Using automatically parsed corpora to discover lexico-grammatical features of English varieties

Schneider, Gerold

Gerold Schneider
Institute of Computational Linguistics and English Department
University of Zurich
gschneid@es.uzh.ch

USING AUTOMATICALLY PARSED CORPORA TO DISCOVER LEXICO-GRAMMATICAL FEATURES OF ENGLISH VARIETIES

Abstract

We employ syntactic parsing to describe and to discover lexico-grammatical features of English regional varieties. In the absence of suitable Treebanks, automatically parsed corpora (tree jungles) can be used. As an example we focus on Indian English, using the International Corpus of English (ICE), and the British National Corpus (BNC). We use a largely corpus-driven method. There are few differences in frequencies of syntactic relations between the corpora, but considerable differences when taking the intricate relations between grammar and lexis into account. We describe differences in the use of zero articles, verb-preposition constructions, and ditransitive verbs. We show that relatively small corpora can be used to discover subtle lexico-grammatical differences.
Keywords: lexico-grammar, syntactic parsing, language variation, Indian English, corpus-driven

1 Introduction

Parsing technology has made considerable advances recently, opening new perspectives for descriptive linguistics. van Noord and Bouma (2009, 37) state that "[k]nowledge-based parsers are now accurate, fast and robust enough to be used to obtain syntactic annotations for very large corpora fully automatically." We apply parsed corpora as a new resource for linguists. Automatically parsed treebanks, also called tree jungles, have been used for e.g. Danish (Bick, 2003) and French (Bick, 2010). The currently available English 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 no treebanks for English regional varieties exist yet. In this situation, automatically parsed corpora can be used as a stopgap to Treebanks.

The detection of regional differences between the various dialects of a language is a major task in synchronic linguistics. We discuss the example of Indian English (IndE), compared to British English (BrE). We use the International Corpus of English (ICE), comparing ICE-India to ICE-GB and partly to the British National Corpus (BNC), when data sparseness problems arise. We use a largely corpus-driven method (Tognini-Bonelli, 2001), paired with manual filtering and linguistic inspection, to detect features of IndE.

The interaction of lexis and grammar has become a linguistic research focus. In computational linguistics, lexicalisation learnt from syntactically annotated corpora has made fast large-scale parsing possible (e.g. Collins (1999)), and in descriptive linguistics, it has given rise to lexicogrammatical and construction grammar theories, for example systemic functional grammar (Halliday, 1994) and collostructions (Stefanowitsch and Gries, 2003). Distinctive phenomena between English varieties typically concentrate at the interface between grammar and lexicon (Schneider, 2004).

	Subject	Object	PP-attachment	clausal
Prec.	92.3% (865/937)	85.3% (353/414)	76.9% (702/913)	74.3% (451/607)
Recall	78.0% (865/1095)	82.5% (353/428)	68.6% (702/1023)	61.7% (451/731)

Table 1: Parser performance on GREVAL test corpus

1.1 Indian English (IndE)

We use IndE as an example variety in this investigation. English is one of the official languages of India. Although there are few native IndE speakers, English is used as *lingua franca* to allow communication between speakers of the many indigenous languages, such as Urdu, Hindhi, Bengali, Marathi, Tamil, and many others. It is therefore an important second or third language for many Indian people, there are over 90 million speakers of IndE. Features of IndE have been described in linguistic research (Gupta and Kapoor, 1991). In the current paper, we are trying to detect regional features in a corpus-driven approach. We do not take previous knowledge as a starting point. The aim is to test a corpus-driven approach as a means of discovering regionalisms.

1.2 Using a syntactic dependency parser

We have used a probabilistic dependency parser, Pro3Gres (Schneider, 2008), which is fast (the BNC parses in one day), close to Tesnière (1959)'s Dependency Grammar conception, and which has been evaluated on several genres and varieties (Haverinen et al., 2008; Lehmann and Schneider, 2009). It is suitable for parsing different varieties of English, as it is robust, so that its output is quite reliable on a number of English varieties (Schneider and Hundt, 2009). For example, it does not enforce subject-verb agreement, it uses statistical preferences instead of strict subcategorisation frames (this entails that e.g. that non-ditransititive verbs can act as ditransitive, a feature that we use in section 3.2, or that prepositional phrases with divergent prepositions get attached, a feature that we need for section 3.3). An evaluation of the performance on subject, object and PP-attachment relations, using the GREVAL gold standard (Carroll, Minnen, and Briscoe, 2003) is given in table 1.

