Identifying Compounds: On The Role of Syntax

Murhaf Fares, Stephan Oepen and Erik Velldal

Language Technology Group, Department of Informatics
University of Oslo
Email: {murhaff|oe|erikve}@ifi.uio.no

Abstract

In this work, we return to a foundational problem related to the interpretation of nominal compounds (in English) that has received comparatively little attention in past research, viz. the identification of instances of the compound construction. We review techniques proposed for this task previously and contrast different approaches along three dimensions of variation, including the contrast of assuming part of speech annotations only vs. using full constituent structure. A first set of quantitative and qualitative experimental results suggest that syntax-based compound identification leads to far better results, at least where gold-standard constituent structures are available.

1 Introduction

In an email among the authors of this paper, one said: "I got a *kitchen update* from Joe." *Kitchen update*, albeit uncommon, is a valid example of nominal compounding in English, where a more typical example could be, say, *apple juice* or *lung cancer*. Downing [4] refers to nominal compounds as "noun plus noun compounds" and adopts the definition by Li [13] as "the concatenation of any two or more nouns functioning as a third nominal." Similarly, our approach defines noun compounds as constructions consisting of two or more nouns that stand in a head–modifier relation.

One of the characteristics of noun compounds is their semantic unpredictability. The aforementioned compound *kitchen update*, for example, may refer to an update about the kitchen status or an update (about whatever) that happened to be given in the kitchen. Furthermore, compounding is a very frequent and productive linguistic process: Baldwin and Tanaka [1] report that 2.6% of the words in the written portion of the British National Corpus (BNC; Burnard [2]) and 3.9% of the Reuters corpus (Rose et al. [20]) are contained in noun–noun compounds. This indicates that a principled and systematic treatment of these constructions will be of potential importance to a wide range of Natural Language Processing (NLP) tasks.

Lauer and Dras [11] identify three tasks related to noun compounds: (1) detection or identification of noun compounds, (2) syntactic analysis of the internal structure, i.e. left vs. right bracketing of compounds with more than two constituents, and finally (3) interpretation of the semantic relation holding between the constituents of the compound. The task of noun compound interpretation has been the focus of many studies (Tratz and Hovy [21], Nakov [16], Ó Séaghdha and Copestake [18]), including several SemEval shared tasks (Girju et al. [7], Butnariu et al. [3], Hendrickx et al. [9]). The bracketing task has also received some attention, either as a separate task (Nakov [16], Pitler et al. [19]) or as part of parsing noun phrases (Vadas and Curran [23]). However, the task of noun compound identification has not received as much attention. This paper presents careful analysis and experimentation directed at the identification task, demonstrating the benefit of using syntactic information. We believe that more accurate noun compound identification will have an effect on the other two tasks of bracketing and interpretation. Further, the three tasks become even more interdependent in the context of our efforts to automatically construct a data set of noun compounds with their semantic interpretation (we will elaborate more on the context of this research in § 6).

In § 2, we briefly review previous work on noun compound identification. In § 3 we define three main variables for noun compound identification strategies. In § 4 we present our approach and experimental setup. In § 5 we report the results of our experiments with a brief analysis. We reflect on the results analysis in § 6, and, finally, in § 7 we conclude the paper.

2 Background

Variations of the heuristic suggested by Lauer [12] comprise some of the most widely used symbolic approaches to noun compound identification (Girju et al. [6], Ó Séaghdha [17], Tratz and Hovy [21]). Lauer [12] defines noun compounds as consecutive pairs of so-called "sure nouns"—nouns that are unambiguous with respect to their part-of-speech (PoS) tags—that are not preceded and not followed by other nouns. Several studies rely on variations of the heuristic of Lauer without mention of the restriction to unambiguous nouns (e.g. Tratz and Hovy [21]). Lauer [12] reports a high precision of 97.9% on a set of 1,068 candidate noun compounds from the Grolier Multimedia Encyclopedia, where an important factor presumably is his limitation of candidate compound constituents to unambiguous nouns.

