Multiword Expressions and Named Entities in the Wiki50 Corpus

Veronika Vincze[1], **István Nagy T.**[2] and **Gábor Berend**[2]
[1]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu
[2]Department of Informatics, University of Szeged
{nistvan,berendg}@inf.u-szeged.hu

Abstract

Multiword expressions (MWEs) and named entities (NEs) exhibit unique and idiosyncratic features, thus, they often pose a problem to NLP systems. In order to facilitate their identification we developed the first corpus of Wikipedia articles in which several types of multiword expressions and named entities are manually annotated at the same time. The corpus can be used for training or testing MWE-detectors or NER systems, which we illustrate with experiments and it also makes it possible to investigate the co-occurrences of different types of MWEs and NEs within the same domain.

1 Introduction

In natural language processing (NLP), a challenging task is the proper treatment of multiword expressions (MWEs). Multiword expressions are lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag et al., 2002; Kim, 2008; Calzolari et al., 2002) thus, they often pose a problem to NLP systems.

Named Entity Recognition (NER) is another widely researched topic in NLP. There are several methods developed for many languages and domains, which are tested on manually annotated databases, e.g. the MUC-6 and MUC-7 and the CoNLL-2002/2003 challenges aimed at identifying NEs in newswire texts (Grishman and Sundheim, 1995; Chinchor, 1998; Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). Multiword named entities can be composed of any words or even characters and their meaning cannot be traced back to their parts. For instance, *Ford Focus* refers to a car and has nothing to do with the original meaning of *ford* or *focus*, thus, it is justifiable to treat the whole expression as one unit.

In this paper, we present our corpus called Wiki50 which contains 50 Wikipedia articles annotated for multiword expressions and named entities. To the best of our knowledge, this is the first corpus in which MWEs and NEs are annotated at the same time. We describe the categories occurring in the database, provide some statistical data on their frequency and finally, we demonstrate how noun compounds and named entities can be automatically detected by applying some dictionary-based and machine learning methods.

2 Related corpora and databases

Several corpora and databases of MWEs have been constructed for a number of languages. For instance, Nicholson and Baldwin (2008) describe a corpus and a database of English compound nouns (BNC dataset in Table 1). As for multiword verbs, corpora and databases for English (Cook et al., 2008), German (Krenn, 2008), Estonian (Muischnek and Kaalep, 2010) and Hungarian (Vincze and Csirik, 2010) have been recently developed. The Prague Dependency Treebank is also annotated for multiword expressions (Bejcek and Stranák, 2010).

As for named entities, several corpora have been constructed, for instance, within the framework of the ACE project (Doddington et al., 2004) and for international challenges such as the CoNLL-2002/2003 datasets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) or the MUC datasets (Grishman and Sundheim, 1995; Chinchor, 1998) – just to name a few.

As can be seen, although there are a number of corpora and databases annotated for MWEs, they typically focus on only one specific type of MWE. That is, there are hardly any corpora that contain manual annotation for several types of English MWEs at the same time. On the other hand, to the best of our knowledge, there exist no corpora where various types of MWEs are an-

Proceedings of Recent Advances in Natural Language Processing, pages 289–295,
Hissar, Bulgaria, 12-14 September 2011.

notated together with NEs. Named entities often consist of more than one word, i.e. they can be seen as a specific type of multiword expressions / noun compounds (Jackendoff, 1997). Although both noun compounds and multiword named entities consist of more than one word, they form one semantic unit and thus, they should be treated as one unit in NLP systems. Taking the example of POS-tagging, the linguistic behavior of compound nouns and multiword NEs is the same as that of single-word nouns, thus, they are preferably tagged as nouns (or proper nouns) even if the phrase itself does not contain any noun (e.g. *has-been* or *Die Hard*). Once identified as such, they can be treated similarly to single words in syntactic parsing for example. Our corpus makes it possible – for the first time – to compare (co-)occurrences of different types of MWEs and NEs and to evaluate the performance of MWE-detectors and NER systems within the same domain.

