Studying the Distribution of Fragments of English Using Deep Semantic Annotation

Camilo Thorne
KRDB Research Centre for Knowledge and Data
3, Piazza Domenicani, 39100 (Italy)
thorne@inf.unibz.it

Abstract

We present a preliminary study on how to use deep semantic annotation, namely the Boxer statistical semantic parser, that is capable of producing FO semantic representations for English sentences, to understand the distribution of families of so-called fragments of English. In particular, we try to answer the questions, relevant for the field of natural logic, of whether and how the semantic complexity of those fragments (viz., the computational complexity of the satisfiability problem of their FO semantic representations) correlates with their frequency.

1 Introduction

Natural logic (Moss, 2010; MacCartney and Manning, 2007; van Eijck, 2005) is a relatively recent area of cognitive science, logic and computational linguistics which has as its main goal to understand which logical formalisms best model common-sense deductive reasoning as "embedded" in spoken and written language.

More recently (Muskens, 2010; Szymanik, 2009), interest has arisen regarding the relationship of such formalisms to, on the one hand, *formal semantics*, the Montagovian HO and FO modelling of natural language semantics and compositionality via logic *meaning representations* (MRs) and, on the other hand, *semantic complexity*, the computational complexity of satisfiability for such MRs. This with two goals in mind: *(i)* Measuring the complexity of natural reasonings. *(ii)* Inferring correlations between complexity and frequency (viz., how often the formal models occur in natural language data) and/or accuracy (viz., what proportion of such formal reasonings are correctly inferred by speakers).

This study purports to contribute to this debate by considering the following two approaches: *(i)* Generating FO MRs from natural language text via *semantic annotation* in the form of deep (compositional and Montagovian-based) semantic parsing. *(ii)* Focusing on so-called *fragments of English*, viz., controlled subsets of English wherein ambiguity has been removed, semantics is compositional and deterministic and that give rise, modulo formal semantics, to fragments of FO (Pratt-Hartmann and Third, 2006).

By studying the semantic complexity of the fragments, via computational complexity analysis and their approximate distribution in corpora, via semantic annotation (compositional semantic parsing), we can, we believe, understand better how complexity correlates with use. For instance, complex, recursive, syntactic structures (e.g., center embedding) are less frequent in English than simpler, non-recursive structures. To see if this also holds for semantic complexity, we try to test the following hypothesis:

Semantic complexity is inversely proportional to frequency. (H)

2 Semantic Complexity and The Fragments of English

A (controlled) fragment of English (Pratt-Hartmann and Third, 2006) is a linguistically salient, ambiguity free subset of English constructed using context-free semantically enriched grammars which generate and recognize, alongside the grammatical utterances of the fragment, their logical (HO and FO) MRs, modulo Montagovian compositional translations $\tau(\cdot)$ (defined via semantic actions attached to the grammar rules). Figure 1 recalls the definition of the base fragment, whose coverage is subsequently expanded to larger subsets of English.

A *positive* fragment is any such fragment *without negation*, and a *negative* fragment is a fragment *with negation*. Each fragment of English

Fragment	Coverage	FO Operators and Relations
COP($\neg$)	Copula ("is a"), nouns ("man"), intransitive verbs ("runs"), "every", "some" names ("Joe"), adjectives ("thin") (+"not"))	$\{\forall, \exists, (\neg)\}$ $\cup$ $\{P_i^1 \mid i \in \mathbb{N}\}$
COP($\neg$)+TV	COP($\neg$) +transitive verbs ("loves")	$\{\forall, \exists, (\neg)\}$ $\cup \{P_i^1, P_j^2 \mid i, j \in \mathbb{N}\}$
COP($\neg$)+DTV	COP($\neg$) +ditransitive verbs ("gives")	$\{\forall, \exists, (\neg)\}$ $\cup \{P_i^1, P_j^3 \mid i, j \in \mathbb{N}\}$
COP($\neg$)+TV+DTV	COP($\neg$)+TV + ditransitive verbs	$\{\forall, \exists, (\neg)\}$ $\cup \{P_i^1, P_j^2, P_k^3 \mid i, j, k \in \mathbb{N}\}$
COP($\neg$)+Rel	COP($\neg$)+relative pronouns ("who", "that", "which") "and", intersective adjectives (+"or")	$\{\forall, \exists, \wedge, (\neg, \vee)\}$ $\cup$ $\{P_i^1 \mid i \in \mathbb{N}\}$
COP($\neg$)+Rel+TV	COP($\neg$)+Rel +transitive verbs	$\{\forall, \exists, \wedge, (\neg, \vee)\}$ $\cup \{P_i^1, P_j^2 \mid i, j \in \mathbb{N}\}$
COP($\neg$)+Rel+DTV	COP($\neg$)+Rel +ditransitive verbs	$\{\forall, \exists, \wedge, (\neg, \vee)\}$ $\cup \{P_i^1, P_j^3 \mid i, j \in \mathbb{N}\}$
COP($\neg$)+Rel+TV+DTV	COP($\neg$)+Rel+TV +ditransitive verbs	$\{\forall, \exists, \wedge, (\neg, \vee)\}$ $\cup \{P_i^1, P_j^2, P_k^3 \mid i, j, k \in \mathbb{N}\}$