Lapata and Lascarides [10] evaluated the heuristic of Lauer on the BNC by inspecting a sample of 800 noun sequences classified as valid compounds and report an accuracy of 71%, which is substantially lower than the original results by Lauer [12]. They mention PoS tagging errors when discussing these results.

In the same article, Lapata and Lascarides [10], also introduce statistical models (based on C4.5 decision tree and naïve Bayes learners) to identify noun compounds. They train and test the models on 1,000 noun sequences that occur only once in the BNC, and experiment with different combinations of features and learn-

ers. Their best model attains an accuracy of 72.3%. In addition to surface form statistics, Lapata and Lascarides [10] use PoS tag information, making it similar to the heuristic of Lauer in terms of the type of information used.

Importantly, Lauer [12] already points out that "there is no guarantee that two consecutive nouns form a compound." For example, bare direct and indirect nominal objects of a transitive verb can occur consecutively without forming a noun compound. In fact, some of the studies that used the heuristic of Lauer resorted to manual inspection of the extracted candidate noun compounds to exclude false positives (Girju et al. [7], Ó Séaghdha [17]). In the present paper we investigate the use of syntactic information to identify noun compounds. As explained in § 3, we expect that a richer linguistic representation may enable one to exclude some of the false positives and include some of the missing false negatives.

3 Noun Compound Identification Strategies

In order to state the problem and our approach more precisely, we define three dimensions of noun compound identification strategies. One dimension is the type of linguistic information used to detect noun compounds, namely PoS tags (*PoS-based*) and syntax trees (*syntax-based*). A second dimension regards the treatment of proper nouns (NNPs), where we can define three options: (a) Simply treat proper nouns like common nouns (i.e. no special treatment), (b) exclude all noun sequences that contain proper nouns or (c) exclude noun sequences that are headed by a proper noun (assuming that the head is always the right-most word in the sequence). We refer to those three strategies as NNP^*, NNP^0 and NNP^h, respectively. A third dimension regards the number of constituents (i.e. nouns) within the noun compound. This is dependent on the type of linguistic information we use to identify noun compounds. In the PoS-based approach, we distinguish between *binary* and *n-ary* strategies for compound identification, where the former identifies noun+noun compounds and the latter identifies compounds that have $n >= 2$ constituents. In the syntax-based approach, we also distinguish between binary and n-ary compounds, but additionally taking into consideration that the bracketed structure of n-ary compounds is available. Hence, we can decompose n-ary noun compounds, where $n > 2$, into 'sub-compounds' including binary ones. We will explain the abovementioned dimensions using the following example sentence from the venerable Wall Street Journal (WSJ) section of the Penn Treebank (PTB; Marcus et al. [14]):

$$\ldots \text{Nasdaq}_{NNP}\ \text{bank}_{NN}\ \text{index}_{NN},\ \text{which}_{WDT}\ \text{tracks}_{VBZ}\ \text{thrift}_{NN}\ \text{issues}_{NNS}\ \ldots$$

First, under a PoS-based binary strategy we will extract *thrift issues*, while an n-ary strategy will extract both *thrift issues* and *Nasdaq bank index*. As for the proper noun treatment, an NNP^0 strategy would exclude *Nasdaq bank index* but NNP^h would not because the proper noun *Nasdaq* is not in the head position. In the syntax-based approach, the same rule for NNP treatment would apply, but there

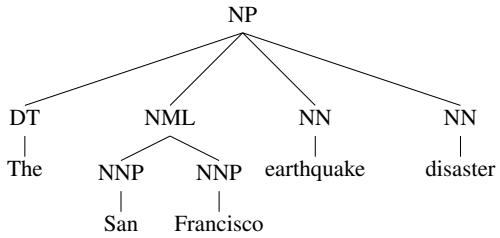

Figure 1: Internal noun phrase structure

will be more binary compounds, namely *bank index*, as syntax gives access to the internal structure of the compound *Nasdaq bank index*.

