.3 to 20.1% and translation quality from 17.8 to 18.2 % NEVA score – a variant of BLEU). Even though the improvements are quite modest, the authors argue that they can be further improved by learning even more rules.

Morin and Daille (2010) obtain an improvement of 33 % in the French–Japanese translation of MWEs. They implement a morphologically-based compositional method for backing-off when there is not enough data in a dictionary to translate a MWE. For example, *chronic fatigue syndrome* can be decomposed as *[chronic fatigue] [syndrome]*, *[chronic] [fatigue syndrome]* or *[chronic] [fatigue] [syndrome]*.

Grefenstette (1999) addresses the translation of noun compounds from German and Spanish into English. He uses web counts to select translations for compositional noun compounds, and achieves an impressive accuracy of 0.86–0.87. Similarly, Tanaka and Baldwin (2003) compare two shallow translation methods for English–Japanese noun compounds. The first one is a static memory-based method where the compound needs to be present in the dictionary in order to be translated correctly. The second is a dynamic compositional method in which alternative translations are validated using corpus evidence. Their evaluation considers the compounds as test translation units (as opposed to traditional sentence-based evaluation). When they combine the two methods, they achieve 95 % coverage and potentially high translation accuracy. This method is further refined by the use of a support vector machine model to rank all possible translations (Baldwin and Tanaka 2004). The model learns the translation scores based on several features coming from monolingual and bilingual dictionaries and corpora.

The popular phrase-based models of the Moses SMT toolkit represent MWEs as flat contiguous sequences of words (Koehn et al. 2007). Bilingual MWEs are bilingual sequences, called "bi-phrases", and have several probabilities associated to them. Carpuat and Diab (2010) propose two complementary strategies in order to add monolingual MWEs from WordNet into an English–Arabic Moses system. The first strategy is a static single-tokenisation that treats MWEs as words-with-spaces. The second strategy is dynamic, adding to the translation model a count for the number of MWEs in the source part of the bi-phrase. They found that both strategies result in improvement of translation quality, which suggests that Moses bi-phrases alone do not model all MWE information.

Nakov (2008a) propose another approach for minimizing data sparseness in MT, based on the generation of monolingual paraphrases to augment the training corpus. Parse trees are used as the basis for generating paraphrases that are nearly-equivalent

semantically (e.g., *ban on beef import* for *beef import ban* and vice versa). The trees are syntactically transformed by a set of heuristics, looking at noun compounds and related constructions. Using Moses' ancestor, Pharaoh, on an English–Spanish task, this technique generates an improvement equivalent to 33–50 % of that of doubling training data.

Automatic word alignment can be more challenging when translating from and to morphologically rich languages. In German and in Scandinavian languages, for instance, a compound is in fact a single token formed through concatenation of words and special infixes (*Hauptbahnhof* is the concatenation of *Haupt (main), Bahn (railway) and Hof (station)*). Stymne (2011) develops a fine-grained typology for MT error analysis which includes concatenated definite and compound nouns. For definiteness, she makes the source text look more like the target text (or vice versa) during training, thus making the learning less prone to errors by using better word alignments. In Stymne (2009), she describes her approach to noun compounds, which she splits into their single word components prior to translation. Then, after translation, she applies some post-processing rules like the reordering or merging of the components.

Pal et al. (2010) explore the extension of a Moses English–Bengali system. Significant improvements are brought by applying preprocessing steps like single-tokenisation for named entities and compound verbs. However, larger improvements (4.59 absolute BLEU points) are observed when using a statistical model for the prior alignment of named entities, allowing for their adequate transliteration.

The domain adaptation of general-purpose MT systems can also be accomplished with the integration of multiword terms. Ren et al. (2009) adapt a Chinese–English standard Moses system using three simple techniques: appending the MWE lexicon to the corpus, appending it to the phrase table, and adding a binary feature to the translation model. They found significant BLEU improvements over the baseline, especially using the extra feature.

In translation memory systems such as Similis, the translation unit can be considered as a MWE as it is an intermediary between words and sentences. The correspondences of word sequences are automatically learned from the translation memory and expressed in a multi-layer architecture including surface forms, lemmas and parts of speech (Planas and Furuse 2000).

Hierarchical and tree-based translation systems like Joshua use tree rewriting rules in order to represent the correspondences between source and target structures (Li et al. 2009). However, it is difficult to implement special rules for MWEs and to distinguish them from rules that should be applied to ordinary word combinations. Promising results in the application of MWE resources such as lexicons and thesauri show that this is a recent and apparently growing topic in the MT community.

Monti et al. (2011) compile a parallel corpus of sentences containing several types of expressions and compare the outputs of rule-based and SMT systems. While their discussion provides insightful examples, it does not help quantify the extent to which multiword expressions pose problems to MT systems. Moreover, it is not possible to know the exact details of the MT paradigms used in their experiments.

The proceedings of the 2013 MUMTTT workshop provide some pointers on different approaches for MWE translation (Mitkov et al. 2013). Our experiments on the evaluation of automatic MWE translation are presented in Chap. 7.

3.4 Summary

The underlying hypothesis in MWE acquisition is that words that form an expression will co-occur more often than if they were randomly combined. This hypothesis is applied in the design of lexical association measures (AMs) for corpus-based MWE acquisition. There are numerous AMs available for MWE acquisition (Evert 2004; Seretan 2008; Pecina 2008). For an arbitrary n-gram, we estimate its probability under MLE and scale this estimate by the total number of n-grams in the corpus, obtaining the expected count. AMs are generally based on the difference between the expected count and the observed count, like t-score, pointwise mutual information and Dice coefficient. More robust and theoretically sound AMs based on contingency tables exist for the special case of 2-grams. Examples of such measures are χ^2 and log-likelihood.

