Design and Implementation of a Lexicon of Dutch Multiword Expressions

Nicole Grégoire
Uil-OTS
University of Utrecht
Utrecht, The Netherlands
Nicole.Gregoire@let.uu.nl

Abstract

This paper describes the design and implementation of a lexicon of Dutch multiword expressions (MWEs). No exhaustive research on a standard lexical representation of MWEs has been done for Dutch before. The approach taken is innovative, since it is based on the Equivalence Class Method. Furthermore, the selection of the lexical entries and their properties is corpus-based. The design of the lexicon and the standard representation will be tested in Dutch NLP systems. The purpose of the current paper is to give an overview of the decisions made in order to come to a standard lexical representation and to discuss the description fields this representation comprises.

1 Introduction

This paper describes the design and implementation of a lexicon of Dutch multiword expressions (MWEs). MWEs are known to be problematic for natural language processing. A considerable amount of research has been conducted in this area. Most progress has been made especially in the field of multiword identification (Villada Moirón and Tiedemann, 2006; Katz and Giesbrecht, 2006; Zhang et al., 2006). Moreover, interesting papers have been written on the representation of MWEs, most of them focusing on a single class of MWEs, see section 2. This paper elaborates on a standard lexical representation for Dutch MWEs developed within the STEVIN IRME project.[1] Part of the project focused on the design and implementation of an electronic resource of 5,000 Dutch expressions that meets the criterion of being highly theory- and implementation-independent, and which can be used in various Dutch NLP systems. The selection of the lexical entries and their properties is corpus-based.

Work has been conducted on collecting Dutch MWEs in the past, yielding one commercial printed dictionary (de Groot, 1999), and an electronic resource called the *Referentiebestand Nederlands* ('Reference Database of The Dutch Language') (Martin and Maks, 2005), both mainly meant for human users. No focus had been put on creating a standard representation for Dutch MWEs that can be converted into any system specific representation. The approach taken is innovative, since it is based on the Equivalence Class Method (ECM) (Odijk, 2004b). The idea behind the ECM is that MWEs that have the same pattern require the same treatment in an NLP system. MWEs with the same pattern form so-called Equivalence Classes (ECs). Having the ECs, it requires some manual work to convert one instance of an EC into a system specific representation, but all other members of the same EC can be done in a fully automatic manner. This method is really powerful since very detailed pattern descriptions can be used for describing the characteristics of a group of MWEs. Besides the description of the MWE patterns, we designed a uniform representation for the description of the individual expressions. Both the pattern descriptions and the MWE descriptions are implemented in the *Lexicon*

[1] http://www-uilots.let.uu.nl/irme/

Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pages 17–24,
Prague, June 2007.

of Dutch MWEs.

The purpose of this paper is to give an overview of the decisions made in order to come to a standard lexical representation and furthermore to discuss the description fields that are part of this representation.

The paper starts with an overview of related research in section 2. This is followed by elaborating the *Lexicon of Dutch MWEs* in section 3, a discussion in section 4, and a conclusion in section 5.

2 Related research: classes and representations

The area of multiword expressions includes many different subtypes, varying from fixed expressions to syntactically more flexible expressions. Sag et al. (2001) wrote a well-known paper on subclasses of MWEs, in which they make a distinction between *lexicalized phrases* and *institutionalized phrases*. Lexicalized phrases are subdivided into fixed, semi-fixed and flexible expressions. The most important reason for this subdivision is the variation in the degree of syntactic flexibility of MWEs. Roughly they claim that syntactic flexibility is related to semantic decomposability. Semantically non-decomposable idioms are idioms the meaning of which cannot be distributed over its parts and which are therefore not subject to syntactic variability. Sag et al. state that "the only types of lexical variation observable in non-decomposable idioms are inflection (*kicked the bucket*) and variation in reflexive form (*wet oneself*)." Examples of non-decomposable idioms are the oft-cited *kick the bucket* and *shoot the breeze*. On the contrary, semantically decomposable idioms, such as *spill the beans*, tend to be syntactically flexible to some degree. Mapping the boundaries of flexibility, however, is not always easy and no one can predict exactly which types of syntactic variation a given idiom can undergo.

