Parser-based analysis of syntax-lexis interactions

Lehmann, Hans Martin ; Schneider, Gerold

Parser-based analysis of syntax-lexis interactions

Hans Martin Lehmann and Gerold Schneider

University of Zurich

1. Introduction

Fixedness in language has been extensively studied in areas like multi-word units, idiomatic expressions, collocations and verb-particle constructions. These have often been treated as relatively fixed non-compositional sequences, which allow for little variation. In our paper we will focus on co-occurrence phenomena between elements in syntactic relations. Specifically, we focus on subject-verb and verb-object relations in active and passive constructions. Looking for fixedness in these syntactic relations where compositionality is expected to hold to a large degree may strike the reader as a strange undertaking. Our main interest lies in establishing how far an open choice principle holds for these relations and to what degree we can find fixedness in these syntactic relations.
The identification of syntactic relations requires syntactically annotated corpora. Most standard corpora of sufficient size are either not annotated at all, or annotated at the non-hierarchical level of part-of-speech tags only. They typically contain no hierarchical information about the syntactic organisation of sentences.
Parsing approaches to fixedness are still quite rare. Exceptions are Lin (1998) and Seretan and Wehrli (2006). Robust broad-coverage syntactic parsers, for example Schneider (2007) or Andersen (2008), have now become available, offering new perspectives for this research.
This paper describes the syntactic annotation of over 160 million running words with the help of *Pro3Gres*, a dependency parser. See Schneider (2007) for a more detailed description. We document the extraction of a database with verb centres and their dependents. We then explore the possibilities and limitations of this dependency database for the study of fixedness in syntactic relations.

2. Previous Work

Most approaches to fixedness in language are based on the use of observation windows or regular expression patterns over large corpora with flat part-of-speech annotation. Typically, collocations and multi-word expressions are investigated. Syntactic analysis has been recognised as a prerequisite for accurately describing the syntax-lexis interface:

> Ideally, in order to identify lexical relations in a corpus one would need to first parse it to verify that the words are used in a single phrase structure. However, in practice, free-style texts contain a great deal of

> nonstandard features over which automatic parsers would fail. This fact is being seriously challenged by current research (...), and might not be true in the near future" (Smadja, 1993, 151)

The currently available corpora which are manually analysed for syntactic structure, for example ICE-GB and the Penn Treebank, are too small for infrequent word-word interactions, and automatic parsers have, until recently, not been robust enough to analyse large corpora. This partly explains why most approaches have been based on observation windows or part-of-speech sequences over corpora without hierarchical syntactic annotation.
These approaches have made a wealth of corpus research possible, but some types of research can profit much from hierarchical syntactic information. Let us consider a simple example. Research on verb subcategorisation and selectional preferences needs to retrieve verb-object relations. The verb-object relation is one of the least problematic relations, since the distance between the verb and the object is quite short in English. In verb-subject relations, for example, relative clauses or appositions may intervene, which further complicates retrieval. In the following, we discuss which types of errors part-of-speech sequences and windows-based methods are typically prone to and motivate our use of a parsing approach.

2.1 Part-of-speech tag sequences

Sequences of part-of-speech tags or regular expressions over part-of-speech tags (e.g. Hoffmann and Lehmann 1998, Heid and Weller 2008) have been used to describe collocations.

Such search strategies may lead to various errors. Regular expression search strings will involve a verb tag followed by a noun tag, typically at some distance, and possibly limiting the context between the verb and the noun tag. Such a search will report many samples. However, many of them will be incorrect (precision errors), and many verb-object relations will not be reported at all (recall errors).
Precision Errors: In sentences such as *Experts fear the Epstein Barr virus will spread.* the regular expression will erroneously report a verb-object relation between *fear* and *virus*. In sentences like *The report arrived Friday.* the regular expression will erroneously report a verb-object relation between *arrived* and *Friday*.
Recall Errors: In sentences such as *John likes swimming*. the regular expression will not find the verb-object relation, because swimming is not a noun, but a verb participle. In sentences like *John likes, but Mary hates Paul.* the regular expression will probably not find the verb-object relation, because the distance is relatively long, and the intervening part-of-speech tags are not the beginning of a noun phrase. Discarding such restrictions on intervening tags would probably lead to a precision error, erroneously reporting a verb-object relation between *likes* and *Mary*. In sentences such as *The potatoes I like are cold*. a regular

expression will not find the verb-object relation which is implicitly contained in the relative clause.

2.2 Windows-based approaches

Windows-based methods (e.g. Stubbs 1995) are still standardly used for collocation detection. N words before and after a key word, for example a verb, are considered. N is typically about 3. The distinction between different types of collocations, for example subject-verb, verb-object and verb-PP is typically left underspecified.

Precision Errors: In addition to the errors of the part-of-speech sequence method, windows-based methods suffer from precision errors due to the lack of implicit head extraction. In the example sentence *Experts fear the Epstein Barr virus will spread.* the windows-based method also reports *fear Epstein* and *fear Barr* as collocation counts.

Recall Errors: Recall is intrinsically low because many of the dependencies appear further then N words away. Recall can be increased by increasing N, but at a forbidding cost of decreasing precision.

2.3 Parsing Approaches

Parsing approaches are still rarely used for investigating collocations, which may be partly related to the fact that some definitions of collocations, in contrast to others, underspecify syntactic relations and are purely surface-based, for example as "sequences of lexical items that habitually co-occur" (Benson 1990). We take the view that syntax-lexis interactions is closely connected to individual syntactic functions and should abstract away from surface sequences as far as it is possible. We therefore base our investigation on syntactic functions. We present results on the subject and object function, but we hope to show that such an approach has a far wider potential.

A second major reason why parsing-based approaches are rare is that parsers which are enough robust, fast and performant, for example Collins (1999), Schneider (2007) or Andersen (2008), have only recently become available. Some of the few parsing approaches to collocation and multi-word expression (MWE) detection are Lin (1998) and Seretan and Wehrli (2006). Seretan and Wehrli (2006) have carefully evaluated their approach. They conclude that, in comparison to windows-based approaches, their parser-based system performs worse for the top-ranked collocations, but better in total.

> As for the MWE precision, the window method performs better for the first 100 pairs; on the remaining part, the parsing-based method is on average 3.7% better. The precision curve for the window method shows a more rapid degradation than it does for the other. Therefore we can conclude that parsing is especially advantageous if one investigates more than the first hundred results (as it seems reasonable for large extraction experiments). (Seretan and Wehrli 2006)

Although we have not conducted an extensive evaluation of our approach as a collocation and MWE finder, our preliminary results support Seretan and Wehrli: some consistent tagging and parsing errors are ranked very high, but low count instances contain considerably less garbage than windows-based approaches typically show. The majority of nonce occurrences are syntactically correct. While parsers make errors, the amount of errors is low enough so that collocation detection can also profit.