MWE acquisition comprises identification (in context) and extraction (out of context). Monolingual MWE acquisition is generally seen as a two-step process: candidate extraction and candidate filtering. In candidate extraction, POS sequences are one of the major approaches, specially for terminology (Justeson and Katz 1995; Daille 2003), but also for noun compounds (Vincze et al. 2011) and verbal expressions (Baldwin 2005). When a parser is available, syntactic patterns can be much more precise than POS sequences, specially in the extraction of non-fixed MWEs (Seretan and Wehrli 2009; Seretan 2008). Tree substitution grammars (Green et al. 2011) and structural regularities in parsing trees (Martens and Vandeghinste 2010) can also be used in order to learn syntactic MWE models from annotated corpora. During candidate filtering, some straightforward procedures are the use of stopword lists and of count thresholds. AMs are also widely employed to rank the candidates and keep only those whose association score is above a certain threshold (Evert and Krenn 2005; Pecina 2005). Supervised learning methods can be used to build a classifier modelling the optimal weights of several AMs and other features (Ramisch et al. 2008; Pecina 2008).

As for bilingual acquisition, automatic word alignments can provide lists of MWE candidates by themselves (de Medeiros Caseli et al. 2010). Bai et al. (2009) present an algorithm capable of mining translations for a given MWE in a parallel aligned corpus. The automatic discovery of non-compositional compounds from parallel data has been explored by Melamed (1997). The English-Hindi language pair presents large word order variation, and it has been shown that MWE-based features that model compositionality can help reducing alignment error rate (Venkatapathy and Joshi 2006). Zarrieß and Kuhn (2009) used syntactically analysed corpora and GIZA++ alignments to extract verb-object pairs from a German–English parallel corpus. Daille et al. (2004) performed multiword term

extraction from comparable corpora in French and in English, and subsequently used the distances between the context vectors to obtain cross-lingual equivalences.

The *syntactic interpretation* of nouns compounds has been explored by Nicholson and Baldwin (2006), who distinguish three syntactic relations in noun–noun compounds: subject, direct object and prepositional object. Three-word or longer noun compounds require syntactic interpretation of the constituent hierarchy. Nakov and Hearst (2005) compare two models, based on adjacency and on dependency. Nakov and Hearst (2008) perform unsupervised *semantic interpretation* of noun compounds by generating a large number of paraphrases involving verbs related to the semantic classes and then retrieving their web counts. Kim and Nakov (2011) used a combination of data bootstrapping and web counts, using paraphrases based on parse trees. Cook and Stevenson (2006) use support vector machines to classify the meaning of the particle *up* in English phrasal verbs. Bannard (2005) investigates the extent to which the components of a phrasal verb contribute their meanings to the interpretation of the whole. A similar work is that of McCarthy et al. (2003), who propose several measures to estimate the idiomaticity of phrasal verbs.

The disambiguation of MWEs is analogous to their interpretation, except that they are considered together with the context in which they appear. Nicholson and Baldwin (2008) present a data set for noun–noun compound disambiguation where a large set of sentences has been manually annotated. Girju et al. (2005) investigate methods for their disambiguation by applying several supervised learning techniques. Fritzinger et al. (2010) manually analyse a large set of ambiguous German preposition–noun–verb constructions retrieved by a parser, classifying them as either literal, compositional or unknown. Light verbs in Japanese have also been studied by Uchiyama et al. (2005), who proposes two disambiguation methods: a statistical approach and a rule-based method. Cook et al. (2007) explore the idiomaticity of verb–noun pairs, where the noun is the direct object of the verb and may have an idiomatic (*make a face*) or literal (*make a cake*) interpretation. Fazly and Stevenson (2007) propose a more fine-grained classification for light verb–noun constructions, using a supervised learning strategy in order to perform a 4-way semantic disambiguation.

Sag et al. (2002) proposed two approaches to represent MWEs in lexicons: words-with-spaces and compositional. However, between these extremes of the compositionality spectrum, there are some other possibilities, explored in related work. Laporte and Voyatzi (2008) describe a dictionary of French adverbial expressions and their corresponding morphosyntactic patterns in the lexicon–grammar format. Graliński et al. (2010) present a qualitative and quantitative comparison between two structured representations, Multiflex and POLENG, for Polish MWEs. Grégoire (2007, 2010) uses an equivalence class method that groups similar expressions according to their syntactic characteristics. Izumi et al. (2010) suggest a rule-based method to normalise Japanese functional expressions in order to optimise their representation. Schuler and Joshi (2011) propose the use of tree-rewriting grammars to describe MWEs.

There are some target applications for which concrete results have been obtained. Constant and Sigogne (2011) present promising results for French parsing. Korkontzelos and Manandhar (2010) obtain impressing improvements by enriching a baseline shallow parser with MWEs. Zhang and Kordoni (2006) and Villavicencio et al. (2007) obtain a significant coverage increase by extending the lexicon of an English HPSG parser with MWEs. Wehrli et al. (2010) demonstrate that MWEs are not a "pain in the neck" but actually a valuable information to reduce syntactic ambiguity.

References

Acosta O, Villavicencio A, Moreira V (2011) Identification and treatment of multiword expressions applied to information retrieval. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Association for Computational Linguistics, Portland, pp 101–109. http://www.aclweb.org/anthology/W/W11/W11-0815

Anastasiou D, Hashimoto C, Nakov P, Kim SN (eds) (2009) Proceedings of the ACL workshop on multiword expressions: identification, interpretation, disambiguation, applications (MWE 2009), Singapore. Association for Computational Linguistics/Suntec. http://aclweb.org/anthology-new/W/W09/W09-29, 70 p.

