8

Improving the lexical coverage of English compound adjectives

Improving the lexical coverage of English compound adjectives in syntactic parsing

Nelleke Oostdijk
Department of Linguistics, Radboud University Nijmegen

Abstract

The present paper addresses the question how in syntactic parsing the coverage of words in previously unseen text may be improved. The adjectives in English are presented here as a case study. Working on the assumption that most new words that are introduced into the language are constructed on the basis of already existing words through the application of word-formation processes, we investigate the role that different word-formation processes play, more specifically in the formation of adjectives in English. An analysis of adjectives in the BNC shows that in the case of adjectives compounding is the word-formation process that is most productive. Moreover, compound adjectives are not formed by combining bases at will; rather, a limited set of fairly simple rules apply that restrict the co-occurrence of bases. This makes it feasible to develop an approach for handling compound adjectives which is rather effective, as is evident from the results from a first implementation where of a set of 30,561 compound adjectives derived from the BNC, 88.68% were correctly identified

Proceedings of the 18th Meeting of Computational Linguistics in the Netherlands, pp. 117–130
Edited by: Suzan Verberne, Hans van Halteren, Peter-Arno Coppen.
 Contact: n.oostdijk@let.ru.nl

as such. Incorporation of the rules in the grammar underlying the Pelican parser accounts for a 7.65% increase in the parser's coverage of a subset of 10,123 sentences taken from the Leipzig corpus.

8.1 Introduction

In many computational linguistic applications involving tagging and parsing the lexicon is critically important to the success of the application. In order to obtain maximal coverage of a text, there is a need for full-coverage of the words that appear in it. For the lexicon this requirement poses a problem as it is impossible to have in advance a complete inventory of all the words that may be encountered in previously unseen text, even for a morphologically 'poor' language such as English. No matter how large a lexicon is, it will at best be near-complete. Therefore, many approaches have resorted to having the lexicon work in tandem with some heuristics that provide a fall-back option for cases where a specific lexical item is unknown in the sense that it has not been included in the lexicon. The heuristics capitalize on the generalization of observed commonalities in items, often morphological features (prefixes and suffixes), but also the surrounding context and spelling cues like capitalization.[1] From experiences in parsing, however, we know that there are still numerous words that are missed out on, while it is not just proper names that are missing: unknown words may occur in each of the open word classes In the present paper we investigate the nature of these words, restricting ourselves to adjectives in English, and how they are best dealt with.
The organization of the paper is as follows: In Section 8.2 we introduce the different word-formation processes and describe how these manifest themselves in data from the British National Corpus (BNC). Since compounding appears to be the most productive word formation process, in Section 8.3 we then focus on compound adjectives. An initial manual classification of a subset of the data suggests that these compounds can be described by means of syntactic rules. We describe briefly how this can actually be done and what results were obtained in the actual implementation. Section 8.4 concludes this paper.

8.2 Word-formation processes

The inventory of lexical items that make up the lexicon of a language is continuously being expanded through the introduction of new words. While occasionally words are introduced that are completely new in the sense that they have not been constructed on the basis of known words, more commonly words are introduced that are constructed on the basis of already existing words through the application of word-formation processes. In the formation of adjectives in English specifically the following processes are involved (cf. Quirk et al. 1985, p. 1520):[2]

[1]For research on unknown word guessing carried out in the area of part-of-speech tagging, see for example Nakagawa et al. (2001), Orphanos and Christodoulakis (1999), Thede (1998), Tseng et al. (n.d.) and Weischedel et al. (1993).
[2]In what follows the term 'base' is used to refer to the minimal free form of a lexical item. We shall consider as base adjectives all lexical items that are adjectives in their minimal free form. Thus *old*

a. prefixation: putting a prefix in front of the base sometimes with, but most usually without, a change of word class; e.g. *un-dead*, *non-empty*, *over-eager*

b. suffixation: putting a suffix after the base, sometimes without, but usually with, a change of word class; e.g. *adjustable*, *financial*, *successful*, *historical*

c. conversion: assigning the base to a different word class with no change of form; in the case of adjectives this process typically concerns the adjectival use of present and past participles; e.g. *crusading*, *pounding*, *slurred*, *validated*

d. compounding: adding one base to another; e.g. *old-age*, *cost-conscious*, *historically-eclectic*, *civil-political*.