Some dependency types are especially hard to recover without parsing approaches: long-range dependencies and passive subjects. Long-range dependencies span more than 5 words in the majority of cases, passive subjects are difficult for a number of reasons: (1) subject-verb distances are often much longer than verb-object dependencies. (2) the recognition of passive forms by means of window methods or regular expressions is difficult, (3) passive verb forms are typically one word longer than active verb forms, introducing additional recall and verb head extraction errors for windows-based methods. The passive subject undergoes selectional restrictions to the same degree as verb-object relations (internal argument), and the passive is construction is a marked construction, which makes it a particularly interesting object of research (cf. Heid and Weller 2008).

3. Data and Method

In this section, we describe the data and the methods that we have used. Our approach is based on a complete syntactic analysis of the entire British National Corpus (World-edition), henceforth BNC (Aston & Burnard 1998) and American newspaper corpora. In our study we will use the written component with 86.5 million words[1]. The American newspaper material used in this study was acquired from Bell and Howell and is available in standard CD editions. Neither the stop-word index nor the compressed format, in which the texts are available on the commercial CD-ROMs, are suitable for corpus linguistics. Upon request Bell and Howell provided the material in super text format, an easily processable ASCII format. The annotation scheme used indicates sections like sports or international news. It also distinguishes the structural elements heading, lead paragraph and body. The present study is based on editorial content of the 1999 issues of *The Boston Globe* (36.3 million words) and *The Times* (35.4 million words). The two papers were selected because they were both available in the annotated super text format and because they represent American and British quality papers. It would have been desirable to base the study on newspaper

[1] Word counts for all corpora are based on a token-count of the tagged corpora excluding punctuation. These numbers may differ from other approaches. However, due to the consistency of the counting method across all our data, they form a solid basis for comparison.

material from the year 1993, the year the BNC was sampled. However, the material for that period was not available in an annotated electronic format.
Our processing pipeline consists of tagging, chunking, head-extracting and parsing the corpus, the parsed material is imported and queried in a database. In the following, we describe the processing steps in detail, and illustrate with examples:

1. In the early days the stigma of being HIV positive had driven away about 60% of my circle of friends. (BNC A00:189)

First, the corpus is tagged and the morphological base form, the lemma, is reported. For this step we have used the decision-tree tagger *Treetagger* (Schmid 1994)[2]. We have chosen to discard the part-of-speech tags included in some of the corpora, for example in the BNC, for reasons of consistency and cross-corpus comparability. Taggers assign morphosyntactic part-of-speech information to each word in the input text. We use the Penn Treebank Tagset (Marcus et al. 1993). After the first step, each word form is followed by the lemma and the tag, separated by commata, the example sentence looks as follows:

2. In_in_IN the_the_DT early_early_JJ days_day_NNS the_the_DT stigma_stigma_NN of_of_IN being_be_VBG HIV_hiv_NNP positive_positive_NN had_have_VBD driven_drive_VBN away_away_RB about_about_IN 60%_CARD_CD of_of_IN my_my_PRP$ circle_circle_NN of_of_IN friends_friend_NNS

In the second step, noun groups and verb groups are recognised by means of a chunker. We use the conditional random fields chunker *Carafe*[3]. After the second step, verb groups and noun groups are marked by double square brackets, the example sentence looks as follows:

3. In_in_IN
[[the_the_DT early_early_JJ days_day_NNS]]
[[the_the_DT stigma_stigma_NN]]
of_of_IN being_be_VBG HIV_hiv_NNP positive_positive_NN
[[had_have_VBD driven_drive_VBN away_away_RB]]
about_about_IN
[[60%_CARD_CD of_of_IN my_my_PRP$ circle_circle_NN]] of_of_IN
friends_friend_NNS

In the third step, the heads of the chunks are extracted. The head of a verb chunk is typically the rightmost verb, the head of a noun chunk is typically the rightmost noun. Then, the corpus is syntactically analysed. The parser *Pro3Gres,* which we

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[3] http://sourceforge.net/projects/carafe

use, is dependency-based, it reports syntactic functions arranged in a tree structure. The parser is very fast and robust, it parses the entire BNC in little over 24 hours. It has been applied in many areas of research, for example information retrieval (Bayer et al. 2004), relation mining in Biomedicine (Rinaldi et al 2007) and psycholinguistics (Schneider 2005). It has been developed by one of the authors and is described in detail in Schneider (2007)[4]. A screenshot of the graphical output of the dependency tree (we exclude relations inside chunks for simplicity) for the example sentence is given in figure 1.

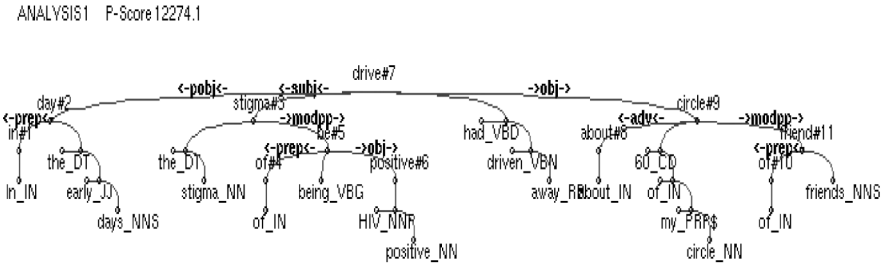

Figure 1. Output of the parser Pro3Gres for our sample sentence (BNC A00:189).

The syntactic analysis of the example sentence conveys, for example, that *drive,* the head of the verb chunk *had driven away*, is the main verb which attaches a prepositional phrase (relation *pobj*) and has a subject (*subj*) and an object (*obj*). The object, which is headed by *circle*, is modified by a prepositional phrase (*modpp*). Inevitably, steps 1 to 3 of our method introduce a certain amount of errors, which affects the results of our experiments. For a detailed analysis and evaluation see section 4.

The fourth step concerns accessing this richly annotated data. The parsed corpora were imported into a large database. We used Prolog to extract the selected data from the corpus and mySQL for storing the data. For each head of predicate identified by the parser, the database contains one record with cells describing the properties of the predicate head as well as its dependents; i.e. head of subject, head of object and PPs with preposition and description noun. For each of these we stored the word-form, the lemma, the part of speech, the direction of the dependency and the position in the sentence. In addition, the predicate is annotated for voice and finiteness. The analysis of the corpora results in databases with 10.5 million records for the written component of the BNC, 4.1 million records for *The Times* and 4.5 million records for *The Boston Globe*. These databases form the basis for our investigations described in section 5.