3 The Wiki50 corpus

When constructing our corpus, we selected 50 random articles from the English Wikipedia. The only selectional criterion applied was that each article should consist of at least 1000 words and they should not contain lists, tables or other structured texts (i.e. only articles with running texts were included). In this section, we present the types of multiword expressions and named entities annotated in our corpus.

3.1 Multiword expressions

A **compound** is a lexical unit that consists of two or more elements that exist on their own. Orthographically, a compound may include spaces (*high school*) or hyphen (*well-known*) or none of them (*headmaster*). We annotated only nominal and adjectival compounds in the database since they are productive and cannot be identified with lists. Our main goal being to develop a corpus for evaluating MWE detectors, we annotated only compounds with spaces since hyphenated compounds (e.g. *self-esteem*) can be easily recognized and are not included in our definition of multiword expressions (i.e. 'words with spaces').

Verb-particle constructions (VPCs, also called phrasal verbs or phrasal-prepositional verbs) are combined of a verb and a particle/preposition (see e.g. Kim (2008)). They can be adjacent (as in *put off*) or separated by an intervening object (***turn** the light **off***). They can be compositional, i.e. it can be computed from the meaning of the preposition and the verb (*lie down*) or non-compositional (*do in* meaning "kill"). VPCs are also marked in the database and their respective parts (i.e. verb and particle) are also annotated in order to facilitate the automatic detection of constructions where the two parts are not adjacent (e.g. ***spit** it **out***).

An **idiom** is a MWE whose meaning cannot (or can only partially) be determined on the basis of its components (Sag et al., 2002; Nunberg et al., 1994). Although most idioms behave normally as syntax and morphology are concerned, i.e. they can undergo some morphological change (e.g. *He spills/spilt the beans*), their semantics is totally unpredictable. **Proverbs** express some important facts thought to be true by most people, e.g. *The early bird catches the worm.* Idioms and proverbs are both annotated in our corpus.

Light verb constructions (LVCs) consist of a nominal and a verbal component where the noun is usually taken in one of its literal senses but the verb usually loses its original sense to some extent e.g. *to give a lecture*, *to come into bloom*, *the problem lies (in)*. The nominal and the verbal component of such constructions are also marked within the light verb construction (hierarchical annotation) for they can be separated from each other within context (e.g. in passive sentences).

There are **other types of MWEs** that do not fit into the above categories (some of them are listed in Jackendoff (1997)) such as *status quo*, *c'est la vie* and *ad hoc*. Although they are composed of perfectly meaningful parts in the original language, in English, these words do no exist on their own hence it is impossible to derive their meaning from their parts and the expression must be stored as a whole. They are also labeled as MWEs in the database. So are multiword verbs that cannot be classified as VPCs or LVCs (e.g. *to voice act* or *drink and drive*).

3.2 Named entities

In our database, the four basic types of named entities are marked: persons (PER), organizations (ORG), locations (LOC) and miscellaneous (MISC). We applied tag-for-meaning annotation in the corpus, that is, occurrences of e.g. country names could refer to an organization and a loca-

Corpus	Sentence	Token
CoNLL-2003	14,987	203,621
Wikipedia	4,350	114,570
BNC dataset	1000	21,631

Table 1: Size of various NE and MWE annotated corpora in terms of sentence and token number

tion as well depending on the context, thus, they were classified as belonging to different categories in such cases.

3.3 Segmentation of data

Sentence boundaries were also manually annotated in the database: sentences ending with an abbreviated form (e.g. *He lives in L.A.*), where the full stop belongs to the named entity and marks the sentence boundary at the same time, are distinctively marked.

The corpus exists in two versions: in the distilled version, segmentation errors (e.g. missing spaces) were corrected manually, and irrelevant parts of the documents (e.g. references or footnotes) were filtered. The other version – being more noisy – can be used in web-mining applications since no such modifications were carried out on the texts collected from the web. Both versions of the corpora are available under the Creative Commons license at `http://rgai.inf.u-szeged.hu/mwe`.

3.4 Statistics on corpus data

In the following, some statistical data on the distilled version of the corpus are provided. The corpus consists of 4350 sentences and 114,570 tokens, which size makes it comparable to other existing corpora (see Table 1). Table 2 summarizes the number of occurrences of the annotated categories and the number of unique phrases (i.e. no multiple occurrences are counted here) as well as the average and variance of the number of the various annotations per token.