One subtype of flexible expressions discussed in Sag et al. (2001) is the type of *Light Verb Constructions* (or *Support Verb Constructions* (SVCs)). SVCs are combinations of a verb that seems to have very little semantic content and a prepositional phrase, a noun phrase or adjectival phrase. An SVC is often paraphrasable by means of a single verb or adjective. Since the complement of the verb is used in its normal sense, the constructions are subject to standard grammar rules, which include passivization, internal modification, etc. The lexical selection of the verb is highly restricted. Examples of SVCs are *give/*make a demo*, *make/*do a mistake*.

As stated, no exhaustive research on a standard representation of MWEs has been done for Dutch before. Work on this topic has been conducted for other languages, which in most cases focused on a single subtype. Both Dormeyer and Fischer (1998) and Fellbaum et al. (2006) report on work on a resource for German verbal idioms, while the representation of German PP-verb collocations is addressed in (Krenn, 2000). Kuiper et al. (2003) worked on a representation of English idioms, and Villavicencio et al. (2004) proposed a lexical encoding of MWEs in general, by analysing English idioms and verb-partical constructions. Except for the SAID-database (Kuiper et al., 2003), which comprises over 13,000 expression, the created resources contain no more than 1,000 high-frequent expressions. Both Fellbaum et al. and Krenn support their lexical annotation with a corpus-based investigation. In our approach, we also use data extracted from corpora as empirical material, see section 3.2.

In most resources addressed, some kind of syntactic analysis is assigned to individual expressions. The most sophisticated syntactic analysis is done in the SAID-database. The approach taken by Kuiper et al. (2003) would have been more theory-independent, if it included a textual description, according to which classes of idioms could be formed. Villavicencio et al. (2004) defined a specific meta-type for each particular class of MWEs. The meta-types can be used to map the semantic relations between the components of an MWE into grammar specific features. Examples of meta-types specified are *verb-object-idiom* and *verb-particle-np*. They state that the majority of the MWEs in their database could be described by the meta-types defined. But since only a sample of 125 verbal idioms was used for the classification, no estimation can be given of how many classes this approach yields, when consulting a larger set of various types of MWEs. Fellbaum et al. (2006) provide a dependency structure for each expression, but not with the intention of grouping the entries accordingly.

To conclude this section, although our approach is in line with some of the projects described, our work is also distinctive because (1) it focuses on Dutch; (2) it does not solely focus on one type of MWEs, but on MWEs in general; (3) the lexicon includes 5,000 unique expressions, and (4) for an initial version of the lexicon a conversion to the Dutch NLP system Alpino[2] has been tested. In the remainder of this paper we discuss our approach to the lexical representation of MWEs.

3 A Lexicon of Dutch MWEs

In our research multiword expressions are defined as a combination of words that has linguistic properties not predictable from the individual components or the normal way they are combined (Odijk, 2004a). The linguistic properties can be of any type, e.g. *in line* is an MWE according to its syntactic characteristics, since it lacks a determiner preceding the singular count noun *line*, which is obligatory in standard English grammar.

Various aspects played a role in the representation as it is in the *Lexicon of Dutch MWEs*. First of all, the main requirement of the standard encoding is that it can be converted into any system specific representation with a minimal amount of manual work. The method adopted to achieve this goal is the Equivalence Class Method (ECM) (Odijk, 2004b). As stated, the ECM is based on the idea that given a class of MWE descriptions, representations for a specific theory and implementation can be derived. The procedure is that one instance of an Equivalence Class (EC) must be converted manually. By defining and formalizing the conversion procedure, the other instances of the same EC can be converted in a fully automatic manner. In other words, having the ECs consisting of MWEs with the same pattern, it requires some manual work to convert one instance of each EC into a system specific representation, but all other members of the same EC can be done fully automatically. In the current approach, a formal representation of the patterns has been added to the pattern descriptions. Since this formal representation is in agreement with a de facto standard for Dutch (van Noord et al., 2006), most Dutch NLP systems are able to use it for the conversion procedure, yielding an optimal reduction of manual labor.

The creation of MWE descriptions is a very time-consuming task and of course we aim at an error-free result. Accordingly, we decided to describe the minimal ingredients of an MWE that are needed for successful incorporation in any Dutch NLP system. For the development of the representation two Dutch parsers are consulted, viz. the Alpino parser and the Rosetta MT system (Rosetta, 1994).

Another requirement of the lexicon structure is that the information needed for the representation is extractable from corpora, since we want to avoid analyses entirely based on speaker-specific intuitions.

3.1 Subclasses

Each MWE in the lexicon is classified as either fixed, semi-flexible or flexible. In general, our classification conforms to the categorization given in Sag et al. (2001), any differences are explicitly discussed below.