1.3 Corpus data

We used the following corpora for our investigation: in section 3.2, we used the written part of ICE-India and compared it to the written part of ICE-GB. In sections 2 and 3.3 we used the entire ICE-India corpus and compared it to BNC. In section 3.1 we used about two thirds of the written part of ICE-India (the parts which fall into the genres that we investigated) and compared to the same subset of other ICE corpora, namely ICE-GB, ICE-NewZealand and ICE-Fiji.

2 Corpus-Driven Diagnostics

As a first step to discovering variety differences, we measured the total number of occurrences of each syntactic dependency relation. There are, for example, considerable differences between different English genres, so differences between English varieties could be expected. It turns out, however, that differences are typically small. Differences are too subtle to leave a visible impact in frequency counts. In fact, the vast majority of sentences in ICE-India could just as well have been produced by a British or American speaker, there is nothing 'unusual' in them.

The differences are intricate. Schneider (2004) observes that, in regional varieties of English

O/E ratio	Trigram	O(BNC)	O(ICE-India)
1575	this_DT court_NN that_IN	3	21
975	the_DT blood_NN group_NN	3	13
810	do_VBP not_RB recollect_VB	5	18
750	the_DT household_NN sector_NN	3	10
731.25	as_RB to_TO why_WRB	4	13
675	statement_NN before_IN the_DT	4	12
675	state_NN government_NN has_VBZ	3	9
675	is_VBZ known_VBN as_RB	3	9
675	in_IN the_DT hostel_NN	8	24
630	proviso_NN to_TO section_NN	5	14
610.7	the_DT best_JJS feature_NN	7	19
600	were_VBD produced_VBN with_IN	3	8
600	the_DT twentieth_NN of_IN	3	8
600	the_DT election_NN commission_NN	3	8
600	submitted_VBD a_DT memorandum_NN	3	8
562.5	in_IN the_DT in_IN	6	15
534.3	a_DT very_RB very_RB	8	19
525	things_NNS are_VBP there_RB	3	7
525	over_IN medium_JJ heat_NN	6	14
525	not_RB to_TO venture_NN	3	7
506.25	on_IN and_CC so_RB	4	9
506.25	both_CC the_DT parties_NNS	4	9
487.5	the_DT rain_NN water_NN	6	13
487.5	for_IN number_NN of_IN	6	13

Table 2: Trigrams that are at least 480 times more surprising in ICE-India than BNC, according to O/E

> distinctive phenomena tend to concentrate at the interface between grammar and lexicon, concerning structural preferences of certain words (like the complementation patterns that verbs allow), co-occurrence and collocational tendencies of words in phrases, and also patterns of word formation. (Schneider, 2004, 229)

It may thus be revealing to investigate the lexical material that is used in syntactic relations. While there are no semantic class restrictions for most relations, some relations have strict restrictions. A case in point is the relation $obj2$, which is only permitted to occur with ditransitive verbs, and with *elect* verbs. The total number of $obj2$ relations in ICE-India is very similar to ICE-GB, but the distribution of lexical verbal heads differs. For example, there are 12 instances where *provide* is used as a ditransitive verb in ICE-India written, while the only one instance in ICE-GB written is a parsing error. We discuss ditransitive verbs in more detail in section 3.2.

A second case in point are prepositions in prepositional phrases. We compared frequency-ordered lists of prepositions in the *prep* relation, but found no obvious difference. The seven most frequent prepositions appear in the same order in both corpora.

While such lists of heads are short in a strongly restricted class situation such as ditransitive verbs or prepositions, open class lists are unwieldy and difficult to interpret without further statistical processing and filtering. In order to detect lexico-grammatical differences in open class relations, we thus try to approach the corpus from the opposite end, the lexical end, since approaching from the global grammatical end, counting frequencies of grammatical relations, showed very few differences.