4 The process of annotation and error analysis

Two linguists carried out the annotation of the corpus. 15 articles out of the 50 were annotated by both of them and differences were later resolved. The agreement rates between the two annotators are represented in Table 3.

As the data show, NEs in general are easier to identify for humans than MWEs. Among MWEs, note that for many categories it was mostly recall that was responsible for the decrease in F-measure. This can be related to the complexity of the annotation task: in 4350 sentences 12,832 elements were marked (i.e. 2.95 elements per sentence), not including the hierarchical categories, which probably led to the fact that annotators were prone to overlook certain expressions in running text. It was especially true for VPCs and MWEs classified as 'other' – VPCs typically consist of short elements, which may make it hard to recognize them in running text. The high precision value of the VPC class suggests that this category is relatively easy to classify (if recognized in text). This is somewhat different for NEs: here the capitalization may be an important feature in detecting NEs while reading, which causes that NEs were almost always annotated (i.e. recall values are higher) hence their agreement rates are higher.

Category	Occurrence	Unique	Avg. frequency
Noun Comp.	2929	2405	0.0263±2.1E-4
Adj. Comp.	78	60	0.0008±1.1E-6
VPC	446	342	0.0038±8.9E-6
LVC	368	338	0.0030±4.9E-6
Idiom	19	18	0.0002±1.2E-7
Other	21	17	0.0002±8.1E-8
MWE sum	3861	3180	0.0342±2.0E-4
PER	4093	1533	0.0352±5.8E-4
ORG	1498	893	0.0133±2.0E-4
LOC	1558	705	0.0150±2.5E-4
MISC	1827	952	0.0166±2.3E-4
NE sum	8976	4083	0.0801±7.5E-4

Table 2: Identified occurrences of categories in the corpus and their relative per-token frequencies

It seems that frequent MWE categories reach higher agreement rates than rare ones (cf. Table 2). In the latter case with only a few tens of examples, one single erroneous annotation or lack of annotation weighed much more than in cases when several hundreds (or even thousands) of examples could be found. As opposed to MWEs there were no underrepresented NE categories, which yields that the overall agreement rate calculated for NEs is higher than that of MWEs.

The κ-measure of the whole annotation (i.e. including NEs and MWEs) is 0.6938, which can be considered as a fairly good agreement rate.

4.1 Errors in MWE annotation

MWE annotation errors can be classified into two groups. The lack of annotation by one of the annotators may be related to conceptual differ-

Category	Precision	Recall	F-score	Jaccard	κ-measure
Noun Comp.	0.7135	0.7089	0.7112	0.5518	0.6414
Adj. Comp.	0.5625	0.4286	0.4865	0.3214	0.4841
VPC	0.8831	0.5620	0.6869	0.5231	0.6792
LVC	0.7454	0.6721	0.7069	0.5467	0.6980
Idiom	0.5556	0.5556	0.5556	0.3846	0.5545
Other	1.0	0.1429	0.25	0.1429	0.2497
MWE sum	0.7320	0.6816	0.7059	0.5329	0.5797
PER	0.9794	0.9802	0.9798	0.9605	0.9708
ORG	0.8322	0.7515	0.7898	0.6526	0.7716
LOC	0.9042	0.9103	0.9073	0.8303	0.8953
MISC	0.8921	0.8986	0.8953	0.8105	0.8781
NE sum	0.9635	0.9544	0.9589	0.8603	0.6789

Table 3: Agreement rates between annotators

ences (e.g. hyphenated noun compounds were not to be marked, however, one annotator occasionally marked phrases like *brother-in-law* as noun compounds) and lack of attention: the annotator simply did not recognize one instance of an element annotated elsewhere in the text. A typical example for the latter case is VPCs, which are usually short and therefore it is hard to catch them in running text for the human annotator. Concerning the second group of errors, here the same expression was marked with two different labels. Interestingly, a common source of error was that certain elements were annotated as noun compounds by one of the annotators and as named entities by the other annotator such as Latinate or botanical names (*Torrey Pine*), buildings (*City Hall*), names of positions or committees (*Board of Trustees*) etc. These issues might be eliminated with more detailed annotation guidelines, however, all of these mismatches were later disambiguated, yielding the gold standard annotation of the corpus.