Apresian J, Boguslavsky I, Iomdin L, Tsinman L (2003) Lexical functions as a tool of ETAP-3. In: Proceedings of the first international conference on meaning-text theory (MTT 2003), Paris

Attia M, Toral A, Tounsi L, Pecina P, van Genabith J (2010) Automatic extraction of Arabic multiword expressions. In: Laporte É, Nakov P, Ramisch C, Villavicencio A (eds) Proceedings of the COLING workshop on multiword expressions: from theory to applications (MWE 2010), Beijing. Association for Computational Linguistics, pp 18–26

Baayen RH (2001) Word frequency distributions, text, speech and language technology, vol 18. Springer, Berlin/New York

Bai MH, You JM, Chen KJ, Chang JS (2009) Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In: Proceedings of the 2009 conference on empirical methods in natural language processing (EMNLP 2009), Singapore. Association for Computational Linguistics/Suntec, pp 478–486

Baldwin T (2005) Deep lexical acquisition of verb-particle constructions. Comput Speech Lang Spec Issue MWEs 19(4):398–414

Baldwin T (2011) MWEs and topic modelling: enhancing machine learning with linguistics. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, p 1. http://www.aclweb.org/anthology/W/W11/W11-0801

Baldwin T, Tanaka T (2004) Translation by machine of complex nominals: getting it right. In: Tanaka T, Villavicencio A, Bond F, Korhonen A (eds) Proceedings of the ACL workshop on multiword expressions: integrating processing (MWE 2004), Barcelona. Association for Computational Linguistics, pp 24–31

Baldwin T, Bannard C, Tanaka T, Widdows D (2003) An empirical model of multiword expression decomposability. In: Bond F, Korhonen A, McCarthy D, Villavicencio A (eds) Proceedings of the ACL workshop on multiword expressions: analysis, acquisition and treatment (MWE 2003), Sapporo. Association for Computational Linguistics, pp 89–96. doi:10.3115/1119282.1119294, http://www.aclweb.org/anthology/W03-1812

Banerjee S, 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, Mexico City, pp 370–381

Bannard C (2005) Learning about the meaning of verb-particle constructions from corpora. Comput Speech Lang Spec Issue MWEs 19(4):467–478

Bejček E, Stranak P, Pecina P (2013) Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the 9th workshop on multiword expressions (MWE 2013), Atlanta. Association for Computational Linguistics, pp 106–115. http://www.aclweb.org/anthology/W13-1016

Bonin F, Dell'Orletta F, Montemagni S, Venturi G (2010a) A contrastive approach to multi-word extraction from domain-specific corpora. In: Proceedings of the seventh international conference on language resources and evaluation (LREC 2010), Valetta. European Language Resources Association

Bonin F, Dell'Orletta F, Venturi G, Montemagni S (2010b) Contrastive filtering of domain-specific multi-word terms from different types of corpora. In: Laporte É, Nakov P, Ramisch C, Villavicencio A (eds) Proceedings of the COLING workshop on multiword expressions: from theory to applications (MWE 2010), Beijing. Association for Computational Linguistics, pp 76–79

Bouamor D, Semmar N, Zweigenbaum P (2012) Identifying bilingual multi-word expressions for statistical machine translation. In: Proceedings of the eigth international conference on language resources and evaluation (LREC 2012), Istanbul. European Language Resources Association

Briscoe T, Carroll J, Watson R (2006) The second release of the RASP system. In: Curran J (ed) Proceedings of the COLING/ACL 2006 interactive presentation sessions, Sidney. Association for Computational Linguistics, pp 77–80. http://www.aclweb.org/anthology/P/P06/P06-4020

Bungum L, Gambäck B, Lynum A, Marsi E (2013) Improving word translation disambiguation by capturing multiword expressions with dictionaries. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the 9th workshop on multiword expressions (MWE 2013), Atlanta. Association for Computational Linguistics, pp 21–30. http://www.aclweb.org/anthology/W13-1003

Burnard L (2007) User reference guide for the British National Corpus. Technical report, Oxford University Computing Services

Butnariu C, Kim SN, Nakov P, Séaghdha DO, Szpakowicz S, Veale T (2010) Semeval-2 task 9: the interpretation of noun compounds using paraphrasing verbs and prepositions. In: Erk K, Strapparava C (eds) Proceedings of the 5th international workshop on semantic evaluation (SemEval 2010), Uppsala. Association for Computational Linguistics, pp 39–44. http://www.aclweb.org/anthology/S10-1007

Carpuat M, Diab M (2010) Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In: Proceedings of human language technology: the 2010 annual conference of the North American chapter of the Association for Computational Linguistics (NAACL 2003), Los Angeles. Association for Computational Linguistics, pp 242–245. http://www.aclweb.org/anthology/N10-1029

Chen SF, Goodman J (1999) An empirical study of smoothing techniques for language modeling. Comput Speech Lang 13(4):359–394

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

Constant M, Sigogne A (2011) MWU-aware part-of-speech tagging with a CRF model and lexical resources. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real World (MWE 2011), Portland. Association for Computational Linguistics, pp 49–56. http://www.aclweb.org/anthology/W/W11/W11-0809

Constant M, Roux JL, Sigogne A (2013) Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. ACM Trans Speech Lang Process Spec Issue Multiword Expr Theory Pract Use Part 2 (TSLP) 10(3):1–24

Cook P, Stevenson S (2006) Classifying particle semantics in English verb-particle constructions. In: Moirón BV, Villavicencio A, McCarthy D, Evert S, Stevenson S (eds) Proceedings of

the COLING/ACL workshop on multiword expressions: identifying and exploiting underlying properties (MWE 2006), Sidney. Association for Computational Linguistics, pp 45–53. http://www.aclweb.org/anthology/W/W06/W06-1207

Cook P, Fazly A, Stevenson S (2007) Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In: Grégoire N, Evert S, Kim SN (eds) Proceedings of the ACL workshop on a broader perspective on multiword expressions (MWE 2007), Prague. Association for Computational Linguistics, pp 41–48. http://www.aclweb.org/anthology/W/W07/W07-1106