In our experiments we compare the PoS-based and syntax-based approaches for both binary and n-ary compounds, and NNP^0 and NNP^h for proper noun treatment.

4 Syntax-based Identification

The PoS-based strategy for noun compound identification requires a sequence of nouns that are not preceded and not followed by other nouns. With richer linguistic representations, such as syntactic trees, the definition of noun compounds goes one step further; the sequence of nouns is also a sequence of leaf nodes in the parse tree, hence the definition of a noun compound becomes a sequence of noun leaf nodes that are dominated by the same parent node—more specifically the same noun phrase parent node (we will amend this definition when we introduce the actual syntactic representation used in our experiments). The requirement of a single parent node stems from the fact that noun compounds act as one nominal, hence their constituents cannot belong to two different phrases.

In order to compare the PoS- and syntax-based strategies, we use the English part of the Prague Czech–English Dependency Treebank 2.0 (PCEDT; Hajič et al. [8]) which contains the WSJ section of the PTB. We chose to use the PCEDT because it includes the internal noun phrase annotations introduced by Vadas and Curran [22], whereas the 'original' PTB leaves the noun phrases flat.

Figure 1 shows an example of the internal annotation of noun phrases in the PCEDT. NML stands for *nominal modifier left-branching* and is one of the nodes introduced by Vadas and Curran [22]. The right-branching noun phrases were left unannotated. Our definition of noun compounds above requires leaf nodes to have an identical parent node, but in Figure 1 we see that *San Francisco* has a different parent node from the *earthquake disaster*, therefore in the implementation of syntax-based noun compound identification we make an exception for the identical-parent condition when the parent node is of type NML. In concrete terms, this means that we extract the following three compounds from the structure in Figure 1:

The ((San Francisco) (earthquake disaster)).

	PoS-Based				Syntax-Based			
	Binary		*N*-Ary		Binary		*N*-Ary	
	NNP^0	**NNP**^{h}	**NNP**^{0}	**NNP**^{h}	**NNP**^{0}	**NNP**^{h}	**NNP**^{0}	**NNP**^{h}
Tokens	27677	33167	30296	39429	29535	36441	34151	42835
Types	15128	18766	17167	23704	15853	20018	19469	25021

Table 1: Total number of noun compounds in PTB WSJ

Note that even though we make an exception for the identical-parent condition for NMLs, we still preserve their (left) bracketing constraints, hence, a compound like *Francisco earthquake* will not be extracted from the example phrase above.

5 Results and Discussion

In order to compare the PoS- and syntax-based approaches we experiment with detecting noun compounds in the full PTB WSJ in the PCEDT with eight different configurations as shown in Table 1, which provides total counts of compound instances (*tokens*) and the numbers of distinct strings (*types*).[1] In all configurations, the syntax-based strategy extracts more compounds than the PoS-based one, and that is because the former has access to the internal structure of the noun compounds and can therefore extract binary compounds out of n-ary ones where $n > 2$. Furthermore, in the binary setup, the PoS-based strategy is limited to strictly two consecutive nouns. The sequence $board_{NN}$ $meeting_{NN}$ $yesterday_{NN}$, for example, is not considered by the binary PoS-based strategy because it contains three consecutive nouns, whereas the syntax-based strategy extracts the sub-compound *board meeting*. Apart from this, the mere numbers do not tell us much in the absence of gold-standard data—to the best of our knowledge there is no gold-standard data set for noun compound identification. We therefore manually inspected a total of 100 random binary NNP^h compounds; 50 of which are only detected by the PoS-based strategy and 50 that are only detected by the syntax-based strategy.