Particularly frequent word-sequences, also known as surface collocations, can be detected by using statistical distribution measures such as mutual information, Z-score or Observed/Expected (O/E). We used O/E as it copes relatively well with sparse data and is easy to interpret. We calculated O/E for all ICE-India trigrams and compared them to British English. When using ICE-GB, data sparseness problems are very serious: very many ICE-India trigrams are unseen in ICE-GB. Due to Zipf's law, data sparseness is typically very serious for lexical items in a one million word corpus. In order to alleviate the problem, we used the 100 million word BNC to

(a) Frequent ICE-India trigrams that are absent in the BNC

Trigram	f(ICE-India)
now_RB a_DT days_NNS	42
special_JJ P_NN P_NN	35
canvassed_VBN before_IN this_DT	32
statement_NN was_VBD recorded_VBN	28
learned_VBN special_JJ P_NN	28
is_VBZ called_VBN as_IN	27
scene_NN of_IN offence_NN	26
the_DT honourable_JJ minister_NN	23
for_IN grain_NN yield_NN	22
the_DT learned_VBN special_JJ	21
in_IN the_DT cyclone_NN	19
delay_NN in_IN reply_NN	18
best_JJS feature_NN film_NN	18
avoid_VB delay_NN in_IN	18
small_JJ circle_NN to_TO	17
of_IN solid_JJ wastes_NNS	17
general_JJ body_NN meeting_NN	17
evidence_NN of_IN P_NN	17
feature_NN film_NN in_IN	16
crores_NNS of_IN rupees_NNS	16
in_IN the_DT nodules_NNS	15
has_VBZ also_RB canvassed_VBN	15
sixty-six_NN and_CC half_NN	14

(b) Ditransitive verb counts in ICE-India written and in ICE-GB written

ICE-India	Count	ICE-GB	Count
give	89	give	104
send	37	send	27
provide	12	offer	12
offer	9	tell	10
grant	6	call	8
show	6	do	8
call	4	show	7
develop	4	cost	6
hand	4	pay	6
pay	4	bring	5
bring	3	ask	4
do	3	allow	3
owe	3	earn	3
ask	2	teach	3
consider	2	consider	2
deny	2	deliver	2
earn	2	find	2
find	2	grant	2
promise	2	hand	2
tell	2	promise	2

Table 3: ICE-India trigrams and ditransitive verbs

compare collocations. We calculated an O/E ratio, O/E(ICE-India) divided by O/E(BNC). We then set a threshold T, for example 100, to report trigrams that are T times more surprising in ICE-India than in the BNC. The lists thus obtained are dominated by proper nouns and punctuation marks. After filtering trigrams containing proper nouns and punctuation, we obtain the results shown in table 2 for a threshold $T = 480$.

The majority of the hits arise from text selection criteria, for example there are relatively many legal texts in ICE-India (*proviso to section*, *statement before the*), many medical texts (*the blood group*), and the spoken data percentage is much larger, showing hesitations etc. (*a very very*, *in the in*). But we also see quite formal expressions (*do not recollect*) and, as it turns out when checking the occurrences in the corpus, zero articles (*for number of*), i.e. expressions involving an NP where BrE or American English speakers would expect an article, but IndE speakers often do not use any. We focus on zero articles in section 3.1. An example of the trigram *for number of* is:

(1) And *for number of* years following the Nehruvian outlook this society has built itself. (ICE-India S1b-054)

We also investigated which frequent ICE-India trigrams are absent in the BNC. After filtering proper names and punctuation, the frequency-ranked top of the list is given in table 3 on the left.[1] Besides text selection, Indian features like archaic spellings (*now a days*), formal language (*the honourable minister*), unusual verb complementation with prepositional phrases (*is called as*), and written numbers (*sixty-six and half*) appear in this list.