The choice of tools used for such a detailed linguistic investigation can have an impact on results. Some of the presented ranked lists are affected by tagging and parsing errors. Different taggers and parsers often make similar mistakes and have similar error rates, so that using a different tagger and parser will probably have little influence. Training a tagger over large manually annotated corpora

[4] http://www.ifi.uzh.ch/cl/gschneid/parser/

from the individual domains would improve the results, but such corpora are not available yet.
While tagsets are standardised, different chunkers may follow different policies. The chunker that we use, *Carafe*, takes a very greedy and semantic approach, as we illustrate in the following. Chunkers typically return noun groups (which are typically unnested NPs) and verb groups. In the sentence

4. The official spokesperson of Bogus Ltd. remained silent.

we have the noun groups *the official spokesperson* and *Bogus Ltd* and the verb group *remained*. There are a number of syntactic configurations where different chunkers report different chunks, however. In the sentence

5. One of the official spokespersons of Bogus Ltd. wanted to remain silent.

our chunker reports the noun groups *one of the official spokespersons* and *Bogus Ltd* and the verb group *wanted to remain*, while many less greedy chunkers report *one* and *the official spokespersons* as separate noun groups and *wanted* and *remain* as separate verb groups. The greedy chunking option typically coincides with being more semantic and less syntactic in nature. The greedy chunker reports a subject relation between *spokesperson* and *remain* in both sentences, abstracting away from surface syntax, while a non-greedy chunker would report a subject relation between *one* and *want* in sentence 5 (and probably a long-range dependency between *one* and *remain*). While our investigation of lexical semantics supports a semantic chunking policy, research on modal verbs would warrant the use of a non-greedy chunker.

4. Parser Evaluation and Error Handling

All automatic annotations face the problem that they are error-prone. Evaluation of the performance of automatic annotations is a major research topic in computational linguistics. We report in detail which error rates the parser has, including errors that stem from processing steps prior to parsing, such as tagging and chunking. First, we motivate why an extensive evaluation of automatic tools is essential for corpus linguistics and quantitative language description in general. Second, we introduce standard evaluation methodology. Third, we give a general evaluation of the parser. Then we give an evaluation of the relations that we have used in our research from sufficiently large random samples of the BNC. Finally, we discuss methods to cope with certain rates of error.

4.1 The Need for Extensive Evaluation

The use of automatic taggers, which introduce about 2-5% errors on average per token, depending on the tagset, the tagger, and the text type, is widespread in corpus linguistics, and this level of error rate is often tacitly acknowledged. While

such a low error rate poses no problem to most frequency-based linguistic research, one needs to consider that errors need not be spread homogenously over the tagset. While some tags reach a performance of 99%, others may have much lower performance. The distinction between prepositions (tag *IN* in the Penn tagset) and verbal particle (tag *RB* in the Penn tagset) is particulrarly difficult, because the context often looks identical. Some taggers achieve only 10% recall and 84% precision on this distinction (Baldwin and Villavicencio 2002). Research on verb particles which is based on tagger output may thus be seriously affected, despite the low error rate on average per token.

A crucial step for assessing the effect that errors are causing thus is to carefully evaluate the performance of the used tools, not only on a general level, but particularly on the linguistic phenomena under investigation, and on the actual corpora. Such evaluations are also known as selective evaluation. Lin (1995) introduces the selective evaluation method for dependency parsers which we use here. An extensive selective evaluation allows one to extrapolate to the number of false positives (precision errors) and to the number of missed instances (recall errors) within reasonable limits. For readers who are not acquainted with the notions of precision and recall, we briefly review them in the following.

4.2 Precision and Recall Errors in Automatic Annotation

The main challenge in using automatically parsed data is the fact that the data contains errors, which means that the analyses reported cannot simply be taken at face value. The success rate of an automatic annotation tool is measured in terms of precision and recall. In an evaluation, one carefully annotates a small random selection (the so-called gold standard) and compares it to the automatic annotation. Precision reports how many of the automatically annotated instances are also contained in the gold standard, i.e. are correct. Recall reports how many of the instances in the gold standard are actually found in the automatic annotation. There is typically a trade-off between precision and recall. For example, a system that automatically annotates only the few “easy cases” of whatever phenomenon has high precision but low recall.

Precision errors can be filtered relatively easily, unless the amount of data reported is too big, by going through the output of an automatic system and discarding the false positives, which are often referred to as unwanted instances or as garbage. This filtering process is a classical corpus linguistics approach. Our data contains several million instances, so that filtering is not an option. Instead, we use the results of our evaluation as an estimate of the number of false positives that we need to expect from our tools.

Recall errors, also known as missed instances, are an even more serious problem. The only perfect way to know what the automatic annotation tool misses is to go through the entire corpus manually. The utter lack of knowledge of what was missed makes it impossible to extrapolate from data based on automatic annotation. With a selective evaluation of the tool’s performance – not just for overall performance but for the specific research question – the corpus linguist can, however, extrapolate from the missed instances in the gold standard to the

new material processed in their research. The careful evaluation of our tools' performance is therefore an essential part of our approach.

4.3 Standard Test Corpus Evaluations

A number of manually annotated corpora are standardly used to compare the performance of syntactic parsers. One of them is the GREVAL corpus (Carroll et al. 2003), which contains 500 near-random sentences from the Suzanne corpus, covering a broad range of news texts. Performance of the Pro3Gres parser on the relations subject, object, PP-attachment to verb and PP-attachment to noun are given in table 1.

Table 1. Performance on Pro3Gres on the 500 GREVAL sentences

Peformance on GREVAL	Subject	Object	Noun-PP	Verb-PP
Precision	92%	89%	74%	72%
Recall	81%	84%	66%	84%

For more detailed evaluations, including a complete mapping to the relation set used in GREVAL and a comparison to several other syntactic parsers, refer to Schneider (2007).
The parser has also been evaluated on one of its application areas, on biomedical texts. 100 random sentences from the domain have been manually annotated and compared to the parser output. The performance numbers are reported in table 2. An independent evaluation mapping Pro3Gres output to the Stanford dependency scheme has been conducted in Haverinen et al. (2008), confirming that the parser has state-of-the-art performance.

Table 2. Performance of Pro3Gres on 100 random biomedical literature sentences

Peformance on GENIA	Subject	Object	Noun-PP	Verb-PP
Precision	90%	93%	85%	82%
Recall	87%	91%	82%	84%

4.4 Evaluation on the BNC

The above evaluations show that for some of the argument structure relations, particularly subject and object, error rates lie between 10% and 20%.

4.4.1 Subjects and Objects

In order to test if these error rates carry over to our application corpora, for example the BNC, we have manually annotated a small random selection of 100 spoken and 100 written sentences from the BNC. Performance results are given in table 3.