Of the first set, 28 instances include a percent sign which is tagged as noun (NN) in the PTB, e.g. *% drop* in "...and a 4% drop in car loadings." In fact, *% stake* and *% increase* are among the top ten most frequent noun compounds identified by the PoS-based strategy, which is unsurprising given the WSJ domain. Such cases are easily excluded in the syntax-based strategy because the percent sign and the following noun belong to different constituents. We also identified five compounds that are due to annotation errors in the PTB on the PoS tag level, but not the syntax level. For example the tag NNS (plural noun) on the verb *amounts* in "one day's trading amounts to $7.6 billion". We also identified subtler annotation errors like annotating the adjective *in vitro* as preposition (IN) and noun (NN), which led the PoS-based strategy to extract *vitro cycles* as a compound in "...after only two in

[1] Note that no linguistic pre-processing (e.g. down-casing or stemming) was applied when calculating the type counts reported in Table 1.

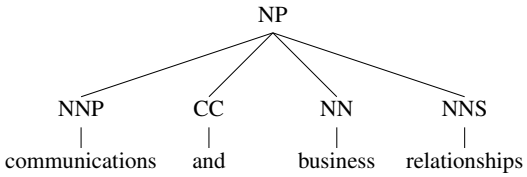

Figure 2: Coordination structure

vitro cycles". The remaining instances involve nouns that are not dominated by the same parent node. There are several linguistic constructions that may lead to such errors, such as the objects of a transitive verb and temporal modifiers like *today* and *yesterday* (tagged as nouns rather than adverbs in the PTB).[2] In sum the 50 compounds detected by only the PoS-based strategy are invalid noun compounds, which suggests that the syntax-based strategy succeeds in excluding some of the false positives referred to by Lauer [12].

Of the 50 noun compounds detected by the syntax-based strategy only, there are 38 compounds that were extracted from other compounds with more than two constituents—cases which could not have been identified by the binary PoS-based strategy. Furthermore, we see seven compounds that are either followed or preceded by other nouns. Such cases are also unidentifiable by the PoS-based strategy because it requires pairs of nouns not surrounded by other nouns. We also found four annotation errors where left-branching noun phrases were annotated as right-branching, for example in the phrase *San Diego home*, which leads to extraction of *Diego home* as a compound. The results analysis revealed that the syntax-based strategy includes arguably incorrect noun compounds when a noun is preceded by a coordinated phrase with noun conjuncts such as "communications and business relationships" in Figure 2. The syntax-based strategy extracts *business relationships*, but this can be either incorrect or incomplete extraction given the nature of coordination structures as we will discuss in the following section.

The results analysis also revealed that our implementation of the identical-parent condition was not fine-grained enough to preserve the left bracketing information in some NML constituencies. For example, in Figure 3 our implementation wrongly extracted the compound *development expenses*. In the following section we report the number of compounds extracted with a finer-grained implementation of the heuristic that handles such errors.

[2] According to the Part-of-Speech Tagging Guidelines of the PTB; "The temporal expressions yesterday, today and tomorrow should be tagged as nouns (NN) rather than as adverbs (RB). Note that you can (marginally) pluralize them and that they allow a possessive form, both of which true adverbs do not." See `http://groups.inf.ed.ac.uk/switchboard/POS-Treebank.pdf`

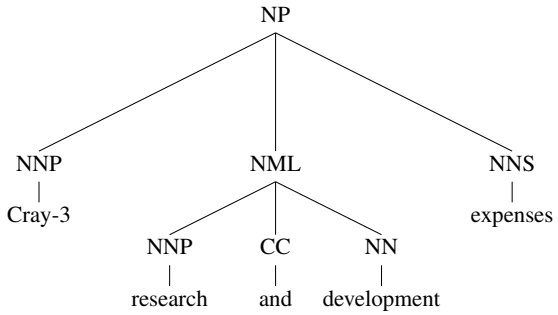

Figure 3: Coordination structure: Left-branching

6 Reflections

As shown in § 5, extracting noun compounds that are partially contained in nominal coordinate structures calls for careful treatment. In order to handle coordinate constructions properly, we need to distinguish between distributive and non-distributive (collective) coordinate structures. Consider the following coordinate constructions:

i Business and nursing programs

ii Research and development expenses

The first construction can be considered distributive and could be paraphrased as *business programs* and *nursing programs*. The second construction, however, is arguably non-distributive, which means that the two nominal conjuncts *research and development* 'jointly' modify the noun *expenses*—though it is also possible that the construction is referring to *research expenses* and *development expenses*, but we will assume that it is clearly non-distributive for the sake of argument. Given this distinction between distributive and non-distributive coordinate structures, it would in principle be possible to extract noun compounds from distributive coordinate structures, as we did with *business and nursing programs*. In practice, however, the PTB annotation does not distinguish between distributive and non-distributive coordinate structures, therefore we decided conservatively to exclude all noun compounds that are part of coordinate structures.

We further refined our implementation of the syntax-based identification heuristic to ensure that left-branching noun phrases are handled correctly. Consider the phrase "regional wastewater system improvement revenue bonds" in Figure 4, which includes an adjectival modifier as part of the initial compound; according to our definition of noun–noun compounds (as strictly nominal sequences), the only compound that can be extracted from this phrase is *revenue bonds*. Given underspecified bracketing information within the first NML constituent, extracting *wastewater system* might be incorrect because, arguably, *wastewater* in this construction may be modified by *regional*, as shown in the following bracketing:

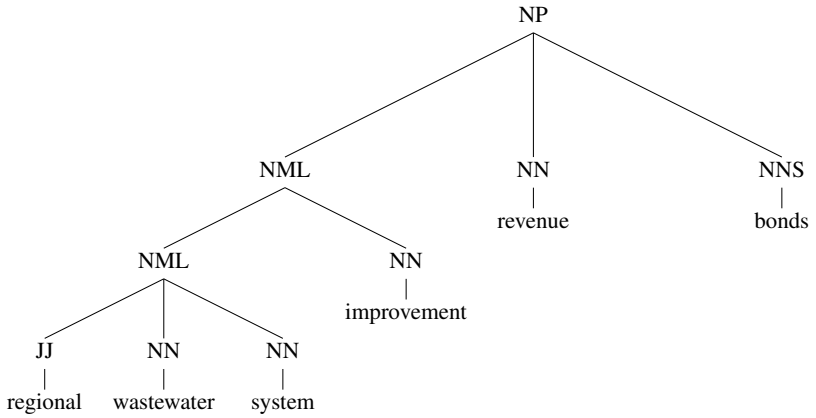

Figure 4: Complex left-branching noun phrase

((((regional wastewater) system) improvement) (revenue bonds))

Our refined implementation of the syntax-based heuristic, which also excludes all noun compounds that are part of a coordinate structure, identifies 33,095 binary NNP^h compounds and 38,925 n-ary NNP^h compounds, comparable in number to the PoS-based method (which would extract *some* compounds from both the conjoined modifier and adjectival modification structures of Figures 3 and 4). However, the trends regarding false positives and false negatives observed in the results analysis of § 5 apply with equal force to this more conservative parameterization of our syntax-based heuristics. We adopt this set of noun compounds as basis for our on-going work to automatically construct a data set of noun compounds with semantic relations based on the so-called PCEDT functors and noun senses and arguments in NomBank (Meyers et al. [15]).

7 Conclusion and Outlook

In this paper we presented two approaches to noun–noun compound identification, syntax-based and PoS-based. We identified three dimensions on which approaches to noun compound identification may vary. Our results and analysis suggest that achieving high-quality noun compound identification requires linguistic representations at least at the level of syntactic structure. We also show, however, that complex cases that include coordinate structures may require even richer linguistic annotations.

One of the challenges for quantifying the accuracy of the different identification strategies is the lack of gold-standard evaluation data. We therefore opted for manual inspection of the extracted compounds, which in turn led to gradual improvement in our implementation of the syntax-based identification heuristic.