3.1.1 Fixed MWEs

Fixed MWEs always occur in the same word order and there is no variation in lexical item choice. Fixed MWEs cannot undergo morpho-syntactic variation and are contiguous, i.e. no other elements can intervene between the words that are part of the fixed MWE. Examples of Dutch fixed MWEs are: *ad hoc*, *ter plaatse* 'on the spot', *van hoger hand* 'from higher authority'.

3.1.2 Semi-flexible MWEs

The following characteristics are applicable to the class of semi-flexible MWEs in our lexicon:

1. The lexical item selection of the elements of the expression is fixed or very limited.

2. The expression can only be modified as a whole.[3]

3. The individual components can inflect, unless explicitly marked otherwise with a parameter.

Examples of Dutch semi-flexible MWEs are: *de plaat poetsen* (lit. 'to polish the plate', id. 'to clear off'), *witte wijn* 'white wine', *bijvoeglijk naamwoord* 'adjective'.

[2] http://odur.let.rug.nl/∼vannoord/alp.

[3] We abstract away from the reason why some external modifiers, such a *proverbial* in *he kicked the proverbial bucket*, may intrude in these semi-flexible expressions.

The characteristics of this class differ on one point from the characteristics of the semi-fixed class discussed in Sag et al. (2001), viz. on the fact that according to Sag et al. semi-fixed expressions are not subject to syntactic variability and the only types of lexical variation are inflection and variation in the reflexive form. This degree of fixedness does not apply to our class of semi-flexible MWEs, i.e. in Dutch (and also in other Germanic languages like German), operations that involve movement of the verb such as verb second, verb first and verb raising, see (1)-(3), are also applicable to the class of semi-flexible expressions (Schenk, 1994).

(1) Hij poetste de plaat.
he polished the plate
'He cleared off.'

(2) Poetste hij the plaat?
polished he the plate
'Did he clear off?'

(3) ... omdat hij de plaat wilde poetsen.
... because he the plate wanted polish
'... because he wanted to clear off'

3.1.3 Flexible MWEs

The main characteristic of flexible MWEs is the fact that, contrary to semi-flexible MWEs, the individual components within flexible MWEs can be modified. This contrast accounts for differences between *de plaat poetsen* versus *een bok schieten* (lit. 'to shoot a male-goat', id. ' to make a blunder') and *blunder maken/begaan* (' to make a blunder'). Although both *een bok schieten* and *blunder maken/begaan* are flexible MWEs, there is a difference between the two expressions. According to the classification proposed by Sag et al. (2001), *een bok schieten* is a decomposable idiom, of which the individual components cannot occur independently in their idiomatic meaning and *een blunder maken* is a support verb construction. We also want to use this classification, and represent these expressions as follows:

1. Expressions of which one part is lexically fixed and the other part is selected from a list of one or more co-occuring lexemes. Dutch examples are: *scherpe/stevige kritiek* ('severe criticism'), *blunder maken/begaan*.
2. Expressions of which the lexical realization of each component consists of exactly one lexeme. A Dutch example is *een bok schieten*.

The difference between the two subtypes is made visible in the representation of the MWE and the MWE pattern.

3.2 The data

We use data extracted from the Twente Nieuws Corpus (TwNC) (Ordelman, 2002) as empirical material.[4] This corpus comprises a 500 million words of newspaper text and television news reports. From the TwNC, a list of candidate expressions is extracted, including for each expression the following properties:

- the pattern assigned to the expression by the Alpino parser
- the frequency
- the head of the expression
- the ten most occurring subjects
- internal complements and for each complement: its head, the head of the complement of the head (in the case of PP complements), its dependency label assigned by Alpino, the number of the noun, whether the noun is positive of diminutive, the ten most occurring determiners, the ten most occurring premodifiers, and the ten most occurring postmodifiers.
- six examples sentences

The use of corpora is necessary but not sufficient. It is necessary because we want our lexicon to reflect actual language usage and because we do not want to restrict ourselves to a linguist's imagination of which uses are possible or actually occur. On the other hand, using the corpora to extract the MWEs is not sufficient for the following reasons: (1) text corpora may contain erroneous usage, and the technique used cannot distinguish this from correct usage; (2) the extraction is in part based on an automatic syntactic parse of the corpus sentences, and these parses may be incorrect; (3) the

[4]The identification of MWEs is done by Begoña Villada Moirón working at the University of Groningen.

extraction techniques cannot distinguish idiomatic versus literal uses of word combinations; (4) the extraction techniques group different expressions that share some but not all words together. Therefore the data extracted were carefully analyzed before creating entries for MWEs.