Table 3. Performance of Pro3Gres on 100 random sentences from the BNC

	BNC written		BNC spoken (conggov)	
	Percent	Count	Percent	Count
Subject precision	86%	108 / 125	88%	125 / 140
Subject recall	83%	108 / 130	89%	125 / 142
Object precision	87%	71 / 82	78%	70 / 90
Object recall	88%	71 / 80	87%	70 / 80

The results are similar to those obtained on the standard test corpora. Performance on the spoken corpus, particularly objects, is affected by our current rudimentary way of filtering hesitation markers (*errm* etc.) and can be expected to improve with a better filtering algorithm.

4.4.2 Passive Subjects

One of the applications that we will discuss in section 5 involves passive subjects, a subgroup of subjects that has special characteristics and may show a different performance. A separate evaluation is thus appropriate. The 100 random sentences from the BNC contained only 5 passive subjects in the spoken and 14 in the written part, these counts are too low to allow a reliable evaluation. In order to attain sufficiently large counts, 100 random sentences from the written BNC which contain verb participles (Penn tag *VBN*) were manually annotated for passive subjects and passive verb forms. The performance thus found is given in table 4. The passive subject error rate (table 5 on the left) is similar to the general subject error rate, although slightly lower. Some passive subject errors are due to the fact that our passive verb form recognition algorithm has only 91% precision and 92% recall, as given in table 4 on the right. Since our random selection method of using only sentences containing a *VBN* tag tacitly assumes that the tag *VBN* is always correct, passive subject performance can in reality be expected to be 1-4 percent below the figures in table 4.

Table 4. Passive subject evaluation, based on 100 random BNC sentences containing verb participles.

	Passive Subject BNC written		Passive Verb Forms BNC written	
	Percent	Count	Percent	Count
Precision	85%	58 / 68	91%	60 / 66
Recall	82%	58 / 71	92%	60 / 65

4.4.3 Local and nonlocal Subjects and Objects

We have hitherto assumed that everybody knows what subjects and objects are. Although we use standard terminology, a defintion is called for. We use the term subject to either denote the explicite subject of a finite verb, or the implicit

subject of an infinite or finite verb. Implicit subjects of finite verbs are relative pronoun resolutions (for example *Girls who like boys*, where *girls* is implicit subject of *like*), implicit subjects of infinite verbs are control structures (for example *Peter is unable to win*, where *Peter* is the implicit subject of *win*). Other types of implicit subjects, for example pronoun resolution (for example *Peter sleeps and he snores*, where *he* is implicitly *Peter*) or indexed gerunds (for example *Peter entered, cheering*, where the implicit subject of *cheer* is *Peter*) are not returned by our parser. The definition of objects is analogous.
The implicit subjects and objects our parser reports are so-called nonlocal dependencies. Nonlocal dependencies are also termed long-distance or long-range dependencies. For example in sentence 6, *procedure* is the implicit subject of the verb *transfer*, and *value* is the implicit subject of *indicate.*

6. The ***procedure*** does not wait for offline modules to be ***transferred***, however a ***value*** is returned in MODULES-ONLINE to ***indicate*** whether any modules are offline awaiting transfer. (BNC HWF:4093)

In this case, the nonlocal dependencies are so-called subject-control relations. Nonlocal dependencies are more difficult to detect by automatic parsers. Separate performance values on the 100 BNC written random sentences (see table 3) broken down into nonlocal and local relations are given in table 5. While the counts on nonlocal relations are too low to deliver reliable results, local subject and object relations are shown to achieve almost 90% precision and recall.

Table 5. Performance of local and nonlocal relations on 100 sentences from the BNC

	Local relations BNC written		Nonlocal relations BNC written	
	Percent	Count	Percent	Count
Subject precision	89%	102 / 104	55%	6 / 11
Subject recall	86%	102 / 119	55%	6 / 11
Object precision	89%	70 / 79	33%	1 / 3
Object recall	89%	70 / 79	100%	1 / 1

4.5 Error Handling

We have shown that error levels between 10% and 20% for the subject and object relations appear consistently both in the standard evaluation corpora as well as in actual BNC data. While such error levels are too high to e.g. report absolute numbers reliably, we suggest that, based on a careful selective evaluation, limited scientific statements are possible, and that it is possible to quantitatively extrapolate to false positives (garbage, precision errors) and to missed instances (recall errors) within reasonable limits.

4.5.1 Large differences, lower error rates

If we want to measure the quantitative difference between two corpora, which we will refer to as A and B, for example to describe sociolinguistic, genre-specific, or dialectal variation, we can reliably assume that there is a difference whenever the difference is larger than the error rate. In other words, if we made the extreme assumption that corpus A has the error rate reported in the random subset evaluation and corpus B had no errors at all, then a real difference between A and B must exist if the differences are bigger than the error rate found in the evaluation.

Let us look at an example to illustrate how one can test if such differences are statistically significant. Table 6 shows important syntactic relation counts from random samples from BNC science on the left and BNC imaginative writing and leisure on the right. Assuming a (slightly simplified) success rate of 80% we can use a chi-square test as follows:

if the smaller value / 80% > larger then let smaller = larger

else let smaller = smaller / 80%

The differences in table 6 are still statistically significant with $P < .0001$.

Table 6. Important syntactic relation counts from 2 BNC genre subsets

	Science	Imag. & Leisure
Relation	Count	Count
Subj	22467	28128
Obj	13990	15882
Verb-PP	13412	14348
Noun-PP	14637	10035
modpart	1453	1262
modrel	1911	1286

The relation *modrel* attaches a relative clause to the noun which it modifies, the relation *modpart* (modification by participle) expresses a reduced relative clause. These two relations are very frequent in the scientific genre.

4.5.2 Expectation of similar error rates

If we can expect similar error rates for two constructions, then one can directly compare frequencies based on parsed data.

First, a negative example: if we wanted to investigate the alternation of agents between active and passive verbs, we cannot expect the same error rates for active verb subjects (80-90%) and for the agent in a PP introduced with *by* in passive verb constructions, because verb-PP attachment has only about 74% precision and about 85% recall (see table 1).

Second, a positive example: if we want to investigate the distribution of objects and verbs inside the object relation, we only need to assume similar error rates for

most lexical items. This assumption holds in most cases, unless the tagger produces an error, as in the following example:

7. This_DT would_MD include_VB ,_, inter_VB alia_NNP ,_, the_DT Take-over_NNP Code_NNP ._SENT (BNC ECD:1637)

The rare word *inter* is consistently mistagged as verb, the equally rare *alia* as proper noun. This has the effect that *inter alia* is reported as a strong verb-object collocation. In the majority of analysed cases, the assumption holds, and we can compare lexical preferences.

5. Exploring the syntax-lexis interface

In section 2 we have described the extraction of a database containing verbs and their dependent subjects, objects and PPs. In this section we are exploring the wealth of data contained in these databases. We will focus on predicates and their subject and object dependencies[5]. The main interest driving our research is a quantitative analysis of the interaction between syntactic structures and the lexicon. We extract and measure lexical preferences in the cline from free choice to collocation and structural preferences in the active-passive alternation using customised databases and statistical measures of surprise.