In addition to these four processes adjectives can be formed with the help of combining forms (e.g. *hispano-*, *bio-*, *climato-*). As Quirk et al. observe such forms "have the semantic characteristics of the first constituent in a compound but they resemble prefixes in mostly (...) being obligatorily initial, in having little or no currency as separate words, and in not normally being the stressed part of a complex word" (1985, p. 1520).

Conversion, it appears, has been perceived as unproblematic and consequently has received very little attention from lexicographers and computational linguists as it is assumed that any participle can occur as adjective. This sharply contrasts with derivation (prefixation and suffixation) and compounding. These processes have been given ample attention with different degrees of success, where it is apparent that they cannot be handled in the same manner. Thus, as regards derivation, in dictionaries derivational affixes usually occur as separate entries, while in morphological analyzers or word form lexicons used for NLP, knowledge about derivational morphology is applied for the analysis or generation of word forms. Compounding does not lend itself for this kind of approach. In actual practice therefore, we find that dictionaries include only compounds with (presumably) high currency, while in NLP heuristics are used to capture what are assumed to be the more common instances relying on the presence of a word-final adjectival suffix and a hyphen signaling apparent compounding.[3] Meanwhile, compound adjectives have repeatedly been reported to be particularly high frequent in some text genres, especially in news reportage, advertising, and also in poetry (Meijs 1975, Salzman n.d., Jackson 2006), where they serve to condense information. In order to gain insight into how different word-formation processes con-

and *young* are considered base adjectives, while *economic* and *useful* are considered to be instances of adjectival suffixation.

[3]Compound adjectives receive quite some attention from usage guidebooks such as the *The American Heritage Book of English Usage*. Here, however, the focus of attention is mostly on whether or not a compound adjective should be hyphenated, and not so much on the composition of compound adjectives.

tribute to the formation of adjectives we decided to investigate the adjectives that occur in the British National Corpus (BNC).

8.2.1 Adjectives in the BNC

We extracted our initial data set from the BNC word frequency list as compiled by Killgariff. The set comprises all 107,657 types with which the POS tag 'aj0' (denoting the word class of adjective) has been associated.[4] Together these types account for 6,264,673 tokens in the corpus. The adjectives show a typical Zipfian distribution. The adjective type *other* which can be found at rank 1 by itself accounts for 2.07% of the tokens, while the 59,591 hapax types (making up 55.35% of the total number of types) together only account for 0.95% of the total number of tokens.

When we consider our data set, we find that the most frequent types of adjectives appear to fall into two groups: one comprising base adjectives (*other*, *new*, *good*, etc.) and the other comprising adjectives that have been arrived at through derivation, more in particular suffixation (*different*, *important*, *national*, etc.). It is only at rank 69 that the first instance of conversion (*following*) can be found, while compounding does not appear until rank 290 where we find *long-term* as the highest ranked compound adjective. These observations, together with the experience we had with unknown words in parsing previously unseen text which most of the time were found to be compounds, led us to hypothesize that there might be a correlation between the word-formation processes on the one hand and the type frequency on the other hand, such that adjective types with higher frequencies can be explained more often in terms of derivation with very little compounding, while with adjective types with lower frequencies more often compounding will be involved at the expense of derivation. Conversion is here hypothesized to be evenly distributed throughout and not to show any frequency effect.

8.2.2 Word-formation processes related to the type frequency distribution

In order to test the word-formation*type-frequency hypothesis we undertook an analysis of adjective types with frequency 1, 5, 10, 15, 20, and 25 respectively. We manually classified each of the adjective types according to the word-formation process involved. The results of the classification are presented in Table 8.1.[5] Compounding indeed appears to be the process that is most productive in the sense that it is responsible for the largest number of hapaxes and thus for the greatest increase in the number of previously unseen types.[6] However, unlike what we hypothesized it is not derivation that gives way to compounding, but conversion:

[4]Thus we excluded all other types where the POS tagger had assigned the tag aj0-av0, aj0-nn1, aj0-vvd, aj0-vvg, or aj0-vvn, indicating a high degree of uncertainty as to the appropriateness of associating the word with the class of adjectives (av0 = adverb, nn1 = common noun singular, vvd = past tense verb, vvg = present participle, vvn = past participle).