Cook P, Fazly A, Stevenson S (2008) The VNC-tokens dataset. In: Grégoire N, Evert S, Krenn B (eds) Proceedings of the LREC workshop towards a shared task for multiword expressions (MWE 2008), Marrakech, pp 19–22

Daille B (2003) Conceptual structuring through term variations. In: Bond F, Korhonen A, McCarthy D, Villavicencio A (eds) Proceedings of the ACL workshop on multiword expressions: analysis, acquisition and treatment (MWE 2003), Sapporo. Association for Computational Linguistics, pp 9–16. doi:10.3115/1119282.1119284. http://www.aclweb.org/anthology/W03-1802

Daille B, Dufour-Kowalski S, Morin E (2004) French-English multi-word term alignment based on lexical context analysis. In: Proceedings of the fourth international conference on language resources and evaluation (LREC 2004), Lisbon. European Language Resources Association, pp 919–922

Déjean H, Gaussier É, Sadat F (2002) An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In: Proceedings of the 19th international conference on computational linguistics (COLING 2002), Taipei. http://aclweb.org/anthology-new/C/C02/C02-1166.pdf

de Medeiros Caseli H, Villavicencio A, Machado A, Finatto MJ (2009) Statistically-driven alignment-based multiword expression identification for technical domains. In: Anastasiou D, Hashimoto C, Nakov P, Kim SN (eds) Proceedings of the ACL workshop on multiword expressions: identification, interpretation, disambiguation, applications (MWE 2009), Singapore. Association for Computational Linguistics/Suntec, pp 1–8

de Medeiros Caseli H, Ramisch C, das Graças Volpe Nunes M, Villavicencio A (2010) Alignment-based extraction of multiword expressions. Lang Resour Eval Spec Issue Multiword Express Hard Going Plain Sail 44(1–2):59–77. doi:10.1007/s10579-009-9097-9, http://www.springerlink.com/content/H7313427H78865MG

Dias G (2003) Multiword unit hybrid extraction. In: Bond F, Korhonen A, McCarthy D, Villavicencio A (eds) Proceedings of the ACL workshop on multiword expressions: analysis, acquisition and treatment (MWE 2003), Sapporo. Association for Computational Linguistics, pp 41–48. doi:10.3115/1119282.1119288. http://www.aclweb.org/anthology/W03-1806

Duan J, Lu R, Wu W, Hu Y, Tian Y (2006) A bio-inspired approach for multi-word expression extraction. In: Curran J (ed) Proceedings of the COLING/ACL 2006 main conference poster sessions, Sidney. Association for Computational Linguistics, pp 176–182. http://www.aclweb.org/anthology/P/P06/P06-2023

Dunning T (1993) Accurate methods for the statistics of surprise and coincidence. Comput Linguist 19(1):61–74

Duran MS, Ramisch C, Aluísio SM, Villavicencio A (2011) Identifying and analyzing Brazilian Portuguese complex predicates. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, pp 74–82. http://www.aclweb.org/anthology/W/W11/W11-0812

Evert S (2004) The statistics of word cooccurrences: word pairs and collocations. PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, 353p

Evert S, Krenn B (2005) Using small random samples for the manual evaluation of statistical association measures. Comput Speech Lang Spec Issue MWEs 19(4):450–466

Fazly A, Stevenson S (2007) Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In: Grégoire N, Evert S, Kim SN (eds) Pro-

ceedings of the ACL workshop on a broader perspective on multiword expressions (MWE 2007), Prague. Association for Computational Linguistics, pp 9–16. http://www.aclweb.org/anthology/W/W07/W07-1102

Finlayson M, Kulkarni N (2011) Detecting multi-word expressions improves word sense disambiguation. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, pp 20–24. http://www.aclweb.org/anthology/W/W11/W11-0805

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

Fritzinger F, Weller M, Heid U (2010) A survey of idiomatic preposition-noun-verb triples on token level. In: Proceedings of the seventh international conference on language resources and evaluation (LREC 2010), Valetta. European Language Resources Association, pp 2908–2914

Gil A, Dias G (2003) Using masks, suffix array-based data structures and multidimensional arrays to compute positional *n*-gram statistics from corpora. In: Bond F, Korhonen A, McCarthy D, Villavicencio A (eds) Proceedings of the ACL workshop on multiword expressions: analysis, acquisition and treatment (MWE 2003), Sapporo. Association for Computational Linguistics, pp 25–32. doi:10.3115/1119282.1119286, http://www.aclweb.org/anthology/W03-1804

Girju R, Moldovan D, Tatu M, Antohe D (2005) On the semantics of noun compounds. Comput Speech Lang Spec Issue MWEs 19(4):479–496

Good IJ (1953) The population frequencies of species and the estimation of population parameters. Biometrika 40(3–4):237–264. doi:10.1093/biomet/40.3-4.237

Graliński F, Savary A, Czerepowicka M, Makowiecki F (2010) Computational lexicography of multi-word units: how efficient can it be? In: Laporte É, Nakov P, Ramisch C, Villavicencio A (eds) Proceedings of the COLING workshop on multiword expressions: from theory to applications (MWE 2010), Beijing. Association for Computational Linguistics, pp 1–9

Green S, de Marneffe MC, Bauer J, Manning CD (2011) Multiword expression identification with tree substitution grammars: a parsing tour de force with French. In: Barzilay R, Johnson M (eds) Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP 2011), Edinburgh. Association for Computational Linguistics, pp 725–735. http://www.aclweb.org/anthology/D11-1067

Grefenstette G (1999) The world wide web as a resource for example-based machine translation tasks. In: Proceedings of the twenty-first international conference on translating and the computer, ASLIB, London

Grégoire N (2007) Design and implementation of a lexicon of Dutch multiword expressions. In: Grégoire N, Evert S, Kim SN (eds) Proceedings of the ACL workshop on a broader perspective on multiword expressions (MWE 2007), Prague. Association for Computational Linguistics, pp 17–24. http://www.aclweb.org/anthology/W/W07/W07-1103