3.3 The lexical represention

3.3.1 Pattern description

In the *Lexicon of Dutch MWEs*, expressions are classified according to their pattern. In the original ECM the pattern is an identifier which refers to the structure of the idiom represented as free text in which the uniqueness of the pattern is described. This description includes the syntactic category of the head of the expression, the complements it takes and the description of the internal structure of the complements. Furthermore it is described whether individual components can be modified. In the current approach the description of the pattern contains besides a textual description also a formal notation, see (4).

(4) Expressions headed by a verb, taking a fixed direct object consisting of a determiner and a noun – [.VP [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)]]

The notation used to describe the patterns is a formalization of dependency trees, in particular CGN (*Corpus Gesproken Nederlands* 'Corpus of Spoken Dutch') dependency trees (Hoekstra et al., 2003). CGN dependency structures are based on traditional syntactic analysis described in the Algemene Nederlandse Spraakkunst (Haeseryn et al., 1997) and are aimed to be as theory neutral as possible.

The patterns are encoded using a formal language, which is short and which allows easy visualization of dependency trees. The dependency labels (in lower case) and category labels (in upper case) are divided by a colon (:), e.g. *obj1:NP*. For leaf nodes, the part-of-speech is represented instead of the category label. Leaf nodes are followed by an index that refers to the MWE component as represented in the CL-field (see section 3.3.2), e.g. (1) refers to the first component of the CL, (2) to the second, etc.

A fixed expression can be represented in two ways depending on its internal structure:

1. For fixed expressions that are difficult to assign an internal structure, we introduced a label *fixed*. The pattern for expressions such as *ad hoc* and *ter plaatste* is [.:Adv fixed(1 2)]

2. Fixed expressions with an analyzable internal structure are represented according to the normal pattern description rules:

(5) *de volle buit* ('everything')
[.NP [.det:D (1)] [.mod:A (2)] [.hd:N (3)]]

Semi-flexible MWEs are also represented according to normal pattern description rules. To make a distinction between (1) an NP of which all elements are fixed, and (2) an NP of which some elements are lexically fixed, but which is still subject to standard grammar rules, a new syntactic category *N1* has been introduced. N1 indicates that the expression can be modified as a whole and can take a determiner as specifier:

(6) *witte wijn*
[.N1 [.mod:A (1)] [.hd:N (2)]]

The pattern of flexible expressions of which the lexical realization of each component consists of exactly one lexeme is encoded using the syntactic category N1. We can use the same category as in (6), since what we want to describe is the fact that the components in the NP are fixed, but can be modified as a whole and can take a determiner as specifier.

(7) *bok schieten*
[.VP [.obj1:N1 [.hd:N (1)]] [.hd:V (2)]]

Expressions of which one part is fixed and the other part is selected from a list of one or more cooccuring lexemes are represented with a so-called LIST-index in the pattern. The fixed part of the expression has its literal sense. The combination of the literal part with other lexemes is not predicable from the meaning of the combining lexeme. Since the meaning of an MWE or its parts is not included in the representation, we can list every single component with which the fixed part can combine in the same MWE entry. For this list of components we created a LISTA-field and LISTB-field in the MWE description. Lists and variables are represented similar to MWE components, attached to the leaf node, in lower case and between (), e.g. [.hd:X (list)], [obj1:NP (var)], [obj2:NP (var)], etc.:

(8) *iemand de helpende hand bieden* (lit. 'offer s.o. the helping hand', id. 'lend s.o. a hand') [.VP [.obj2:NP (var)] [.obj1:NP [.det:D (1)] [.mod:A (2)] [.hd:N (3)]] [.hd:V (4)]]

Our characterization of the classes of MWEs and the formal notation of the patterns do not fully cover the range of different types of MWEs that are described in the lexicon. The strength of the ECM is, however, that any expression can be included in the lexicon, regardless of whether it fits our classification, because of the textual description that can be assigned. Expressions that cannot be assigned a dependency structure, because of the limitations of the notation, are classified according to the textual description of its pattern. A revision of the formal notation might be done in the future.