[5]The number of instances reported here are the number of instances that remain after we have discarded from our data all apparent 'rubbish' (tokenization errors, spelling errors, foreign language data, etc.)

[6]On productivity, see also Plag (2003).

Table 8.1: Proportion of types with frequency (F) 1, 5, 10, 15, 20 and 25 resp. explained through word-formation processes

Word formation process	Type F1 *N*=54,591	Type F5 *N*=2,324	Type F10 *N*=818	Type F15 *N*=429	Type F20 *N*=234	Type F25 *N*=193
Conversion	4.26	17.77	22.74	25.87	21.79	31.61
Derivation	28.61	30.25	34.72	33.33	36.32	33.68
Compounding	64.09	47.59	37.53	35.43	35.90	33.16
Combining	2.92	3.87	4.40	3.03	4.27	0.52
Base ADJ	0.12	0.52	0.61	2.33	1.71	1.04
total	100.00	100.00	100.00	100.00	100.00	100.00

while derivation is distributed rather evenly across different frequencies, conversion occurs increasingly less frequently among low-frequent types. Now one could speculate that the present limitation to only the set of items that have been tagged unambiguously as aj0 might have a serious impact on the relative frequency of conversion, as often the tags aj0-vvn and aj0-vvg are assigned to tokens which exhibit conversion. However, this does not appear to be the case: even if we include all instances of tokens tagged aj0-vvn or aj0-vvg and consider these as conversion, the picture essentially remains the same (cf. Table 8.2).

Table 8.2: Proportion of types with frequency (F) 1, 5, 10, 15, 20 and 25 resp. that can possibly be explained through conversion (tags aj0, aj0-vvn, aj0-vvg)

Word formation process	Type F1 *N*=60,198	Type F5 *N*=3,142	Type F10 *N*=1,171	Type F15 *N*=620	Type F20 *N*=349	Type F25 *N*=302
Conversion	9.31	26.03	30.15	30.81	32.95	36.09

In the next section we investigate compound adjective types more closely, as they play a key role in the appearance of previously unseen words.

8.3 Compound adjectives

In the definition provided by Quirk et al. (1985) compounds are formed by combining one base with another. This seems to suggest that bases can be combined freely, without being bound by any restrictions. In order to check whether this is indeed the case we investigated all compound hapaxes in our data. To this end we extracted from our initial data set the subset of compound adjectives with frequency 1. The set comprises 34,987 items, only 3,829 of which are complex (i.e. multi-word) compounds combining more than two words,[7] 31,158 are simple

[7]Examples of complex compound adjectives are *easy-to-grasp*, *fun-to-wear*, *red-and-white-striped*, *suddenly-made-redundant*, and *very-low-fat*.

compounds combining two words, nearly all of which are hyphenated.[8] In what follows we focus exclusively on the set of 30,561 hyphenated simple compounds.

8.3.1 Simple compounds

The simple compound adjectives in our data roughly fall into five main groups.[9] A brief characterization of each of these groups is given below.

Group 1 comprises compound adjectives that take an adjective base as head and some other word class as first part. The adjective base is either a base adjective or an adjective arrived at by way of derivation. The head typically combines with a noun, numeral or an adverb. Typical examples are *application-dependent*, *cabinet-wide*, *four-dimensional*, *climate-relevant*, *overly-sensitive*, and *pharmalogically-active*. Quite frequently time adverbs occur as in *once-blind*, *ever-reluctant*, *still-resident*, *then-arthritic*.
Group 2 is formed by compound adjectives that are formed by combining two adjective bases. Examples are *cognitive-affective*, *classical-scholarly*, *chemical-physical*, *electric-acoustic*, and *Egyptian-Syrian*.
Group 3 comprises compound adjectives that are headed by adjective bases that have been arrived at through conversion. One subgroup consists of items where the head combines with a noun, adjective or adverb. Examples are *panic-driven*, *bug-infested*, *fresh-caught*, *money-generating*, *posh-looking*, *duly-authorised*, *forever-changing*. Another subgroup is formed by compounds headed by a present or past participle where the head combines with a particle. Examples: *agreed-upon*, *signed-off*, *trimmed-down*, *turning-away*, and *coming-down*.
Group 4 is made up of derivational compounds. The head of the compound is always a noun which is combined with an adjective, a noun or a numeral. To the combination the adjectival suffix *-ed* is added, giving the resulting word its adjectival status. Examples are *sunken-cheeked*, *missing-toothed*, *bare-fisted*, *single-platformed*, *metal-cased*, *leopard-sized*, *4-cornered*, and *six-fingered*.
Group 5 comprises compounds that are considered to be adjectives but are more peripheral to the class of adjectives than items falling within any of the other groups above. They are headed by a noun which is preceded by an adjective or a numeral. Examples are *close-attack*, *big-league*, *four-sensor*, *sixteen-page*.[10]