Examples that show the trigram *is called as* are:

(2) A substance which is helping in chemical reaction *is called as* a reagent. (ICE-India S1b-004)

(3) Thus the intermediate state between crystaline annd isotopic state *is called as* the mesophase or liquid crystals. (ICE-India W1a-020)

[1] A few of these trigrams appear both in the BNC and ICE-India, but the tagger assigned them different tags

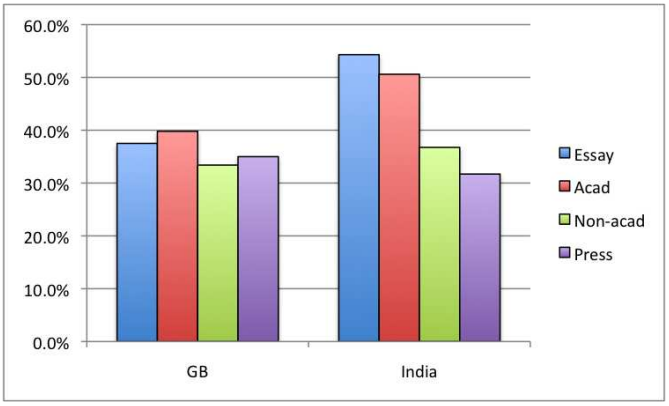

Figure 1: Zero-form article percentages per _NN-tagged chunk head noun (singular common noun) across genre and variety

We investigate verb complementation by prepositional phrases in section 3.3.

Although lists like tables 2 and 3 contain true positives, they contain a high level of garbage, hits that are rare or absent in the BNC due to data sparseness. Larger corpora, and more sophisticated methods are sought for. As for more sophisticated methods, we analyse the parsed material in the following section. Before doing so, let us summarise: The corpus-driven approach with additional manual filtering has uncovered the following potential features of IndE.

- IndE seems to leave out determiners in some situations (e.g. *for number of*). We discuss this in section 3.1.
- Ditransitive verbs have a different distribution in IndE, which we discuss in section 3.2.
- Verb complementation may also involve unusual prepositional phrases (e.g. *is called as*). We discuss this in section 3.3.

3 Analysis

3.1 Zero articles

While the number of articles per noun is only slightly higher in ICE-GB, the number of nouns that have a zero article are considerably higher in ICE-India, as we discuss now.

We have tested a large subset, consisting of two thirds of the written part of the ICE corpora. In ICE-GB, 10,034 of the 27,360 singular common nouns, or 36.7%, have no article. In ICE-India, 12,633 of the 29,032 singular common nouns, or 43.5% have no article. The difference is statistically highly significant (chi-square contingency test, $p < 0.01\%$). In Figure 1 we have broken down zero articles by genre. While the percentage is spread quite homogenously across genres in ICE-GB, ICE-India shows a peak in the least edited genre, student essays, and a tendency towards over-correction in the most edited genre, press.

The need to include zero articles in corpus studies is widely acknowledged in descriptive linguistics: "... no study of article use is truly complete without the discussion of zero articles" (Sand, 2004, 295). Unfortunately, in surface-based approaches it is very difficult to detect zero-forms (e.g. Sedlatschek (2009, 198)).

In a syntactic approach, a zero article form is simply a noun chunk without an article. There are potential complications, however. Quirk et al. (1985, 246) point out that zero articles are only present with nouns that can also be used with a definite article. In e.g. *I like Richard* there is no zero article, but a zero form, as "the zero form is only a label denoting the absence of any article" (Berezowski, 2009, 7). In order to increase the correspondence between zero form

O/E ratio	Head	Prep	f(India)	O/E (India)	O/E(BNC)	manual inspection comment
80.6962	discuss	about	10	148.012	1.83419	You come we will *discuss about* it.
51.3664	study	about	7	67.7127	1.31823	Today we are *studying about* rotation and revolution of the earth.
705.33	advise	into	7	279.731	0.396597	no, consistent parsing error
39.8306	result	into	5	55.3685	1.3901	This *resulted into* a deep sense of growing loneliness
78.7867	burst	of	5	234.214	2.97276	no
53.0517	arrest	from	5	59.374	1.11917	five more terrorists *were arrested from* his home
93.5978	etch	at	3	147.232	1.57303	no
67.2343	withstand	to	2	139.353	2.07265	no
46.6381	significant	on	2	33.1642	0.711096	no
45.8399	nice	on	2	70.0133	1.52734	no
84.4974	line	of	2	120.453	1.42552	no
47.4123	land	into	2	102.124	2.15396	Atul's tendency of worrying too much ... *landed him into* trouble
107.968	exciting	on	2	315.06	2.9181	no
214.685	benefit	out	2	128.156	0.596949	yes: So they*'ll benefit out of the faculty teaching*

Table 4: Candidates for Indian verb-PP constructions, obtained with $\frac{O}{E}\ ratio > 35$ and $f(BNC) < 3$

and zero article, we only measure zero forms of singular common nouns, because few singular nouns, unlike proper names or plural nouns, occur exclusively without article.