Table 1: The ("positive" and "negative") fragments of English. See (Pratt-Hartmann and Third, 2006; Thorne, 2010) for more detailed definitions. Please note that we have modified slightly the former notation of the fragments, for readability reasons.

gives rise to *(i)* a distinct combination of FO operators (i.e., $\forall, \exists, \vee\neg$ and $\wedge$) *(ii)* a distinct combination of unary, binary and ternary relation symbols (together with individual constants). See Table 1. More in general, it generates a unique FO fragment, whose computational complexity for satisfiability constitutes the semantic complexity of the fragment (Pratt-Hartmann and Third, 2006), which can be studied in general, viz., *combined complexity*, or relatively to the number of constants occurring in the MRs, viz., *data complexity* (Thorne, 2010).

The fragment's content lexicon (nouns, common nouns, verbs, adjectives, names) will thus convey the signature (constants and relations) of the engendered FO fragment, while the function lexicon will convey the logical operators. Semantic complexity will be, in general, correlated to the fragment's function lexicon. Two big classes of fragments can be observed:

- "Non-Boolean-closed" fragments: are fragments that cannot express Boolean functions, viz., the positive fragments, together with COP$^\neg$, COP$^\neg$+TV, COP$^\neg$+DTV and COP$^\neg$+TV+DTV.
- "Boolean-closed" fragments: fragments expressive enough to encode Boolean satisfiability, viz., COP$^\neg$+Rel, COP$^\neg$+Rel+TV, COP$^\neg$+Rel+TV and COP$^\neg$+Rel+TV+DTV.

Table 2 summarizes the computational properties (data and combined) that arise for the fragments (for the proofs, we refer the reader to (Pratt-Hartmann and Third, 2006) and (Thorne, 2010). As the reader can see, "Boolean-closedness" gives rise, in general, to an exponential blowup in complexity. Fragments that are "non-Boolean-closed", such as the positive fragments and the fragments without relatives and transitive verbs, have *tractable* (at most PTIME) combined or data complexity, whereas "Boolean-closed" fragments have *intractable* (at least NPTIME-hard) combined or data complexity.

3 Empirical Analysis

3.1 Corpus Analysis

In this section we summarize our analysis regarding the co-occurence of negations, conjunctions, disjunctions, and universal and existential quantification in English question and sentence corpora via semantic annotation.

More precisely, we consider the frequency of sentences expressing, modulo formal semantics, *positive* (not containing $\neg$) and *negative* (contain-

Phrase Structure Rules	
$\mathbf{S} \rightarrow \mathbf{NP\ VP}$	$\tau(\mathbf{S}) = \tau(\mathbf{NP})(\tau(\mathbf{VP}))$
$\mathbf{VP} \rightarrow$ is a $\mathbf{N}$	$\tau(\mathbf{VP}) = \tau(\mathbf{N})$
$\mathbf{VP} \rightarrow$ is $\mathbf{Adj}$	$\tau(\mathbf{VP}) = \tau(\mathbf{Adj})$
$\mathbf{VP} \rightarrow \mathbf{IV}$	$\tau(\mathbf{VP}) = \tau(\mathbf{IV})$
$\mathbf{NP} \rightarrow \mathbf{Pn}$	$\tau(\mathbf{NP}) = \tau(\mathbf{Pn})$
$\mathbf{NP} \rightarrow \mathbf{Det\ N}$	$\tau(\mathbf{NP}) = \tau(\mathbf{Det})(\tau(\mathbf{N}))$
$(\mathbf{VP} \rightarrow$ is $\mathbf{Ng}$ a $\mathbf{N}$	$\tau(\mathbf{VP}) = \tau(\mathbf{Ng})(\tau(\mathbf{N})))$
$(\mathbf{VP} \rightarrow$ does $\mathbf{Ng\ IV}$	$\tau(\mathbf{VP}) = \tau(\mathbf{Ng})(\tau(\mathbf{IV})))$