As is apparent from this description, there are clear indications that in combining bases to form compound adjectives there are underlying syntactic rules and semantic restrictions to be observed. This suggests that it should be feasible to develop a set of rules that describe how single token lexical items that are listed the lexicon may be combined so as to form compound adjectives.

[8] Non-hyphenated compound adjectives are always written as single items: e.g. *timesharing*, *soft-hearted*, *windswept*.
[9] There are some minor types that we shall not discuss here in detail.
[10] In our data we also find a large number of instances where a noun base is combined with another noun base (e.g. *author-subscriber*, *rugby-soccer*). Whether to consider these modifying compound nouns as adjectives in the literature is subject of discussion (cf. Bauer (1983): 210; Meijs (1975): 194).

In the next sections we follow up on this idea as we describe how it was implemented in the context of the Pelican parser and lexicon.[11]

8.3.2 Compounding rules

The grammar underlying the Pelican parser is an attribute grammar that in terms of rewrite rules describes the syntactic structures that occur in English. The grammar operates in tandem with a lexicon in which in principle should account for all possible word forms in the language. The interface between the grammar and the lexicon constitutes of a set of defining rules in the grammar which specify the different word classes that are included in the lexicon. In the case of adjectives, several subclasses are distinguished: apart from the subclass of common adjectives, adjectives are distinguished that originate from a present participle (*-ingp*) or past participle (*-edp*) form of a verb, while also the adjectives *such*, *very* and *worth* are distinguished as separate subclasses on the grounds of their idiosyncratic syntactic behaviour.

Where the grammar stops at the point where the lexicon is called upon, a typical rule looks as follows:[12]

```
c ADJ ADJECTIVE (AJP_TYPE, GRADABILITY):
   n listed token ADJ (AJP_TYPE, GRADABILITY).
```

The rule states that with a lexical category or word class adjective information is associated which informs us about the type of adjective ('AJP_TYPE'), whether or not the adjective is gradable and if it is, whether the form of adjective is 'absolute', 'comparative' or 'superlative' ('GRADABILITY'). The definition of the adjective as a 'listed token' refers to the specification of the individual items that belong to this particular word class as it is given in the lexicon.

Lexical entries in the lexicon take the following form:

```
"gorgeous"  n listed token ADJ (ajp_type_attributive|
              ajp_type_predicative, gradability_absolute)

"singing"   n listed token ADJ ingp(ajp_type_attributive,
              gradability_absolute|gradability_non_gradable)
```

In order to account for compound adjectives, we extended our grammar with a set of rules that was based on the observations we had made in the analysis of the

[11] The Pelican parser is an English wide-coverage rule-based parser that is being developed at Nijmegen University. See also http://lands.let.ru.nl/projects/pelican

[12] The initial 'c' in the example rule indicates that it concerns a (lexical) category. The abbreviation and long name are used for the sake of convenience. While the long name enhances the readability of the grammar, the abbreviation is used when the eventual output is represented in the form of a tree. All nodes that are associated with rules that have been prefixed with the letter 'n' will not appear in the eventual output.