3.2 Ditransitive verbs

We mentioned that a frequency-ordered list of ditransitive verb occurrences from the written components of ICE-India and ICE-GB shows considerable differences. The list of all occurrences except for hapax legomena is given in table 3 on the right. Marked differences are in boldprint. An example of *provide* from ICE-India is:

(4) I am enclosing herewith a detailed resume of my professional career and feel that I can *provide you the best possible services* in the areas required. (ICE-India W1b-024)

Grant occurs twice in ICE-GB written and six times in ICE-India written, all syntactic analyses are correct. *Hand* occurs twice in ICE-GB written and four times in ICE-India, all syntactic analyses are correct. These differences may thus arise from a text selection coincidence just as well as represent an Indian feature. All instances of *develop* are parser errors.

Differences in ditransitive verbs, particularly *provide*, are confirmed in the corpus linguistics literature, for example Mukherjee and Hoffmann (2006). They list 5 new ditransitive verbs that occur in ICE-India, but only *provide* occurs more than 4 times in the one-million word corpus (we only used the written component, i.e. 400,000 words). Mukherjee (2009, 125) writes that "as most of the new ditransitives are relatively rare, only few of them can be detected in the 1-million-word ICE-India corpus".

Verb complementation is often described as particularly important for linguistic variation: "Verb complementation is an all-pervading structural feature of language and thus likely to be more significant in giving a variety its character than, for example, lexis."(Olavarrìa de Ersson and Shaw, 2003, p. 118)

3.3 Verb-preposition constructions

For this investigation, we leave the distinction between preposition and verbal particle underspecified. All verb-PP constructions are included, irrespective of whether they are complements or adjuncts. To retrieve unusual verb-preposition combinations, we use the O/E measure. O/E is a probabilistic measure of surprise, it tends to give particularly high scores to rare events, and it works well on rare collocations. We used the BNC instead of ICE-GB because of sparse data problems, which can partly be alleviated by using a large comparison base. The O/E ratio that

we use expresses how much more surprising a collocation is in ICE-India than in the BNC. It is calculated as follows:

$$O/E\ ratio = \frac{O/E(India)}{O/E(BNC)} = \frac{\frac{O(India)}{E(India)}}{\frac{O(BNC)}{E(BNC)}} = \frac{\frac{O_{India}(R,w_1,w_2)\cdot N_{India}}{O_{India}(R,w_1)\cdot O_{India}(R,w_2)}}{\frac{O_{BNC}(R,w_1,w_2)\cdot N_{BNC}}{O_{BNC}(R,w_1)\cdot O_{BNC}(R,w_2)}} \quad (1)$$

where N is corpus size, R is the relation ($pobj$), w_1 the head (verb), w_2 the preposition or verbal particle. We then apply variable thresholds to generate candidates for specifically Indian verb-PP constructions. For $O/E\ ratio > 35$ and $f(BNC) < 3$ we get the candidates shown in table 4. In the last column, we give a comment, assessing whether the candidate is a true positive, based on manual inspection of all occurrences.

Using lower thresholds leads to lower precision, but more instances are recalled, e.g:

(5) So he was using the stones and *preparing instruments out* of it (ICE-India S1b-008)

(6) And he has *described all about* that. (ICE-India S1a-092)

(7) Then from government aided school I *switched over* to government school. (ICE-India S1a-024)

(8) You had the guts of your blighted mother to *complain against* us to the Governor. (ICE-India W2f-018)

(9) ... he tried to enlighten the people and be *aware towards* all these irregularities and if possible try to remove them. (ICE-India S1a-007)

(10) Wings are *absent to* apterygotes. (ICE-India W1a-019)

Counts are very low, too low for reaching statistical significance. Although a one-million word corpus is very small for lexical research, particularly for lexical interaction research, valid insights can be obtained, the amount of manual filtering required is easily manageable.