4.2 Errors in NE annotation

In NE annotation, most of the cases where only one of the annotators marked the phrase were related to fictional objects. Many articles described a video game or a fantasy world in which a lot of special objects (e.g. *Trap Cards*) were annotated as miscellaneous by one of the annotators while the other did not mark them. On the other hand, different labels were also assigned to the same phrases. Besides NE-MWE differences, certain NEs were annotated as different NE-subtypes, which is partly connected to metonymic annotation. For instance, names of countries or states could be annotated as locations and organizations according to context but sometimes the two annotators did not agree on whether it should be marked as a location or an organization. Another example was person-like fictional characters (e.g. *Zombie Werewolf*): one annotator labeled them as a person while the other as miscellaneous. Again, these cases are hard to determine without deeper knowledge of the story and more refined guidelines are necessary for their annotation.

4.3 Nested expressions

A special type of annotation differences concerned nested expressions. A multiword expression may contain another multiword expression (***carbon monoxide*** *leak*), a named entity may include another named entity (***New York*** *City*) or a MWE may include a NE (***FBI*** *special agent*) and may be part of an NE (*Tallulah* ***High School***). Although it was assumed that in each case the longest unit is marked, sometimes this principle was not observed by one of the annotators, which resulted in annotation errors. This issue may be resolved with hierarchical annotation where elements within the longest unit are also annotated, which we plan to carry out in the future.

5 Experiments

In this section, we describe our dictionary-based and machine learning based approaches to identify noun compounds and named entities in the corpus.

5.1 Dictionary based approaches

We used several Wikipedia-based approaches to automatically identify noun compounds, which

Threshold	Match	Merge	POS rules	Combined
100	0.2915	0.3093	0.2742	0.2901
50	0.3391	0.3599	0.3289	0.3479
20	0.4133	0.4332	0.4104	0.4295
10	0.4560	0.4751	0.4597	0.4801
5	0.4715	0.4883	0.4872	0.5051
2	0.4749	0.4976	0.5158	0.5420
1	0.4528	0.4751	0.5334	0.5609

Table 4: Effect of various heuristics using dictionary based methods

were evaluated on the above described corpus. Our methods were motivated by the encyclopedic nature of Wikipedia: as opposed to dictionaries, it mostly contains nominal concepts. Thus, we assumed that by using internal links of Wikipedia[1], a list of possible noun compounds can be gathered. This list consists of the anchor texts of all internal links with their frequencies (how many times this text span occurred as a link) comprising 2-4 lowercase tokens. The list was later filtered for special (non-English) characters and words not typical of noun compounds (such as auxiliaries or quantifiers).

In the first approach we marked a phrase as a noun compound if it occurred in the list and its frequency exceeded the current threshold (Match). In the second case, we assumed that if *a b* is a possible noun compound and *b c* too, they can be merged, so *a b c* is also a noun compound. In this case, *a b c* was only accepted as a noun compound if *a b* and *b c* occurred in the list and the frequency of the whole phrase exceeded the current threshold (Merge). As for the the third approach, several part-of-speech based patterns such as `JJ (NN|NNS)` were created. A potential noun compound in the text was accepted if it appeared in the list, its POS code sequence matched one of the patterns and its frequency exceeded the current threshold (POS rules). POS codes were determined using Stanford POS Tagger (Toutanova and Manning, 2000). Finally, we combined these approaches: we accepted a potential noun compound if it appeared in the list, its POS code sequence matched our patterns, furthermore, merges were allowed too (Combined). Results in terms of F-measure are shown in Table 4, from which it can be seen that best results were obtained when all the above mentioned extensions were combined and no frequency threshold was considered.