BNC data (see Section 8.3.1). We included rules of the form[13]

```
c ADJ ADJECTIVE (AJP_TYPE, GRADABILITY):
(1) n listed token N (N_TYPE, N_CLASS, NUMBER),
    "-\--",
    n listed token ADJ (AJP_TYPE, GRADABILITY);
(2) n listed token ADJ (AJP_TYPE, GRADABILITY),
    "-\--",
    n listed token ADJ (AJP_TYPE, GRADABILITY1).
(3) n listed token NUM (NUM_TYPE, NUMBER),
    "-\--",
    n listed token N (n_type_common, n_class_other,
                                          number_sing),
    "-ed";
(4) n listed token N (N_TYPE, N_CLASS, NUMBER),
    "-\--",
    n listed token LV (complementation_motr,
         finiteness_pastpart, MOOD, NUMBER1, PERSON);
(5) n listed token ADJ (AJP_TYPE1, GRADABILITY),
    "-\--",
    n listed token N (N_TYPE, N_CLASS, NUMBER).
```

The features associated with the various word classes were used to restrict the possible combinations, as for example in the case of compound adjectives formed on the basis of a past participle form of a verb preceded by an adjective, where we required that the lexical verb should be mono transitive ('complementation_motr', rule 4).

8.3.3 Applying the rules

In order to measure to what extent incorporation of the rules in the grammar contributes to an improvement of the parser's coverage, we carried out two evaluations: one in which we applied the adapted version of our parser to the set of 30,561 simple compound adjectives that we had derived from the BNC, the other in which we applied the parser to a subset of 10,123 sentences that were taken from the Leipzig Corpus. In both cases the lexicon remained unchanged. The results are presented below.

Results obtained on compound adjectives from the BNC

In the case of the compound adjectives from the BNC, the fact that all items in the data set were lower case presented a problem since in the approach taken by the Pelican parser and lexicon the distinction between upper and lower case is taken to be significant.[14] Prior to parsing, therefore, we manually restored upper case characters where necessary. The overall coverage we obtained was 88.68%, i.e.

[13]The abbreviations are as follows: ADJ adjective, LV lexical verb, N noun, and NUM numeral. Examples of the compound adjectives covered by each of the alternatives are *user-adjustable*, *cultural-political*, *four-stringed*, *mafia-controlled*, *direct-action*.

[14]The distinction between upper case and lower case is particularly relevant for distinguishing between proper nouns/names and common nouns (John vs john).

27,100 compound adjectives were accounted for by the rules we had developed (cf. Table 8.3). 17,707 of these were analyzed unambiguously, that is, in each case a single rule applied. In the 30.74% of the cases (9,399 items) more than one rule was applied, thus yielding ambiguity at sub-word level. For 3,461 items the set of rules fails and apparently needs to be extended.

Table 8.3: Coverage of compound adjectives

		# items	% of total
Success	single analysis	17,707	57.94
	multiple analyses	9,393	30.74
	subtotal	27,100	88.68
Failure	no analyses	3,461	11.32
	total	30,561	100.00

Discussion

In what follows we shall first discuss the set of items which were unambiguously identified as compound adjectives. Next we turn to the ambiguous cases. We conclude our discussion with an analysis of the failures.

Unambiguous cases

Among the cases that were unambiguously identified as compound adjectives, the distribution over the various groups we distinguished in Section 8.3.1 appears to be rather unbalanced. As Table 8.4 shows, group 3 compounds are the largest group by far, making up 46.46% of the total number of compound adjectives in our data set. When we look at this group in more detail, we find that past participle forms occur much more frequently than present participle forms, while compounds are much more frequently formed by combining a noun and a participle than by combining an adverb and a participle (for details see Table 8.5). Where with group 3 compounds the adverb follows the participle, it is always a particle (e.g. *about*, *away*, *back*, *by*, *down*, *in*, *on*, *off*, *through*): *dozing-off*, *gearing-up*, *hidden-away*, *nailed-down*, *backed-off*, etc. Adverbs that precede the participle are always general adverbs: *freely-existing*, *rapidly-expanding*, *publicly-approved*, *freshly-scrubbed*, *cylindrically-shaped*, etc.