Grégoire N (2010) DuELME: a Dutch electronic lexicon of multiword expressions. Lang Resour Eval Spec Issue Multiword Expr Hard Going Plain Sail 44(1–2):23–39. doi:10.1007/s10579-009-9094-z. http://www.springerlink.com/content/7308605442W17698

Grégoire N, Evert S, Krenn B (eds) (2008) Proceedings of the LREC workshop towards a shared task for multiword expressions (MWE 2008), Marrakech, 57p. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W20_Proceedings.pdf

Gurrutxaga A, Alegria I (2011) Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, pp 2–7. http://www.aclweb.org/anthology/W/W11/W11-0802

Haugereid P, Bond F (2011) Extracting transfer rules for multiword expressions from parallel corpora. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, pp 92–100. http://www.aclweb.org/anthology/W/W11/W11-0814

Hendrickx I, Kim SN, Kozareva Z, Nakov P, Séaghdha DO, Padó S, Pennacchiotti M, Romano L, Szpakowicz S (2010) Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. In: Erk K, Strapparava C (eds) Proceedings of the 5th international workshop on semantic evaluation (SemEval 2010), Uppsala. Association for Computational Linguistics, pp 33–38. http://www.aclweb.org/anthology/S10-1006

Hoang HH, Kim SN, Kan MY (2009) A re-examination of lexical association measures. In: Anastasiou D, Hashimoto C, Nakov P, Kim SN (eds) Proceedings of the ACL workshop on multiword expressions: identification, interpretation, disambiguation, applications (MWE 2009), Singapore. Association for Computational Linguistics/Suntec, pp 31–39

Hogan D, Foster J, van Genabith J (2011) Decreasing lexical data sparsity in statistical syntactic parsing – experiments with named entities. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, pp 14–19. http://www.aclweb.org/anthology/W/W11/W11-0804

Izumi T, Imamura K, Kikui G, Sato S (2010) Standardizing complex functional expressions in Japanese predicates: applying theoretically-based paraphrasing rules. In: Laporte É, Nakov P, Ramisch C, Villavicencio A (eds) Proceedings of the COLING workshop on multiword expressions: from theory to applications (MWE 2010), Beijing. Association for Computational Linguistics, pp 63–71

Jurafsky D, Martin JH (2008) Speech and language processing, 2nd edn. Prentice Hall, Upper Saddle River, 1024p

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

Keller F, Lapata M (2003) Using the web to obtain frequencies for unseen bigrams. Comput Linguist Spec Issue Web Corpus 29(3):459–484

Kim SN, Baldwin T (2013) A lexical semantic approach to interpreting and bracketing English noun compounds. Nat Lang Eng Spec Issue Noun Compd 19(3):385–407. doi:10.1017/S1351324913000107, http://journals.cambridge.org/article_S1351324913000107

Kim SN, Nakov P (2011) Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In: Barzilay R, Johnson M (eds) Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP 2011), Edinburgh. Association for Computational Linguistics, pp 648–658. http://www.aclweb.org/anthology/D11-1060

Kneser R, Ney H (1995) Improved backing-off for M-gram language modeling. In: Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP 1995), Detroit, vol 1, pp 181–184. doi:10.1109/ICASSP.1995.479394, http://dx.doi.org/10.1109/ICASSP.1995.479394

Koehn P (2005) Europarl: a parallel corpus for statistical machine translation. In: Proceedings of the tenth machine translation summit (MT Summit 2005), Phuket. Asian-Pacific Association for Machine Translation, pp 79–86

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, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the Association for Computational Linguistics (ACL 2007), Prague. Association for Computational Linguistics, pp 177–180

Korkontzelos I, Manandhar S (2010) Can recognising multiword expressions improve shallow parsing? In: Proceedings of human language technology: the 2010 annual conference of the North American chapter of the Association for Computational Linguistics (NAACL 2003), Los Angeles. Association for Computational Linguistics, pp 636–644. http://www.aclweb.org/anthology/N10-1089

Kulkarni N, Finlayson M (2011) jMWE: a java toolkit for detecting multi-word expressions. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, pp 122–124. http://www.aclweb.org/anthology/W/W11/W11-0818