[1]Articles included in our corpus were not considered when collecting links.

leave-one-out	R	P	F
MWE	58.07	69.86	63.42
MWE + NE	65.65	72.44	68.68
NE	85.58	86.02	85.81
NE + MWE	87.07	87.28	87.18

Table 5: Results of leave-one-out approaches in terms of precision (P), recall (R) and F-measure (F). MWE: our CRF extended with automatically collected MWE dictionary, MWE+NE: our CRF with MWE features extended with NEs as feature, NE: our CRF trained with basic feature set, NE+MWE: our CRF model extended with MWEs as feature.

5.2 Machine Learning approaches

In addition to the above-described approach, we defined another method for automatically identifying noun compounds. The Conditional Random Fields (CRF) classifier was used (MALLET implementations (McCallum, 2002)) with the feature set used in Szarvas et al. (2006). To identify noun compound MWEs we used Wiki50 to train CRF classification models (they were evaluated in a leave-one-document-out scheme). Results are shown in the *MWE* row of Table 5.

In order to use Wiki50 only for testing purposes, we automatically generated a train database for the CRF trainer. The train set consists of 5,000 randomly selected Wikipedia pages and we ignored those containing lists, tables or other structured texts. Since this document set has not been manually annotated, dictionary based noun compound labeling was considered as the gold standard. As a result, we had a less accurate but much bigger training database. The CRF model was trained on the automatically generated train database with the above presented feature set. Results can be seen in the *CRF* row of Table 6. However, the database included many sentences without any labeled noun compounds hence negative examples were overrepresented. Therefore, we thought it necessary to filter the sentences: only those with at least one noun compound label were retained in the database (*CRF+SF*). With this filtering methodology the CRF could build a better model. The above-described feature set was completed with the information that a token is a named entity or not. The *MWE+NE* row of Table 5 shows that this feature proved very effective in the leave-one-document-out scheme, so we used it in the auto-

Approach	R	P	F
DictCombined	52.47	59.45	55.75
CRF	44.38	58.42	50.44
CRF+SF	53.39	56.66	54.98
CRF+NE	45.81	58.37	51.33
CRF+NE+SF	53.12	55.89	54.47
CRF+OwnNE+SF	53.29	57.60	55.36
CRF+OwnNELeft+SF	53.44	57.60	55.44
CRF+MWELeft+SF	53.53	58.74	56.02

Table 6: Results of different methods for noun compounds in terms of precision (P), recall (R) and F-measure (F). DictCombined: combination of dictionary based methods, CRF: our CRF model trained on automatically generated database, SF: sentences without any MWE label filtered, NE: NEs marked by Stanford NER used as feature, OwnNE: NEs marked by our CRF model (trained on Wikipedia) used as feature, OwnNELeft: the NE labeling selected as feature and the standard noun compound notation removed, MWELeft: the NE feature deleted and the standard noun compound notation selected.

matically generated train database too. As shown in the *CRF+NE* row of Table 6, the CRF model which was trained on the automatic training set could achieve better results with this feature than the original *CRF*.

First, the Stanford NER model was used for identifying NEs. However, we assumed that a model trained on Wikipedia could identify NEs more effectively in Wikipedia (i.e. in the same domain). Therefore, we merged the four NE classes marked in Wiki50 into one NE class to train the CRF with the above described common features set. Results are shown in the *NE* row of Table 5.

The *CRF+OwnNE+SF* row in Table 6 represents results achieved when the NEs identified by using the entire Wiki50 as the training dataset functioned as a feature. Although the *CRF+NE+SF* (when NEs were identified by the Stanford model) did not achieve better results than the *CRF+SF*, our Wikipedia based CRF model to identify NEs in the automatically generated training dataset (*CRF+OwnNE SF*) yielded better F-score than *CRF+SF*, which means that NEs are useful in the identification of noun compounds.

Sometimes it was not unequivocal to decide whether a multiword unit is a noun compound or a NE. However, we assumed that a term can occur either as a NE or a noun compound. Therefore, if the dictionary method marked a particular word as noun compound and the NE model also marked it as NE, we had to decide which mark to delete. The *CRF+OwnNELeft+SF* row in Table 6 shows results we achieved if the NE labeling was selected as feature and the standard noun compound notation was removed, whereas the row *CRF+MWELeft+SF* refers to the scenario when the NE feature was deleted, and the standard noun compound notation remained.