Ambiguous cases

In 9,393 cases (30.74% of the total number of 30,561 items) multiple rules could be applied. As a result, ambiguity was generated at the sub-word level. In itself sub-word level ambiguity is not considered to be problematic as it does not appear in the eventual output.[15] However, ambiguity may point to unforeseen

[15]Recall that the objective was to improve on the lexical coverage of previously unseen words, in this

Table 8.4: Distribution of compound adjectives over different groups (cf. Section 8.3.1)

group	description	# items	% of total
1	headed by ADJ	1,985	11.21
2	ADJ-ADJ	298	1.68
3	headed by -ingp or edp	8,226	46.46
4	'derivational compounds'	842	4.76
5a	headed by N: ADJ-NUM-N	3,273	18.48
5b	headed by N: N-N	2,926	16.54
rest	minor types	157	0.89
total		***17,707***	***100.00***

Table 8.5: Sub-groups for group 3

combination	# items	% of total
N-LVingp	869	10.56
N-LVedp	5,118	62.22
ADV-LVingp	181	2.20
ADV-LVedp	1,737	21.52
LVingp-ADV	21	0.26
LVedp-ADV	300	3.65
total	***8,226***	***100.00***

interaction,between rules, in which case we might want to adapt our rules.
What we found was that, where there was ambiguity, in most cases there were two competing rules (cf. Table 8.6). One major source of ambiguity is of course the multiple word class membership of various tokens. For example, a great many words can be either an adjective or a noun (e.g. *private*, *light*, *public*, *current*), a noun or a present participle (e.g. *sinking*, *smelling*, *counting*, *dressing*), or an adverb or an adjective (e.g. *half*, *well*) Ambiguity that arises from the rules themselves is found in cases like *dramatic-coloured* where forms ending in *-ed* (like *coloured*) can be retrieved in full from the lexicon or—alternatively—are described as formed on the basis of a noun (*colour*) to which the adjectival suffix *-ed* has been added.

Pairs of rules that were most frequently in competition with each other are the following:
These results did not give rise to wanting to adapt the present set of rules.

Failures

case adjective compounds.

Table 8.6: Ambiguous cases

# parses	# items	% of total (N=30,561)
2	7,237	23.68
3	1,315	4.30
4	758	2.48
5	74	0.24
6	9	0.03
sub total	***9,393***	***30.74***

Table 8.7: Competing rules

rule 1 vs rule 2	example	# items	% of total
N-LVingp vs N-N	*news-reporting*	1,281	17.70
ADJ-N vs N-N	*cold-chain*	1,215	16.79
N-LVedp vs ADJ-N-ed	*explosive-tipped*	1,028	14.20
N-ADJ vs N-N	*policy-variant*	981	13.56
N-ADJ vs ADV-ADJ	*half-hysterical*	391	5.40
ADV-LVedp vs N-LVedp	*well-invested*	228	3.15
other		2113	29.20
	total	***7,237***	***100,00***

In those cases where the set of rules failed to identify the item as a compound adjective, failure could be attributed to either of two causes:

1. (at least one) part of the compound was not in the lexicon. This was the case for 1,460 items (4.78% of the 30,561 items).

2. while in principle both constituent parts of the compounds were accounted for in the lexicon, there was no rule describing the particular combinations. This held for 2,001 items (6.55% of the 30,561 items).

As regards tokens that were (as yet) missing from the lexicon, many of these tokens that were acronyms or chemical substances. Another class of tokens that was frequently missing is constituted by adjectives deriving from proper names, such as *Trotskyite* as encountered in *half-Trotskyite*, or *Wagnerian* as in *near-Wagnerian*). Other missing tokens did not point to any systematic omissions.

Inspection of the group of 2,001 items where the rules failed to identify the compound adjective as such led to the observation that some of the failures were due to a too rigid formulation of particular rules. Thus 109 items failed due to fact

that the rule describing compounds that consisted of a numeral followed by a common noun excluded the class of common nouns that denote some kind of measure. Examples are *12-ft*, *36-km*, *40-lb*, and *100-mph*. In a similar fashion, the restriction to have a noun followed by the adjectival suffix *-ed* preceded by a numeral failed to take into account instances where instead of a numeral the quantifier *many* occurs: *many-stranded*, *many-tiered*, *many-tentacled*. Other failed items suggested that additional rules might be formulated. These include the following:[16]