Lapata M (2002) The disambiguation of nominalizations. Comput Linguist 28(3):357–388
Laporte É, Voyatzi S (2008) An electronic dictionary of French multiword adverbs. In: Grégoire N, Evert S, Krenn B (eds) Proceedings of the LREC workshop towards a shared task for multiword expressions (MWE 2008), Marrakech, pp 31–34
Laporte É, Nakamura T, Voyatzi S (2008) A French corpus annotated for multiword nouns. In: Grégoire N, Evert S, Krenn B (eds) Proceedings of the LREC workshop towards a shared task for multiword expressions (MWE 2008), Marrakech, pp 27–30
Li Z, Callison-Burch C, Dyer C, Ganitkevitch J, Khudanpur S, Schwartz L, Thornton WNG, Weese J, Zaidan OF (2009) Joshua: an open source toolkit for parsing-based machine translation. In: Proceedingsof the fourth workshop on statistical machine translation (WMT 2009), Athens. Association for Computational Linguistics, pp 135–139
Manber U, Myers G (1990) Suffix arrays: a new method for on-line string searches. In: SODA '90: proceedings of the first annual ACM-SIAM symposium on discrete algorithms, San Francisco. Society for Industrial and Applied Mathematics, Philadelphia, pp 319–327
Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT, Cambridge, 620p
Martens S (2010) Varro: an algorithm and toolkit for regular structure discovery in treebanks. In: Huang CR, Jurafsky D (eds) Proceedings of the 23rd international conference on computational linguistics (COLING 2010)—posters, Beijing. The Coling 2010 Organizing Committee, pp 810–818. http://www.aclweb.org/anthology/C10-2093
Martens S, Vandeghinste V (2010) An efficient, generic approach to extracting multi-word expressions from dependency trees. In: Laporte É, Nakov P, Ramisch C, Villavicencio A (eds) Proceedings of the COLING workshop on multiword expressions: from theory to applications (MWE 2010), Beijing. Association for Computational Linguistics, pp 84–87
McCarthy D, Keller B, Carroll J (2003) Detecting a continuum of compositionality in phrasal verbs. In: Bond F, Korhonen A, McCarthy D, Villavicencio A (eds) Proceedings of the ACL workshop on multiword expressions: analysis, acquisition and treatment (MWE 2003), Sapporo. Association for Computational Linguistics, pp 73–80. doi:10.3115/1119282.1119292, http://www.aclweb.org/anthology/W03-1810
McCarthy D, Venkatapathy S, Joshi A (2007) Detecting compositionality of verb-object combinations using selectional preferences. In: Eisner J (ed) Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007), Prague. Association for Computational Linguistics, pp 369–379. http://www.aclweb.org/anthology/D/D07/D07-1039
Melamed ID (1997) Automatic discovery of non-compositional compounds in parallel data. In: Proceedings of the 2nd conference on empirical methods in natural language processing (EMNLP-2), Brown University, Providence. Association for Computational Linguistics, pp 97–108
Michou A, Seretan V (2009) A tool for multi-word expression extraction in modern Greek using syntactic parsing. In: Proceedings of the demonstrations session at EACL 2009, Athens. Association for Computational Linguistics, pp 45–48
Mikheev A (2002) Periods, capitalized words, etc. Comput Linguist 28(3):289–318
Mirroshandel SA, Nasr A, Roux JL (2012) Semi-supervised dependency parsing using lexical affinities. In: Proceedings of the 50th annual meeting of the Association for Computational Linguistics (vol 1: long papers), Jeju Island. Association for Computational Linguistics, pp 777–785. http://www.aclweb.org/anthology/P12-1082
Mitkov R, Monti J, Pastor GC, Seretan V (eds) (2013) Proceedings of the MT summit 2013 workshop on multi-word units in machine translation and translation technology (MUMTTT 2013), Nice. European Association for Machine Translation, 71p. http://www.mtsummit2013.info/workshop4.asp
Monti J, Barreiro A, Elia A, Marano F, Napoli A (2011) Taking on new challenges in multi-word unit processing for machine translation. In: Proceedings of the second international workshop on free/open-source rule-based machine translation, Barcelona

Morin E, Daille B (2010) Compositionality and lexical alignment of multi-word terms. Lang Resour Eval Spec Issue Multiword Express Hard Going Plain Sail 44(1–2):79–95. doi:10.1007/s10579-009-9098-8, http://www.springerlink.com/content/30264870R1K04744

Nakov P (2007) Using the web as an implicit training set: application to noun compound syntax and semantics. PhD thesis, EECS Department, University of California, Berkeley, 392p

Nakov P (2008a) Improved statistical machine translation using monolingual paraphrases. In: Ghallab M, Spyropoulos CD, Fakotakis N, Avouris NM (eds) Proceedings of the 18th European conference on artificial intelligence (ECAI 2008), Patras. Frontiers in Artificial Intelligence and Applications, vol 178. IOS Press, pp 338–342

Nakov P (2008b) Paraphrasing verbs for noun compound interpretation. In: Grégoire N, Evert S, Krenn B (eds) Proceedings of the LREC workshop towards a shared task for multiword expressions (MWE 2008), Marrakech, pp 46–49

Nakov P (2013) On the interpretation of noun compounds: syntax, semantics, and entailment. Nat Lang Eng Spec Issue Noun Compd 19(3):291–330. doi:10.1017/S1351324913000065, http://journals.cambridge.org/article_S1351324913000065

Nakov P, Hearst MA (2005) Search engine statistics beyond the *n*-gram: application to noun compound bracketing. In: Dagan I, Gildea D (eds) Proceedings of the ninth conference on natural language learning (CoNLL-2005), University of Michigan, Ann Arbor. Association for Computational Linguistics, pp 17–24. http://www.aclweb.org/anthology/W/W05/W05-0603

Nakov P, Hearst MA (2008) Solving relational similarity problems using the web as a corpus. In: Proceedings of the 46th annual meeting of the Association for Computational Linguistics: human language technology (ACL-08: HLT), Columbus. Association for Computational Linguistics, pp 452–460

Nasr A, Bechet F, Rey JF, Favre B, Roux JL (2011) MACAON an NLP tool suite for processing word lattices. In: Proceedings of the ACL 2011 system demonstrations, Portland. Association for Computational Linguistics, pp 86–91. http://www.aclweb.org/anthology/P11-4015

Newman MEJ (2005) Power laws, pareto distributions and zipf's law. Contemp Phys 46:323–351

Nicholson J, Baldwin T (2006) Interpretation of compound nominalisations using corpus and web statistics. In: Moirón BV, Villavicencio A, McCarthy D, Evert S, Stevenson S (eds) Proceedings of the COLING/ACL workshop on multiword expressions: identifying and exploiting underlying properties (MWE 2006), Sidney. Association for Computational Linguistics, pp 54–61. http://www.aclweb.org/anthology/W/W06/W06-1208

Nicholson J, Baldwin T (2008) Interpreting compound nominalisations. In: Grégoire N, Evert S, Krenn B (eds) Proceedings of the LREC workshop towards a shared task for multiword expressions (MWE 2008), Marrakech, pp 43–45

Nulty P, Costello F (2010) UCD-PN: Selecting general paraphrases using conditional probability. In: Erk K, Strapparava C (eds) Proceedings of the 5th international workshop on semantic evaluation (SemEval 2010), Uppsala. Association for Computational Linguistics, pp 234–237. http://www.aclweb.org/anthology/S10-1052