Our findings are confirmed in the previous literature but also list new pairs. Differences in verb-preposition and verbal particle use in IndE are described in Sedlatschek (2009), Mukherjee (2009), Nesselhauf (2009). The former two authors hypothesize on the reasons for the differences; for example analogy to existing, semantically related particle verbs (e.g. in 10) or noun-verb conversion (e.g. in 8). Concerning articles, many Indian substrate languages do not have articles which makes it difficult for language learners to acquire the concept. However, while such explanations sound reasonable, they are empirically almost impossible to prove.

4 Conclusions and Outlook

We have demonstrated the benefits of using NLP techniques to help descriptive linguistic studies. In particular, we have shown that automatically parsed corpora can be used to detect regional English variety features and subtle lexico-grammatical differences using a largely corpus-driven method. As the data inspection phase involves analyzing, commenting and subcategorizing instances, the overhead which manual filtering creates is a manageable disadvantage compared to a fully automatic approach. We are not aware of any fully automatized approach to this task. The features that we found are all confirmed in the descriptive linguistic literature. Concerning research on zero articles, only a syntactic approach offers the appropriate tools to measure zero article frequency.

We have shown that with small corpora (1 million words or even less) many regional features can be discovered. We have conducted similar investigations on other corpora of the ICE family, and we have investigated additional features, such as differences in tense, aspect and modality. We will use our method to discover regional features from large web-collected corpora.

References

Berezowski, Leszek. 2009. *The Myth of the Zero Article*. Continuum, London.

Bick, Eckhard. 2003. A CG & PSG hybrid approach to automatic corpus annotation. In Kiril Simow and Petya Osenova, editors, *Proceedings of SProLaC2003*, pages 1–12, Lancaster.

Bick, Eckhard. 2010. FrAG, a hybrid constraint grammar parser for French. In *Proceedings of LREC 2010*, Valletta, Malta.

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*. Kluwer, Dordrecht, pages 299–316.

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

Gupta, R.S. and Kapil Kapoor. 1991. *English in India: Issues and Problems*. Academic Foundation, Delhi.

Halliday, M.A.K. 1994. *An Introduction to Functional Grammar, 2nd ed.* Arnold, London.

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.

Lehmann, Hans Martin and Gerold Schneider. 2009. Parser-based analysis of syntax-lexis interaction. In Andreas H. Jucker, Daniel Schreier, and Marianne Hundt, editors, *Corpora: Pragmatics and discourse: papers from the 29th International conference on English language research on computerized corpora (ICAME 29)*, Language and computers 68. Rodopi, Amsterdam/Atlanta, pages 477–502.

Mukherjee, Joybrato. 2009. The lexicogrammar of present-day Indian English. Corpus-based perspectives on structural nativisation. In Ute Römer and Rainer Schulze, editors, *Exploring the Lexis-Grammar Interface*. John Benjamins, Amsterdam, pages 117–135.

Mukherjee, Joybrato and Sebastian Hoffmann. 2006. Describing verb-complementational profiles of New Englishes: A pilot study of Indian English. *English World-Wide*, 27(2):147–173.

Nesselhauf, Nadja. 2009. Co-selection phenomena across New Englishes. Parallels (and differences) to foreign learner varieties. *English World-Wide*, 30(1):1–26.

Olavarrìa de Ersson, Eugenia and Philip Shaw. 2003. Verb complementation patterns in Indian Standard English. *English World-Wide: A Journal of Varieties of English*, 24(2):137–161.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language. 11th edn.* Longman, London.

Sand, Andrea. 2004. Shared morpho-syntactic features in contact varieties of English: Article use. *World Englishes*, 23:281–98.

Schneider, Edgar. 2004. How to trace structural nativization: Particle verbs in World Englishes. *World Englishes*, 23:2:227–249.

Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Schneider, Gerold and Marianne Hundt. 2009. Using a parser as a heuristic tool for the description of New Englishes. In *Proceedings of Corpus Linguistics 2009*, Liverpool.

Sedlatschek, Andreas. 2009. *Contemporary Indian English: variation and change*. Varieties of English around the world. John Benjamins, Amsterdam / Philadelphia.

Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, pages 209–43.

Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Librairie Klincksieck, Paris.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. John Benjams, Amsterdam.

van Noord, Gertjan and Gosse Bouma. 2009. Parsed corpora for linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 33–39, Athens, Greece. Association for Computational Linguistics.