Any study of the interaction between lexical choices and syntactic choices in subject and object NPs and their governing verbs will have to take into account the active-passive alternation. Lexical choices will heavily depend on thematic roles. The use of parsed data allows us to deal with subjects of active verbal constructions in separation from subjects in passive constructions. In the same way, we can limit our observations to objects in active constructions.

Windows-based approaches to syntax-lexis interaction (e.g. Stubbs 1995) take into account all content words present in the vicinity of each other (inside the observation window), irrespective of their syntactic function, and irrespective whether they are syntactically connected at all. As a consequence such non-hierarchical approaches are forced to base the expected value (E, null-hypothesis) on the assumption of a corpus in which the words appear in random order. As Evert (2008) points out such a null-hypothesis is not unproblematic.

> …the null hypothesis of independence is extremely unrealistic. Words are never combined at random in natural language, being subject to a variety of syntactic, semantic and lexical restrictions. For a large corpus, even a small deviation from the null hypothesis may lead to highly significant rejection and inflated association scores calculated

[5] We are aware of the intriguing possibilities offered by the dependent PPs stored with preposition and description noun in our database. However we feel that such an analysis is beyond the scope of this paper.

> by significance measures. Effect-size measures are also subject to this problem and will produce inflated scores, e.g. for two rare words that always occur near each other (such as *déjà* and *vu*). A possible solution would be to specify a more realistic null hypothesis that takes some of the restrictions on word combinatorics into account, but research along these lines is still at a very early stage. (Evert 2008)

For these reasons, we avoid a null hypothesis (E) based on a random shuffling of words. Instead, we base our expectation on a random shuffling of the head verbs, and head nouns actually observed in the subject-verb and verb-object dependencies in our data. This offers a more realistic expectation, by taking the syntactic restrictions on word combinatorics into account. The observed lexical head-verb preferences inside a fixed structure allow us to investigate selectional preferences and native-like selection (Pawley and Syder 1983). Comparing the use of the same head verbs and head nouns across different syntactic variants also permits us to investigate alternation preferences.

Our approach has the advantage of focusing on the construction under examination; for example verb-object relations ignoring frequent lexical items found in PPs and other constructions, which would skew our results otherwise.

We use O/E in the following because it has a clear probabilistic definition and is directly related to information-theoretic measures of surprise such as mutual information.

Unlike many other approaches we use the lemma and not the word-forms in producing generalizations about the distribution of lexical items. On the one hand, this introduces a higher degree of generalization; on the other hand, we may miss some interesting phenomena concerned with the co-occurrence of word-forms. The present empirical setup would easily allow for an investigation based on word-forms. However, in this paper we will focus on the lemmas for establishing types of structural co-occurrence.

In order to avoid subject complements, all our calculations exclude the lemmas *be* and *seem* as head of predicate.

In the case of subject-verb combinations we calculate the measure of surprise on the basis of a random shuffling of all subject heads and predicate heads found in the analysed corpus. This will result in an expected frequency for individual subject-verb combinations, E. The observed frequency of each subject-verb combination, O, is then used to calculate the measure of surprise, O/E, which expresses the factor by which the actual occurrence of the combination exceeds the expected frequency. Given the focus on subject-verb combinations the measure of surprise here does not express the surprise of finding individual lexemes in subject position in general. The expectation is based on the assumption of free combination within the limits of the syntactic construction, i.e. any subject head can combine with any predicate head. In other words E expresses free selection as opposed to selectional restriction.

Table 7. Selectional restriction in active subject-verb combinations, f(sub_verb) > 50. BNC-world written component.

subject_verb heads	f(sub_verb)	f(subject)	f(verb)	O/E
tentacle_pore	77	136	243	8.78576e+10
onion_chop	56	139	776	1.9577e+10
egg_hatch	58	396	456	1.21116e+10
doorbell_ring	65	80	4220	7.26013e+09
interview(s)_record	136	136	5389	6.99722e+09
bomb_explode	158	652	1326	6.89128e+09
rumour_circulate	56	402	848	6.19441e+09
telephone_ring	195	356	4220	4.89447e+09
dog_bark	81	1739	506	3.47111e+09
phone_ring	142	374	4220	3.39264e+09
relation_deteriorate	62	950	738	3.33461e+09
sun_shine	294	1981	1690	3.31139e+09
lip_part	63	825	872	3.3022e+09
lifespan_display	111	420	3391	2.93886e+09
god_bless	135	3349	603	2.52078e+09
wind_blow	239	1381	2986	2.18549e+09
thief_steal	103	590	3015	2.18339e+09

Table 7 shows the tendency of subject heads to co-occur with certain verb heads in subject-verb constructions. While as linguists we will not be surprised of the fact that dogs bark, phones ring and thieves steal, this data reminds us that selectional restrictions not only occur with internal arguments, but also with external arguments.

As we have shown in section 4, the annotation process is far from infallible. The combination *tentacle pore* is a case where the tagger failed and assigned a verb reading to *pores* in the compound *tentacle pores*. The combination *onion chop* is produced by a parsing error that failed to interpret postmodification by participle in sentences like 8, found in a recipe.

8. 1 medium onion, chopped. (BNC BPG:1001)

Sentences 9 and 10 show that the parser not only introduces sources of error but also contributes considerably to the coverage of our approach by including long-distance dependencies.

9. For a number of reasons, however, the ***eggs*** failed to ***hatch***. (BNC AM2:103)

10. Here the female may produce 800,000 ***eggs*** which ***hatch*** within 36 hours into larvae. (BNC A3Y:43)

The list in table 7 has to be seen as raw material for closer analysis. It could for instance be used for extracting dictionary entries together with typical examples.

Table 8. Selectional restriction in passive subject-verb combinations, f(sub_verb) > 50. BNC-world written component.

subject_verb heads	f(sub_verb)	f(subject)	f(verb)	O/E
seed_sow	51	172	147	1.6762e+09
shot_fire	89	233	474	6.69664e+08
lesson_learn	73	217	604	4.62836e+08
duty_owe	59	431	282	4.03391e+08
battle_fight	55	235	521	3.733e+08
breakfast_serve	55	113	1375	2.94159e+08
offence_commit	193	398	1459	2.76198e+08
warrant_issue	73	150	1511	2.67651e+08
day_number	62	989	207	2.51667e+08
power_vest	124	1564	274	2.40456e+08
battle_win	53	235	788	2.37839e+08
prise_award	72	192	1345	2.31691e+08
attention_focus	149	1145	486	2.22508e+08
war_fight	57	427	521	2.12917e+08
treaty_sign	92	330	1160	1.99718e+08
reliance_place	52	69	3175	1.97248e+08
car_park	68	1121	263	1.91668e+08
interview_conduct	60	232	1189	1.80752e+08

Table 8 shows the association between subject and verb in passive constructions; seeds are sowed, shots are fired and lessons learnt. In terms of the active passive alternation it is more interesting to compare passive subjects with active objects than with active subjects.