The pattern is part of the MWE pattern description which includes, besides a pattern name, a pattern and a textual description, five additional fields, which are both maintenance field and fields needed for a successful implementation of the standard representation into a system specific representation. Examples of MWE pattern descriptions stored in the *Lexicon of Dutch MWEs* are given in Table 1.

3.3.2 MWE description

In addition to the MWE pattern descriptions, the lexicon contains MWE descriptions, see Table 2 for a list of examples. An MWE description comprises 8 description fields. The PATTERN_NAME is used to assign an MWE pattern description to the expression. The EXPRESSION-field contains the obligatory fixed components of an MWE in the full form.

The Component List (CL) contains the same components as the EXPRESSION-field. The difference is that the components in the CL are in the canonical (or non-inflected) form, instead of in the full form. Parameters are used to specify the full form characteristics of each component. The term *parameter* is a feature and can be defined as an occurrence of the pair <parameter category,parameter value>, where *parameter category* refers to the aspect we parameterize, and *parameter value* to the value a parameter category takes. Examples of parameters are <nnum,sg> for singular nouns, <afrm,sup> for superlative adjectives, <vfrm,part> for particle verbs (Grégoire, 2006). Parameter values are realized between square brackets directly on the right of the item they parameterize.

The LISTA-field and LISTB-field are used to store components that can be substituted for the LIST-index in the pattern, yielding one or more expressions. The reason for using two LIST-fields is to separate predefined list values from special list values. The predefined list values are high frequent verbs that are known to occur often as so-called light verbs, especially with PPs. Two sets of verbs are predefined:

1. blijken ('appear') blijven ('remain') gaan ('go') komen ('come') lijken ('appear') raken ('get') schijnen ('seem') vallen ('be') worden ('become') zijn ('be')

2. brengen ('bring') doen ('do') geven ('give') hebben ('have') houden ('keep') krijgen ('get') maken ('make') zetten ('put')

A complement co-occurs either with verbs from set 1 or with verbs from set 2. Each verb from the chosen set is checked against the occurrences found in the corpus data. If a verb does not occur in the corpus data and also not in self-constructed data, it is deleted from the LISTA-field. The LISTB-field contains lexemes that are not in the predefined set but do co-occur with the component(s) in the EXPRESSION-field. The information in the LISTB-field is merely based on corpus data and therefore may not be exhaustive.

The EXAMPLE-field contains an example sentence with the expression. The only requirement of this field is that its structure is identical for each expression with the same PATTERN_NAME. The POLARITY-field is *none* by default and takes the value *NPI* if an expression can only occur in negative environments, and *PPI* if an expression can only occur in positive environments. Finally, the MWE description contains a field with a reference to a plain text file in which the information extracted from the corpora is stored.

4 Discussion

We have given an overview of the decisions made in order to come to a standard lexical representation for Dutch MWEs and discussed the description

NAME	PATTERN	DESCRIPTION
EC1	[.VP [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)]]	Expressions headed by a verb, taking a fixed direct object contisting of a determiner and a noun.
EC2	[.VP [.obj1:N1 [.hd:N (1)]] [.hd:V (list)]]	Expressions headed by a verb, taking a direct object consisting of a fixed modifiable and inflectable noun (list).
EC9	[.VP [.obj1:N1 [.hd:N (1)]] [.hd:V (list)] [.pc:PP [.hd:P (2)] [obj1:NP (var)]]]	Expressions headed by a verb, taking (1) a direct object consisting of a fixed modifiable noun, and (2) a PP-argument consisting of a fixed preposition and a variable complement (list).

Table 1: List of MWE pattern descriptions.

PATTERN	EXPRESSION	CL	LIST
EC1	zijn kansen waarnemen (‘to seize the opportunity’)	zijn kans[pl] waarnemen	-
EC2	blunder (‘mistake’)	blunder	begaan (‘commit’) maken (‘make’)
EC9	kans op (‘to stand a change of s.th.’)	kans op	lopen (‘get’) maken

Table 2: List of MWE descriptions.

fields this representation comprises. Contrary to related work, we did not solely focus on one type of MWEs, but on MWEs in general. The *Lexicon of Dutch MWEs* includes 5,000 unique expressions and for an initial version a conversion to the Dutch NLP system Alpino has been tested. The strength of our method lies in the ability of grouping individual expressions according to their pattern, yielding multiple classes of MWEs. The advantage of creating classes of MWEs is that it eases the conversion of the standard representation into any system specific representation.