Nulty P, Costello F (2013) General and specific paraphrases of semantic relations between nouns. Nat Lang Eng Spec Issue Noun Compd 19(3):357–384. doi:10.1017/S1351324913000089, http://journals.cambridge.org/article_S1351324913000089

Pal S, Naskar SK, Pecina P, Bandyopadhyay S, Way A (2010) Handling named entities and compound verbs in phrase-based statistical machine translation. In: Laporte É, Nakov P, Ramisch C, Villavicencio A (eds) Proceedings of the COLING workshop on multiword expressions: from theory to applications (MWE 2010), Beijing. Association for Computational Linguistics, pp 45–53

Pearce D (2002) A comparative evaluation of collocation extraction techniques. In: Proceedings of the third international conference on language resources and evaluation (LREC 2002), Las Palmas. European Language Resources Association, pp 1530–1536

Pecina P (2005) An extensive empirical study of collocation extraction methods. In: Proceedings of the ACL 2005 student research workshop, Ann Arbor. Association for Computational Linguistics, pp 13–18. http://www.aclweb.org/anthology/P/P05/P05-2003

Pecina P (2008) Reference data for Czech collocation extraction. In: Grégoire N, Evert S, Krenn B (eds) Proceedings of the LREC workshop towards a shared task for multiword expressions (MWE 2008), Marrakech, pp 11–14

Pedersen T, Banerjee S, McInnes B, Kohli S, Joshi M, Liu Y (2011) The n-gram statistics package (text::NSP): a flexible tool for identifying n-grams, collocations, and word associations. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, pp 131–133. http://www.aclweb.org/anthology/W/W11/W11-0821

Planas E, Furuse O (2000) Multi-level similar segment matching algorithm for translation memories and example-based machine translation. In: Proceedings of the 18th international conference on computational linguistics (COLING 2000), Saarbrücken. http://aclweb.org/anthology-new/C/C00/C00-2090.pdf

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

Ramisch C, Villavicencio A, Moura L, Idiart M (2008) Picking them up and figuring them out: verb-particle constructions, noise and idiomaticity. In: Clark A, Toutanova K (eds) Proceedings of the twelfth conference on natural language learning (CoNLL 2008), Manchester. The Coling 2008 Organizing Committee, pp 49–56. http://www.aclweb.org/anthology/W08-2107

Ramisch C, de Medeiros Caseli H, Villavicencio A, Machado A, Finatto MJ (2010) A hybrid approach for multiword expression identification. In: Proceedings of the 9th international conference on computational processing of Portuguese language (PROPOR 2010), Porto Alegre. Lecture notes in computer science (Lecture notes in artificail intelligence), vol 6001. Springer, pp 65–74. doi:10.1007/978-3-642-12320-7_9, http://www.springerlink.com/content/978-3-642-12319-1

Ren Z, Lü Y, Cao J, Liu Q, Huang Y (2009) Improving statistical machine translation using domain bilingual multiword expressions. In: Anastasiou D, Hashimoto C, Nakov P, Kim SN (eds) Proceedings of the ACL workshop on multiword expressions: identification, interpretation, disambiguation, applications (MWE 2009), Singapore. Association for Computational Linguistics/Suntec, pp 47–54

Roller S, im Walde SS, Scheible S (2013) The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the 9th workshop on multiword expressions (MWE 2013), Atlanta. Association for Computational Linguistics, pp 32–41. http://www.aclweb.org/anthology/W13-1005

Sag I, Baldwin T, Bond F, Copestake A, Flickinger D (2002) 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), Mexico City. Lecture notes in computer science, vol 2276/2010. Springer, pp 1–15

SanJuan E, Dowdall J, Ibekwe-SanJuan F, Rinaldi F (2005) A symbolic approach to automatic multiword term structuring. Comput Speech Lang Spec Issue MWEs 19(4):524–542

Schmid H (1994) Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the international conference on new methods in language processing, Manchester, pp 44–49. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.1139

Schone P, Jurafsky D (2001) Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In: Lee L, Harman D (eds) Proceedings of the 2001 conference on empirical methods in natural language processing (EMNLP 2001), Pittsburgh. Association for Computational Linguistics, pp 100–108

Schuler W, Joshi A (2011) Tree-rewriting models of multi-word expressions. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, pp 25–30. http://www.aclweb.org/anthology/W/W11/W11-0806

Séaghdha DÓ, Copestake A (2013) Interpreting compound nouns with kernel methods. Nat Lang Eng Spec Issue Noun Compd 19(3):331–356. doi:10.1017/S1351324912000368, http://journals.cambridge.org/article_S1351324912000368

Seretan V (2008) Collocation extraction based on syntactic parsing. PhD thesis, University of Geneva, Geneva, 249p

Seretan V (2011) Syntax-based Collocation extraction, text, speech and language technology, vol 44, 1st edn. Springer, Dordrecht, 212p

Seretan V, Wehrli E (2006) Multilingual collocation extraction: issues and solutions. In: Witt A, Sérasset G, Armstrong S, Breen J, Heid U, Sasaki F (eds) Proceedings of the ACL workshop on multilingual language resources and interoperability, Sydney. Association for Computational Linguistics, pp 40–49. http://www.aclweb.org/anthology/W/W06/W06-1006

Seretan V, Wehrli E (2009) Multilingual collocation extraction with a syntactic parser. Lang Resour Eval Spec Issue Multiling Lang Resour Interoper 43(1):71–85. doi:10.1007/s10579-008-9075-7, http://www.springerlink.com/content/341877K50497682X

Seretan V, Wehrli E (2011) Fipscoview: on-line visualisation of collocations extracted from multilingual parallel corpora. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, pp 125–127. http://www.aclweb.org/anthology/W/W11/W11-0819

Silva J, 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), Orlando, pp 369–381