Table 9. Selectional restriction in active verb-object combinations, f(sub_verb) > 50. BNC-world written component.

verb-object heads	f(sverb_obj)	f(verb)	f(subject)	O/E
inter_alia	259	348	269	8.26219e+10
wreak_havoc	88	157	248	6.7493e+10
whet_appetite	70	85	419	5.86938e+10
rick_sky	83	91	500	5.44746e+10

extol_virtue	56	132	379	3.34274e+10
programme_tdy	55	1068	55	2.79612e+10
clench_fist	82	399	403	1.52287e+10
beg_pardon	145	1320	216	1.51868e+10
grit_tooth	146	227	1363	1.40915e+10
purse_lip	135	184	1680	1.30417e+10
wrinkle_nose	82	202	1039	1.16674e+10
bridge_gap	162	321	1367	1.10247e+10
sow_seed	107	469	637	1.06954e+10
heave_sigh	74	512	430	1.00374e+10
buck_trend	58	184	1004	9.3757e+09
enclose_sae	68	1148	211	8.38324e+09
ratify_treaty	99	419	859	8.21401e+09
reap_reward	73	394	680	8.13664e+09

Table 9 shows the top 18 verb-object combinations. The combinations *inter alia*, *rick sky* and *programme tdy* are reported due to tagging errors. Rather than as result in itself this list can serve as raw material and starting point for further investigation. We also find similar items to the ones found in table 8; e.g. *sow seeds* vs. *seeds are sowed*, where we find both alternations, active and passive.
Due to the size of the corpus and the extended coverage provided by the parser it is possible to investigate the semantic prosody of low frequency items. Taking the combinations *wreak havoc* and *extol virtue* as a starting point, table 10 shows the semantic prosodies of *extol* and *wreak*.

Table 10. Semantic prosody of extol and wreak.

extol		*wreak*	
verb-object	*n*	*verb-object*	*n*
extol_virtue	56	wreak_havoc	88
extol_benefit	5	wreak_vengeance	10
extol_beauty	4	wreak_revenge	9
extol_man	2	wreak_destruction	5
extol_courage	1	wreak_damage	4
extol_approach	1	wreak_kind	2
extol_brilliance	1	wreak_mayhem	2
extol_authority	1	wreak_assault	1
extol_success	1	wreak_distortion	1
extol_achievement	1	wreak_pain	1
extol_leader	1	wreak_carnage	1
extol_riches	1	wreak_spite	1

The dependency database developed for the purpose of this study allows for the exploration of phenomena like semantic prosody. It also provides links to the

original corpus. Sentence 11 shows an example where the semantic prosody of wreak is used for creating a contrary effect.

11. David Baddiel extolling hardcore porn? (TLN955826475)

Table 10, as all our co-occurrence tables, is truncated. Unlike other tables it contains nonce occurrences. Our approach based on parsed corpora provides astonishingly clean data even at extremely low levels of frequency. Given these promising result we may investigate phenomena occurring at even lower frequencies.

It is typically recommended to avoid using O/E, the statistical measure that we have used here, because it has the tendency to rank combinations where both words are rare very high. In windows-based and tag-sequence based approaches, this has the undesirable side effect that false positives dominate a large area at the top of the lists. The parser-based approach suffers from this to a much lesser degree, opening up new possibilities for investigating combinations of rare words.

Table 11 shows the measure of surprise of finding subject-verb-object triplets. We filtered out occurrences of *page omitted advertisement* and *page omitted photograph*, which quite obviously are due to coding errors in the BNC corpus, where they are not consistently set off as encoding comments.

Table 11. Fixedness in active subject-verb-object combinations, f(svo) > 20. BNC-world written component.

subject-verb-object heads	f(svo)	f(s)	f(v)	f(o)	O/E
coroner_record_verdict	33	284	5389	517	6.08756e+12
spine_form_fan	23	113	12659	547	4.29051e+12
heart_miss_beat	26	1590	7393	222	1.45428e+12
clause_exclude_liability	25	744	3061	1126	1.42302e+12
jury_return_verdict	45	669	16513	517	1.15005e+12
sale_start_monday	23	1667	17732	159	7.14306e+11
female_lay_egg	29	683	6985	1317	6.73708e+11
sale_start_december	31	1667	17732	250	6.12315e+11
republic_achieve_independence	24	596	11145	1130	4.66717e+11
error_occur_error	21	583	13564	961	4.03354e+11
court_grant_injunction	22	7229	3247	345	3.96543e+11
inc_report_profit	145	1802	11740	2711	3.69031e+11
plc_report_profit	22	288	11740	2711	3.50332e+11
index_close_point	204	947	13999	8451	2.6578e+11
tenant_pay_rent	21	865	23676	666	2.24733e+11
corp_report_profit	65	1380	11740	2711	2.16015e+11
price_include_breakfast	195	3462	45097	948	1.92308e+11
history_repeat_itself	29	1186	3841	6172	1.50553e+11

Such triplets would be extremely difficult to retrieve with window based or pattern based approaches and the low number of instances found in a 90 million word corpus shows the advantage of our parser based retrieval. Instances like sentence 12 are extremely difficult to locate with non-hierarchical strategies.

12. It was a foregone conclusion that the ***jury***, carefully selected beforehand, would ***return*** their immediate and unanimous ***verdict*** of "Guilty". (BNC ALK:796)

In the following we are focusing on the study of the active-passive alternation and its interaction with lexical choices. We decided to investigate cases where the same pair of lexical items is involved in active as well as passive constructions as in *sow seeds* vs. *seeds are sowed*. Table 12 shows a ranking of such pairs ordered according to their preference for passive constructions.

Table 12. Preference for passive constructions for word pairs occurring in alternation in the written BNC. f(active)>2, f(passive>2, f(total)> 100

pair of lemmas	**f(active)**	**f(passive)**	**f(total)**	**% passive**
baby_bear	3	141	144	97.9167
study_carry	4	118	122	96.7213
committee_set	5	137	142	96.4789
power_vest	6	124	130	95.3846
test_carry	7	100	107	93.4579
research_carry	10	125	135	92.5926
system_base	10	106	116	91.3793
work_carry	29	274	303	90.429
example_show	47	253	300	84.3333
case_adjourn	23	94	117	80.3419
election_hold	112	442	554	79.7834
people_arrest	31	113	144	78.4722
detail_obtain	33	105	138	76.087
decision_base	45	118	163	72.3926
people_injure	45	102	147	69.3878
detail_find	43	95	138	68.8406
service_hold	38	78	116	67.2414
soldier_kill	38	75	113	66.3717

Table 12 shows the top 18 word pairs in terms of preference for the passive. The restriction to pairs occurring more than 100 times ensures a minimal number of observations for the comparison between active and passive. The restriction to pairs that occur at least 3 times in the active as well as in the passive is applied in order to limit our observation to pairs for which the active passive alternation is relevant.