5.3 Named Entity Recognition with MWEs

We investigated the usability of noun compounds in named entity recognition. So we used Wiki50 to train CRF classification models with the basic feature set, which was extended with the feature noun compound MWE for NE recognition and they were evaluated in a leave-one-document-out scheme. Results of these approaches are shown in the *NE+MWE* row of Table 5. Comparing these results to those of the *NE* method (when the CRF was trained without the noun compound feature), noun compounds are also beneficial in NE identification.

6 Discussion

For identifying noun compounds we examined dictionary based and machine learning based methods too. The approaches we applied heavily rely on Wikipedia. The dictionary based approach made use of the automatically collected list from Wikipedia. The machine learning method exploited automatically generated training data. These were less accurate but much bigger than the available manually annotated training sets.

Our results demonstrate that previously known noun compounds are beneficial in NER and identified NEs enhance MWE detection. This may be related to the fact that multiword NEs and noun compounds are similar from a linguistic point of view as discussed above – moreover, in some cases, it is not easy to determine even for humans whether a given sequence of words is a NE or a MWE (capitalized names of positions such as *Prime Minister* or taxonomic names, e.g. *Torrey Pine*). In the test databases, no unit was annotated as NE and MWE at the same time, thus, it was necessary to disambiguate cases which could be labeled by both the MWE and the NE systems. By fixing the label of such cases, disambiguity is

eliminated, that is, the training data are less noisy, which leads to better overall results.

7 Conclusions

In this paper, Wiki50, the first corpus in which multiword expressions and named entities are annotated at the same time was presented. Corpus data make it possible to investigate the co-occurrences of different types of MWEs and NEs within the same domain. The corpus consists of 50 Wikipedia articles (4350 sentences) and is freely available for research purposes. We also conducted various experiments on the identification of noun compounds and named entities in the corpus by dictionary-based and machine learning methods as well. We hope that our corpus will enhance the training and testing of MWE-detectors and NER systems.

Acknowledgments

This work was supported by the Project "TÁMOP-4.2.1/B-09/1/KONV-2010-0005 – Creating the Center of Excellence at the University of Szeged", supported by the European Union and co-financed by the European Regional Development Fund and by the project MASZEKER financed by the National Innovation Office of the Hungarian government.

References

Eduard Bejcek and Pavel Stranák. 2010. Annotation of multiword expressions in the Prague Dependency Treebank. *Language Resources and Evaluation*, 44(1-2):7–21.

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of LREC-2002*, pages 1934–1940, Las Palmas.

Nancy A. Chinchor. 1998. Overview of MUC-7/MET-2. In *Proceedings of MUC-7*.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22, Marrakech, Morocco, June.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program – Tasks, Data and Evaluation. In *Proceedings of LREC 2004*, pages 837–840.

Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 evaluation. In *Proceedings of MUC-6*, pages 1–12, Stroudsburg, PA, USA. ACL.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Su Nam Kim. 2008. *Statistical Modeling of Multiword Expressions*. Ph.D. thesis, University of Melbourne, Melbourne.

Brigitte Krenn. 2008. Description of Evaluation Resource – German PP-verb data. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 7–10, Marrakech, Morocco, June.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Kadri Muischnek and Heiki Jaan Kaalep. 2010. The variability of multi-word verbal expressions in Estonian. *Language Resources and Evaluation*, 44(1-2):115–135.

Jeremy Nicholson and Timothy Baldwin. 2008. Interpreting Compound Nominalisations. In *LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 43–45, Marrakech, Morocco.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70:491–538.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of CICLing-2002*, pages 1–15, Mexico City, Mexico.

György Szarvas, Richárd Farkas, and András Kocsor. 2006. A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In *Discovery Science*, pages 267–278.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of EMNLP 2000*, pages 63–70, Stroudsburg, PA, USA. ACL.

Veronika Vincze and János Csirik. 2010. Hungarian corpus of light verb constructions. In *Proceedings of Coling 2010*, pages 1110–1118, Beijing, China. Coling 2010 Organizing Committee.