Silva J, Lopes G (2010) Towards automatic building of document keywords. In: Huang CR, Jurafsky D (eds) Proceedings of the 23rd international conference on computational linguistics (COLING 2010)—posters, Beijing. The Coling 2010 Organizing Committee, pp 1149–1157. http://www.aclweb.org/anthology/C10-2132

da Silva JF, Dias G, Guilloré S, Lopes JGP (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, London. EPIA 1999, pp 113–132. Springer. http://dl.acm.org/citation.cfm?id=645377.651205

Smadja FA (1993) Retrieving collocations from text: xtract. Comput Linguist 19(1):143–177

Stymne S (2009) A comparison of merging strategies for translation of German compounds. In: Proceedings of the student research workshop at EACL 2009, Athens, pp 61–69

Stymne S (2011) Pre- and postprocessing for statistical machine translation into Germanic languages. In: Proceedings of the ACL 2011 student research workshop, Portland. Association for Computational Linguistics, pp 12–17. http://www.aclweb.org/anthology/P11-3003

Szpakowicz S, Bond F, Nakov P, Kim SN (2013) On the semantics of noun compounds. In: Nat Lang Eng Spec Issue Noun Compd 19(3):289–290. Cambridge Univesity Press, Cambridge

Tanaka T, Baldwin T (2003) Noun-noun compound machine translation a feasibility study on shallow processing. In: Bond F, Korhonen A, McCarthy D, Villavicencio A (eds) Proceedings of the ACL workshop on multiword expressions: analysis, acquisition and treatment (MWE 2003), Sapporo. Association for Computational Linguistics, pp 17–24. doi:10.3115/1119282.1119285. http://www.aclweb.org/anthology/W03-1803

Tsvetkov Y, Wintner S (2010) Extraction of multi-word expressions from small parallel corpora. In: Huang CR, Jurafsky D (eds) Proceedings of the 23rd international conference on computational linguistics (COLING 2010)—posters, Beijing. The Coling 2010 Organizing Committee, pp 1256–1264. http://www.aclweb.org/anthology/C10-2144

Tsvetkov Y, Wintner S (2011) Identification of multi-word expressions by combining multiple linguistic information sources. In: Barzilay R, Johnson M (eds) Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP 2011), Edinburgh. Association for Computational Linguistics, pp 836–845. http://www.aclweb.org/anthology/D11-1077

Uchiyama K, Baldwin T, Ishizaki S (2005) Disambiguating Japanese compound verbs. Comput Speech Lang Spec Issue MWEs 19(4):497–512

Uresova Z, Hajic J, Fucikova E, Sindlerova J (2013) An analysis of annotation of verb-noun idiomatic combinations in a parallel dependency corpus. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the 9th workshop on multiword expressions (MWE 2013), Atlanta. Association for Computational Linguistics, pp 58–63. http://www.aclweb.org/anthology/W13-1009

Venkatapathy S, Joshi AK (2006) Using information about multi-word expressions for the word-alignment task. In: Moirón BV, Villavicencio A, McCarthy D, Evert S, Stevenson S (eds) Proceedings of the COLING/ACL workshop on multiword expressions: identifying and exploiting underlying properties (MWE 2006), Sidney. Association for Computational Linguistics, pp 20–27. http://www.aclweb.org/anthology/W/W06/W06-1204

Villavicencio A, Bond F, Korhonen A, McCarthy D (2005) Introduction to the special issue on multiword expressions: having a crack at a hard nut. Comput Speech Lang Spec Issue MWEs 19(4):365–377

Villavicencio A, Kordoni V, Zhang Y, Idiart M, Ramisch C (2007) Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In: Eisner J (ed) Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007), Prague. Association for Computational Linguistics, pp 1034–1043. http://www.aclweb.org/anthology/D/D07/D07-1110

Vincze V, Nagy TI, Berend G (2011) Detecting noun compounds and light verb constructions: a contrastive study. In: Kordoni V, Ramisch C, Villavicencio A (eds) Proceedings of the ALC workshop on multiword expressions: from parsing and generation to the real world (MWE 2011), Portland. Association for Computational Linguistics, pp 116–121. http://www.aclweb.org/anthology/W/W11/W11-0817

Wehrli E (1998) Translating idioms. In: Proceedings of the 36th annual meeting of the Association for Computational Linguistics and 17th international conference on computational linguistics, Montreal, vol 2. Association for Computational Linguistics, pp 1388–1392. doi:10.3115/980691.980795. http://www.aclweb.org/anthology/P98-2226

Wehrli E, Seretan V, Nerima L (2010) Sentence analysis and collocation identification. In: Laporte É, Nakov P, Ramisch C, Villavicencio A (eds) Proceedings of the COLING workshop on multiword expressions: from theory to applications (MWE 2010), Beijing. Association for Computational Linguistics, pp 27–35

Wermter J, Hahn U (2006) You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics (COLING/ACL 2006), Sidney. Association for Computational Linguistics, pp 785–792

Xu Y, Goebel R, Ringlstetter C, Kondrak G (2010) Application of the tightness continuum measure to Chinese information retrieval. In: Laporte É, Nakov P, Ramisch C, Villavicencio A (eds) Proceedings of the COLING workshop on multiword expressions: from theory to applications (MWE 2010), Beijing. Association for Computational Linguistics, pp 54–62

Yamamoto M, Church K (2001) Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. Comput Linguist 27(1):1–30

Zarrieß S, Kuhn J (2009) Exploiting translational correspondences for pattern-independent MWE identification. In: Anastasiou D, Hashimoto C, Nakov P, Kim SN (eds) Proceedings of the ACL workshop on multiword expressions: identification, interpretation, disambiguation, applications (MWE 2009), Singapore. Association for Computational Linguistics/Suntec, pp 23–30

Zhang Y, Kordoni V (2006) Automated deep lexical acquisition for robust open texts processing. In: Proceedings of the sixth international conference on language resources and evaluation (LREC 2006), Genoa. European Language Resources Association, pp 275–280