The top-ranked pair *baby* and *bear* shows a massive preference for the passive. More than 97 percent of all observations occur in the passive, as in sentence 14. Active constructions, as in sentence 13, are extremely rare.

13. They say drug-abusing mothers who would previously have had an abortion are ***bearing*** sickly ***babies*** with low chances of survival. (BNC A1G:487)

14. Overall, three in ten ***babies*** are ***born*** outside marriage in the UK to mothers of all age groups. (BNC K3S:91)

The results in table 12 also reflect our conservative approach to MWEs (multi word entities). Phrasal verbs and their particles are analyzed as separate word tokens. As a consequence the predicate head *carry* represents both *carry out* as well as *carry forward*.
The reported instances at the top of the list show a marked difference for the distribution of active and passive constructions, which is generally found at a level of 6-13%, depending on text-type. However, we have to take into account the restriction of our observation to 100 occurrences in total, with at least 3 occurrences for both variants. Combinations that do occur only in the active or only in the passive voice are excluded. Based on the average of the pairs actually observed in table 12 we expect 16 percent of the instances realised as passives. Given the range of observed passive percentages from 98% for *bear baby* down to 0.6% for *make sense*, we can observe a strong interaction between active-passive constructions and lexical choices. The fact that in table 12 we only consider the middle ground between pairs that exclusively occur in either the active or the passive voice makes this observation more remarkable. To complete the picture at both ends of the cline we present the pairs at the extreme ends in table 13.

Table 13. Active verb-object pairs that do not have a passive counterpart and passive verb-subject pairs that do not have an active counterpart

exclusively active		**exclusively passive**	
object verb	**n**	**verb subject**	**n**
time_have	3620	scroll_area	45
chance_have	2106	situate_hotel	43
power_have	1980	approve_study	43
way_go	1456	know_little	39
difficulty_have	1434	bear_william	38
interest_have	1353	base_figure	36
access_have	1229	base_diagnosis	34
opportunity_have	1228	base_some	32
lot_have	1207	call_fireman	30
impact_have	1185	enter_correspondence	29

reason_have	1100	age_cent	28
home_come	1065	set_council	28
choice_have	1017	wind_company	28
role_have	972	hand_judgment	25
money_have	970	situate_house	24
look_have	949	bear_george	24
sense_have	944	bear_thomas	23
implication_have	889	announce_date	23
influence_have	881	bear_james	22
job_get	851	suspend_share	22

Not surprisingly we find the middle verb *have* dominating the most frequent combinations not occurring in the passive voice. The combinations exclusively occurring in the passive are more varied and occur at very low frequency. Besides *be based on* we find *be born*, *be situated*, *be suspended* and *be announced*. Of course such passive verbs can be found in the active voice in our data, as shown in sentence 15 for *be situated*.

15. If you ***situate*** the *cable* tidy as close to your tank as possible , you can cut the wires on your equipment fairly short , to get rid of all those unsightly trailing cables . (BNC C97:1725)

A quick glance through the type list of objects occurring with *situate* showed no instance for an object of the type building, whereas a type list for the passive subject immediately reveals *hotel, house, village, premise, property, school, office, station, centre, college* etc.

In table 14 we list the verbs occurring in the combinations presented in table 13 in abstraction from the objects or subjects they occurred with.

Table 14. Frequency of verbs in verb-object combinations not occurring in the passive and of verbs in subject-verb combinations not occurring in the active voice in the written component of the BNC.

exclusively active		**exclusively passive**	
object verb	***n***	***verb subject***	***n***
have	136302	base	353
get	25436	bear	260
become	14054	associate	154
see	8305	situate	144
give	6172	set	123
do	5787	know	73
include	5090	approve	63
take	4983	deal	53
make	4672	scroll	45

want	4452	confine	44
come	3847	carry	41
feel	3813	concern	39
receive	3308	account	35
go	3053	hold	32
like	2811	call	30
meet	2708	enter	29
follow	2568	injure	29
show	2542	aim	28
leave	2521	age	28
allow	2438	wind	28

Among the verbs occurring in exclusively active combinations we find verbs that do not form the passive in general. These represent a bundle of features like stative vs. dynamic, agent subject vs. non-agent subject etc. However, we also find verbs with asymmetric preferences for passive subjects and active objects. Such asymmetries are due to a variety of causes. A closer analysis of these could be used for creating a corpus-driven classification of verbs. In some cases the view tacitly taken here that we observe the active-passive alternations as a system of only two possibilities is wrong. There may be semantically close variants, e.g. *decision be made* vs. *make decision* vs. *decide*.

The verbs occurring in exclusively active combination contain many semantically weak verbs. We can see that semantically weak verb combinations are subject to restricted flexibility.

In a next step we explore the active-passive alternation with regard to the different preferences for individual pairs of lexemes in the two major varieties, American and British English. For American English we have no comparable data at our disposal. We decided to use the editorial content of the year 1999 of the *The* [London] *Times* (TLND) and *The Boston Globe*. For both daily papers, we calculated the preferences for the passive construction, as shown in table 12 for the written component of the BNC. We then combined the two tables and calculated the difference between American and British English, shown in table 15. The top of the table shows passive preference in American English, the bottom preference for the passive in British English.

Table 15. Differences in the preferences for active and passive in *The* [London] *Times* and *The Boston Globe*. f(total) > 50.

	The Boston Globe		***The Times***		
	f(total)	***%passive***	**f(total)**	***%passive***	**diff %**
material_use	779	*94*	77	*18*	75
information_use	211	*66*	61	*21*	45
name_use	238	*39*	123	*18*	22
talk_hold	57	*25*	209	*7*	17
award_present	116	*53*	71	*37*	17

people_involve	84	*39*	62	*24*	15
bridge_build	83	*20*	54	*6*	15
order_issue	88	*28*	54	*15*	14
ground_break	183	*15*	81	*1*	14
man_kill	209	*51*	90	*38*	13
...	…	...	...	...	…
jury_tell	93	*9*	178	*33*	-25
advice_offer	92	*1*	231	*26*	-25
proceed_use	58	*34*	55	*60*	-26
case_bring	62	*15*	87	*40*	-26
people_put	102	*3*	116	*29*	-26
reference_make	75	*7*	115	*34*	-27
people_give	70	*17*	59	*49*	-32
shot_block	102	*5*	59	*37*	-32
charge_make	56	*9*	60	*45*	-36
legislation_introduce	53	*8*	64	*47*	-39
court_tell	59	*5*	933	*58*	-53

The differences observed may have extra-linguistic reasons. At the extreme ends of the scale we observe massive differences for the total frequency of the observed combinations. *court tell* occurs 59 times in *The Boston Globe* and 933 times in *The Times*. In the case of *charges make* we can exclude such a difference in overall frequency. Nevertheless we observe a marked difference in the preference for the passive. In *The Times* the combination *charges make* occurs in the passive in 47 percent of all cases whereas in *The Boston Globe* we only find 9 percent of the instances in the passive.