Function Lexicon	
$\mathbf{Det} \rightarrow$ every	$\tau(\mathbf{Det}) = \lambda P.Q.\forall x(P(x) \rightarrow Q(x))$
$\mathbf{Det} \rightarrow$ some	$\tau(\mathbf{Det}) = \lambda P.Q.\exists x(P(x) \wedge Q(x))$
$(\mathbf{Ng} \rightarrow$ not	$\tau(\mathbf{Ng}) = \lambda P.\lambda x.\neg P(x))$

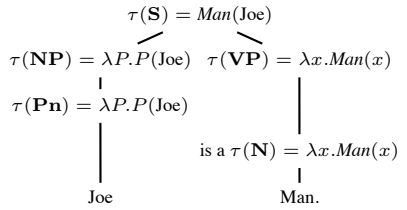

Figure 1: Top: COP(¬). Bottom: COP(¬) parse tree for "Joe is a man.". We omit the content lexicon. Notice how to each grammar rule a semantic action is attached, defining $\tau(\cdot)$.

ing ¬) classes $c \subseteq \{\forall, \exists, \neg, \wedge, \vee\}$ of FO operators. Each such class approximates MRs belonging, modulo logical equivalence, to a distinct fragment of FO and expressible by a distinct fragment of English. For instance the class $\{\forall, \exists, \wedge, \vee\}$ identifies MRs from the positive fragment of FO. But it also identifies MRs belonging to English fragments such as, e.g., COP(+Rel)+TV+DTV. Specifically, after semantically annotating the corpora we observed the frequency of

- 4 "Boolean-closed" classes viz.: $\{\exists, \wedge, \neg\}$, $\{\exists, \wedge, \neg, \forall\}$, $\{\exists, \wedge, \neg, \forall, \vee\}$ and $\{\neg, \forall\}$, and of
- 4 "non-Boolean-closed" classes viz.: $\{\exists, \wedge\}$, $\{\exists, \wedge, \forall\}$, $\{\exists, \wedge, \vee\}$ and $\{\exists, \wedge, \forall, \vee\}$,

where by "Boolean-closed" and "non-Boolean-closed", we mean, by abuse, classes, resp., expressive or not expressive enough to encode Boolean satisfiability.

To obtain a representative sample, we considered corpora of multiple domains and with sentences of arbitrary type (declarative and interrogative). We considered: *(i)* a subset (A: press articles) of the Brown corpus[1]; *(ii)* a subset (Geoquery880) of the Geoquery corpus[2]; *(iii)* a corpus of clinical questions[3]; and *(iv)* a sample from the TREC 2008 corpus[4]. Table 3 summarizes their main features.

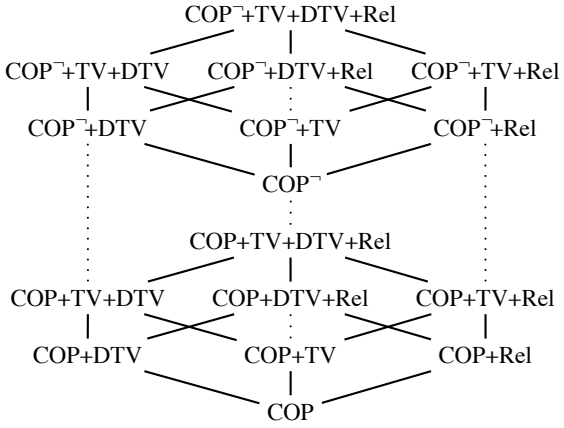

Figure 2: Relative expressive power of the fragments.

We used two methods, that we describe below. The left column of Figure 3 provides plots of the statistics collected with both methods. They also plot the mean class frequencies across the corpora, and the mean cumulative frequency.