Describing a class of MWEs using free text is already very useful in its current form. To help speeding up the process of converting the standard representation into a system specific representation, we introduced a formal notation using dependency structures, which are aimed to be as theory neutral as possible. However, our current notation is unable to cover all the patterns described in the lexicon. The notation can be extended, but we must make sure that it does not become too ad hoc and more complicated than interpreting free text.

We have created a resource that is suited for a wide variety of MWEs. The resource describes a set of essential properties for each MWE and classifies each expression as either fixed, semi-flexible or flexible. The set of properties can surely be extended, but we have limited ourselves to a number of core properties because of resource limitations. We are confident that this resource can form a good basis for an even more complete description of MWEs.

5 Conclusion

This paper described the design and implementation of a lexicon of Dutch multiword expressions. No exhaustive research on a standard representation of MWEs has been done for Dutch before. Data extracted form large Dutch text corpora were used as empirical material. The approach taken is innovative, since it is based on the Equivalence Class Method (ECM). The ECM focuses on describing MWEs according to their pattern, making it possible to form classes of MWEs that require the same treatment in natural language processing. The *Lexicon of*

Dutch MWEs constitutes 5,000 unique expressions and for an initial version of the lexicon a conversion to the Dutch NLP system Alpino has been tested.

Acknowledgements

The IRME project is carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (http://taalunieversum.org/stevin).

The author would like to thank Jan Odijk and two anonymous reviewers for their valuable input to this paper.

References

Hans de Groot. 1999. *Van Dale Idioomwoordenboek*. Van Dale Lexicografie, Utrecht.

Ricarda Dormeyer and Ingrid Fischer. 1998. Building lexicons out of a database for idioms. In Antonio Rubio, Nativiad Gallardo, Rosa Castro, and Antonio Tejada, editors, *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 833 – 838.

Christiane Fellbaum, Alexander Geyken, Axel Herold, Fabian Koerner, and Gerald Neumann. 2006. Corpus-Based Studies of German Idioms and Light Verbs. *International Journal of Lexicography*, 19(4):349–361.

Nicole Grégoire. 2006. Elaborating the parameterized equivalence class method for dutch. In Nicoletta Calzolari, editor, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1894–99, Genoa, Italy. ELRA.

W. Haeseryn, K. Romijn, G. Geerts, J. de Rooij, and M.C. van den Toorn. 1997. *Algemene Nederlandse Spraakkunst*. Martinus Nijhoff and Wolters Plantyn, Groningen en Deurne.

Heleen Hoekstra, Michael Moortgat, Bram Renmans, Machteld Schouppe, Ineke Schuurman, and Ton van der Wouden. 2003. Cgn syntactische annotatie.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties.*, Sydney, Australia.

Brigitte Krenn. 2000. CDB - a database of lexical collocations. In *2nd International Conference on Language Resources & Evaluation (LREC '00), May 31 - June 2*, Athens, Greece. ELRA.

Koenraad Kuiper, Heather McCann, Heidi Quinn, Therese Aitchison, and Kees van der Veer. 2003. SAID: A syntactically annotated idiom dataset. Linguistic Data Consortium, LDC2003T10, Pennsylvania.

Willy Martin and Isa Maks. 2005. Referentie bestand nederlands documentatie. Technical report, INL.

Jan Odijk. 2004a. Multiword expressions in NLP. Course presentation, LOT Summerschool, Utrecht, July.

Jan Odijk. 2004b. A proposed standard for the lexical representation of idioms. In *EURALEX 2004 Proceedings*, pages 153–164. Université de Bretagne Sud, July.

R.J.F. Ordelman. 2002. Twente nieuws corpus (TwNC).

M T. Rosetta. 1994. *Compositional Translation*. Kluwer Academic Publishers, Dordrecht.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for NLP. LinGO Working Paper, (2001-03).

André Schenk. 1994. *Idioms and collocations in compositional grammars*. Ph.D. thesis, University of Utrecht.

Gertjan van Noord, Ineke Schuurman, and Vincent Vandeghinste. 2006. Syntactic annotation of large corpora in stevin. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa - Italy.

Begona Villada Moirón and Joerg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context*, Trento, Italy.

Aline Villavicencio, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. The lexical encoding of MWEs. In T. Tanaka, A. Villavicencio, F. Bond, and A. Korhonen, editors, *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proceedings of the ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties.*, Sydney, Australia.