All the results presented here are of exploratory nature. Any of the phenomena explored could and should be studied in a more detailed analysis. As shown, the parser-based approach may help us in the study of identified co-occurrences, like *charge make*. However, the main impact for studying fixedness in language and the interface between syntax and the lexicon consists in the new possibility of a mainly corpus-driven rationale for the selection of individual co-occurrences in syntactic structures. While, as corpus linguists we can easily explore identified co-occurrences, we were largely limited to approaches with a lexical node for studying these. The selection of a lexical node itself was motivated by hunches, intuition and previous work done in the field of study.

Pattern based approaches may reach a similar recall at identifying specific phenomena (see e.g. Lehmann 1997), however, they tend to incur a very low level of precision, which severely limits the usefulness of the results. The only alternative for producing a more reliable ranking of the combinations, is manual annotation, which for the BNC would imply the manual creation of a database of more than 10 million records.

The selection of large corpora is not only a pretext for using a parser. The precariously low numbers in the cells of our tables clearly shows the necessity of analysing large corpora for this type of study.

6. Conclusions and Outlook

In this paper we have described the compilation of a verb dependency database and shown its potential in several areas of research. We have explored selectional preferences in subject-verb and verb-object combinations. For the active-passive alternation we have found and partly described the gradient area where lexical choices coincide with preferences for the active and passive construction. In the case of the active-passive alternation our approach allowed us to quantify the gradient lexico-grammatical phenomena at the interface between syntax and lexis on a largely corpus-driven basis. We have also explored the possibility of comparing these preferences in two major varieties of English. We have outlined several applications exploiting the data compiled for this study.

We are currently including further syntactic relations and investigate descriptions of English argument structure. Besides the exploitation of the database for lexicographical purposes like rich lexicon entries indicating combinatorial preferences, we see the potential of the database in corpus-driven studies of other alternations and the extraction of special classes like ergative verbs as future applications.

7. References

Andersen, Øistein E., Julien Nioche, Ted Briscoe and John Carroll, 2008. The BNC Parsed with RASP4UIMA. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

Aston, Guy and Lou Burnard, 1998. The BNC Handbook. Exploring the British National Corpus with SARA.Edinburgh: Edinburgh University Press.

Baldwin, Timothy and Aline Villavicencio, 2002. Extracting the unextractable: A case study on verb-particles. In Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL 2002), pages 98–104, Taipei, Taiwan.

Bayer, Samuel, John Burger, Warren Greiff and Ben Wellner, 2004. The MITRE Logical Form Generation System. In Proceedings of Senseval-3: The Third International Workshop on Evaluation of Systems for the Semantic Analysis of Text. Barcelona, Spain.

Benson, Morton, 1990. Collocations and general-purpose dictionaries. International Journal of Lexicography, 3(1):23–35. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, Mass.

Carroll, John, Guido Minnen, and Edward Briscoe, 2003. Parser evaluation: using a grammatical relation annotation scheme. In Anne Abeillé, editor, Treebanks: Building and Using Parsed Corpora, pages 299–316. Kluwer, Dordrecht,.

Collins, Michael, 1999. Head-Driven Statistical Models for Natural Language Parsing. University of Pennsylvani, Ph. D. thesis, Philadelphia, PA.

Evert, Stefan, to appear 2008. Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), Corpus Linguistics. An International Handbook, article 58. Mouton de Gruyter, Berlin.

Haverinen, Katri, Filip Ginter, Sampo Pyysalo and Tapio Salakoski, 2008. Accurate conversion of dependency parses: targeting the Stanford scheme. In Proceedings of Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland.

Heid, Ulrich and Marion Weller, 2008. Tools for Collocation Extraction: Preferences for Active vs. Passive. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

Hoffmann, Sebastian and Hans Martin Lehmann, 2000. Collocational Evidence from the British National Corpus. In Kirk, J. (Ed.), Corpora Galore. Analyses and Techniques in Describing English. Amsterdam/Atlanta: Rodopi. 17-32.

Lehmann, Hans Martin, 1997. Automatic Retrieval of Zero Elements in a Computerised Corpus. In Ljung, M. (Ed.), *Corpus-based Studies in English. Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora.* Amsterdam: Rodopi. 179–194.

Lin, Dekang, 1995. A dependency-based method for evaluating broad-coverage parsers. In Proceedings of IJCAI-95, Montreal.

Lin, Dekang, 1998. Extracting collocations from text corpora. In Proceedings of First Workshop on Computational Terminology, pages 57–63, Montreal, Canada.

Pawley, Andrew and Syder, Frances Hodgetts (1983). Two Puzzles for Linguistic Theory: Native-like selection and native-like fluency. In Richards, J. C. & Schmidt, R. W. (Eds.), *Language and Communication*. London: Longman. 191–226.

Rinaldi, Fabio, Gerold Schneider, Kaarel Kaljurand, Michael Hess, Christos Andronis, Ourania Konstanti and Andreas Persidis, 2007. Mining of Functional Relations between Genes and Proteins over Biomedical Scientific Literature using a Deep-Linguistic Approach. *Journal of Artificial Intelligence in Medicine*, 39: pages 127-136.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger, 2001. Multi-word expressions: A pain in the neck for NLP. Technical Report LinGO Working Paper No. 2001-03, Stanford University, CA.

Schmid, Helmut, 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing.

Schneider, Gerold, 2007. Hybrid Long-Distance Functional Dependency Parsing. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Schneider, Gerold, Fabio Rinaldi, Kaarel Kaljurand and Michael Hess, 2005. Closing the Gap: Cognitively Adequate, Fast Broad-Coverage Grammatical Role Parsing. ICEIS Workshop on Natural Language Understanding and Cognitive Science (NLUCS 2005, Miami, FL.

Seretan, Violeta and Eric Wehrli, 2006. Accurate collocation extraction using a multilingual parser. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 953–960, Sydney, Australia, July. Association for Computational Linguistics.

Smadja, Frank, 1993. Retrieving collocations from text: Xtract. Computational Linguistics, 19(1):143–177.

Stubbs, Michael, 1995. Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language*, 2, 1: 23-55.