In future work, we seek to extend our investigation into the utility of syntactic

structure for the task of compound identification in two ways; by (a) evaluating the recent re-annotation of the WSJ Corpus in DeepBank (Flickinger et al. [5]) as a candidate gold standard, and by (b) gauging the effects on compound identification accuracy when moving from gold-standard syntactic structures to those available from state-of-the-art syntactic parsers. Also, we have started to combine our high-quality compound identification over PTB trees with thematic annotations over the same underlying text from resources like PCEDT and NomBank, aiming to fully automatically create comprehensive and high-quality gold-standard data for the thematic interpretation of relations among compound members.

References

[1] Timothy Baldwin and Takaaki Tanaka. Translation by machine of complex nominals. Getting it right. In *Second ACL Workshop on Multiword Expressions: Integrating Processing*, page 24 – 31, Barcelona, Spain, 2004.

[2] Lou Burnard. Reference guide for the British National Corpus version 1.0, 2000.

[3] Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. SemEval-2010 Task 9. The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, page 100 – 105, 2009.

[4] Pamela Downing. On the creation and use of English compound nouns. *Language*, 53(4):810 – 842, 1977.

[5] Dan Flickinger, Yi Zhang, and Valia Kordoni. DeepBank. A dynamically annotated treebank of the Wall Street Journal. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, page 85 – 96, Lisbon, Portugal, 2012. Edições Colibri.

[6] Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. On the semantics of noun compounds. *Computer Speech & Language*, 19(4):479 – 496, 2005.

[7] Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. SemEval-2007 Task 04. Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, page 13 – 18, Prague, Czech Republic, 2007.

[8] Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and

Zdeněk Žabokrtský. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, page 3153–3160, Istanbul, Turkey, May 2012.

[9] Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. SemEval-2013 Task 4. Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, page 138–143, Atlanta, Georgia, USA, 2013.

[10] Mirella Lapata and Alex Lascarides. Detecting novel compounds. The role of distributional evidence. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, page 235–242, 2003.

[11] M. Lauer and M. Dras. A probabilistic model of compound nouns. In *Proceedings of the 7th Australian Joint Conference on AI*, 1994.

[12] Mark Lauer. *Designing Statistical Language Learners. Experiments on Noun Compounds*. Doctoral dissertation, Macquarie University, Sydney, Australia, 1995.

[13] Charles Na Li. *Semantics and the Structure of Compounds in Chinese*. PhD thesis, University of California, Berkeley, 1972.

[14] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpora of English. The Penn Treebank. *Computational Linguistics*, 19:313–330, 1993.

[15] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. Annotating noun argument structure for NomBank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, page 803–806, Lisbon, Portugal, 2004.

[16] Preslav Nakov. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19(3):291–330, 2013.

[17] Diarmuid Ó Séaghdha. Learning compound noun semantics. Technical Report UCAM-CL-TR-735, University of Cambridge, Computer Laboratory, Cambridge, UK, 2008.

[18] Diarmuid Ó Séaghdha and Ann Copestake. Interpreting compound nouns with kernel methods. *Journal of Natural Language Engineering*, 19(3):331–356, 2013.

[19] Emily Pitler, Shane Bergsma, Dekang Lin, and Kenneth Church. Using web-scale n-grams to improve base NP parsing performance. In *Proceedings of the 23rd International Conference on Computational Linguistics*, page 886–894, Beijing, China, 2010.

[20] Tony Rose, Mark Stevenson, and Miles Whitehead. The Reuters Corpus Volume 1. From yesterday's news to tomorrow's language resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, volume 2, page 827–832, Las Palmas, Spain, 2002.

[21] Stephen Tratz and Eduard Hovy. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics*, page 678–687, Uppsala, Sweden, 2010.

[22] David Vadas and James Curran. Adding Noun Phrase Structure to the Penn Treebank. In *Proceedings of the 45th Meeting of the Association for Computational Linguistics*, page 240–247, Prague, Czech Republic, 2007.

[23] David Vadas and James R Curran. Parsing noun phrases in the Penn Treebank. *Computational Linguistics*, 37(4):753–809, 2011.