- a compound adjective may be formed on the basis of an adjective and a present participle verb. Especially the verbs sound and smell appear particularly productive here: together they account for 82 failures, including for example *metallic-sounding*, *soppy-sounding*, *unusual-sounding*, *Irish-sounding*, *authentic-sounding*, *corrupt-smelling*, *fishy-smelling*, *milky-smelling*.
- a compound adjective may be formed by combining a preposition and a singular common noun; for example *before-tax*, *between-species*, *up-river* (69 items of the 2,001)

While an analysis of the set of items that the set of rules failed to identify as compound adjectives suggests that some rules should be formulated less restrictive while also additional rules may be formulated, it is important to keep in mind that this should be done with care since there is the risk that this may have an undesired side-effect in generating (additional) ambiguity.

Results obtained on sentences from the Leipzig Corpus

We are currently in the process of analyzing the Leipzig Corpus, a collection of one million sentences originating from various newspapers (Associated Press, Wall Street Journal, Financial Times, OTS News Ticker). In order to see to what extent the compounding rules for adjectives contribute to the success of the parser, we investigated a subset of 10,123 sentences that have been parsed successfully. In 774 cases (7.65%) the correct parse is arrived at by means of one of the compounding rules. From this we may conclude that the compounding rules are quite effective: without these rules, the parser would have failed to produce the correct parse.[17]

8.4 Conclusion

Our analysis of adjectives as they occur in the BNC shows that in the case of adjectives, compounding is the word-formation process that is most productive. Moreover, we find that compounds are not formed by combining bases at will;

[16] It proved difficult to generalize over larger sets of failed items. Below, the two examples of rules that one might consider adding cover a fair number of instances. Any other rule would at best account for a handful of items.

[17] Note that the impact that the compounding rules have here may be biased by the particular genre (news reportage). Future research should give insight to what extent this is indeed the case.

rather, a limited set of fairly simple rules apply that restrict the co-occurrence of bases. The introduction of handcrafted rules that call upon information that is already present in the lexicon provides a means to maintain control and guard over the quality of the lexical information while a substantial improvement is obtained in the lexical coverage of compound adjectives.

References

Bauer, L. (1983), *English Word-formation*, Cambridge University Press, Cambridge.

Jackson, M. (2006), Compound Adjectives in Arden of Faversham, *Notes and Queries* **53** (1), pp. 51–55.

Meijs, W. (1975), *Compound Adjectives and the Ideal Speaker-Listener. A Study of Compounding in a Transformational-Generative Framework*, North-Holland, Amsterdam.

Nakagawa, T., T. Kudoh, and Y. Matsumoto (2001), Unknown Word Guessing and Part-ofSpeech Tagging Using Support Vector Machines, *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001)*, Tokyo, pp. 325–331.

Orphanos, G. and D. Christodoulakis (1999), POS Disambiguation and Unknown Word Guessing with Decision Trees, *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL–99)*, pp. 134–141.

Plag, I. (2003), *Word-formation in English*, Cambridge University Press, Cambridge.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985), *A Comprehensive Grammar of the English Language*, Longman, London.

Salzman, A. (n.d.), Using Compound Adjectives to Give Physical or Metaphorical Descriptions. Available at `http://www.iei.uiuc.edu/structure/Structure1/haired.html`.

Thede, S. (1998), Predicting Part-of-Speech Information about Unknown Words using Statistical Methods, *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (ACL/COLING–98)*, pp. 1505–1507.

Tseng, H., D. Jurafsky, and C. Manning (n.d.), Morphological features help POS tagging of unknown words across language varieties. Available at `http://www.stanford.edu/~jurafsky/sighan_pos.pdf`.

Weischedel, R., M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci (1993), Coping with Ambiguity and Unknown Words through Morphological Models, *Computational Linguistics* **19** (2), pp. 359–382.

Resources consulted

BNC database and word frequency lists (n.d.). Compiled by Adam Kilgariff. Available at `http://www.kilgarriff.co.uk/bnc-readme.html`.

Oxford English Dictionary (2003), Oxford University Press, Oxford.

The American Heritage Book of English Usage. A Practical Guide to Contemporary English (1996). Available at `http://www.bartleby.com/64/84.htm`.