Semantic Parsing with Boxer. We exploited the availability of wide-coverage (statistical) deep semantic parsers and annotators. In particular, the Boxer and Clark & Curran tools (Bos, 2008), based on combinatorial categorial grammar and discourse representation theory (DRT), that output first-order MRs. The pipeline of this system consists in the following three basic steps: *(i)* each part of speech in a sentence is annotated with its most likely (categorial grammar) syntactic category; *(ii)* the most likely of the resulting possible combinatorial categorial grammar derivations (or proofs) is computed and returned; and *(iii)* a neo-Davidsonian semantically weakened[5] FO meaning representation is computed using DRT.

For instance, when parsing Wh-questions from

[1] http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml

[2] http://www.cs.utexas.edu/users/ml/nldata/geoquery.html

[3] http://clinques.nlm.nih.gov

[4] http://trec.nist.gov

[5] In this setting, the semantics of verbs is represented in terms of events connected via thematic roles to verb arguments (agents, themes, etc.). In addition, the semantics of non-FO constructs such as "most" is weakened to some FO representation.

	Combined	Data
COP	LSPACE	LSPACE
COP+TV	PTIME	LSPACE
COP+DTV	PTIME	LSPACE
COP+TV+DTV	PTIME	LSPACE
COP+Rel	PTIME-c	LSPACE
COP+Rel+TV	PTIME-c	PTIME
COP+Rel+DTV	PTIME-c	PTIME
COP+Rel+DTV+TV	PTIME-c	PTIME

	Combined	Data
$\text{COP}^{\neg}$	LSPACE	LSPACE
$\text{COP}^{\neg}$+TV	NLSPACE-c	LSPACE
$\text{COP}^{\neg}$+DTV	PTIME	LSPACE
$\text{COP}^{\neg}$+TV+DTV	PTIME	LSPACE
$\text{COP}^{\neg}$+Rel	NPTIME-c	LSPACE
$\text{COP}^{\neg}$+Rel+TV	EXPTIME-c	NPTIME-c
$\text{COP}^{\neg}$+Rel+DTV	NEXPTIME-c	NPTIME-c
$\text{COP}^{\neg}$+Rel+DTV+TV	NEXPTIME-c	NPTIME-c

Table 2: Semantic complexity of the fragments of English, positive and otherwise (Pratt-Hartmann and Third, 2006; Thorne, 2010).

Corpus	Size	Domain	Type
Brown	19,741 sent.	Open (news)	Decl.
Geoquery	364 ques.	Geographical	Int.
Clinical ques.	12,189 ques.	Clinical	Int.
TREC 2008	436 ques.	Open	Int.

Table 3: Corpora used in this study.

the TREC 2008 corpus such as "What is one common element of major religions?", Boxer outputs

$$\exists y \exists z \exists e \exists u(\texttt{card}(y,u) \wedge \texttt{c1num}(u) \wedge \texttt{nnumeral1}(u) \wedge \texttt{acommon1}(y) \wedge \texttt{nelement1}(y) \wedge \texttt{amajor1}(z) \wedge \texttt{nreligions1}(z) \wedge \texttt{nevent1}(e) \wedge \texttt{rof1}(y,z))$$

where $\wedge$ and $\exists$ co-occur, but not $\vee$, $\neg$, or $\rightarrow$.

After semantically annotating each corpus with Boxer, we checked for each MR produced, whether it belongs to a "Boolean-closed" or a "non-Boolean-closed" class.

Pattern-based. Boxer is considered to have a reasonably good performance (covering over 95% of English, with approx. 75% accuracy), when parsing and annotating declarative sentences and corpora, but not necessarily so over interrogative sentences and corpora. It also commits us to a neo-Davidsonian semantics, whose event-based verb semantics biases interpretation towards positive existential MRs, making its output somewhat noisy.

To understand how useful Boxer (or similar deep semantic annotators) can be to discover statistical trends of the kind stated in our hypothesis (H), we decided compare to its results to those that one may obtain using a simple methodology based on patterns. Indeed, modulo formal semantics, English function words convey or express FO operators. As such we considered the following patterns, *(i)* for $\neg$: "not", "no" *(ii)* for $\exists$: "some", "a" *(iii)* for $\forall$: "all", "every, "each" *(iv)* for $\wedge$: "who", "what", "which", "and" *(v)* for $\vee$: "or", and their combinations/co-occurences *within sentences* to approximate the"Boolean-" and "non-Boolean-closed" classes that interest us.

3.2 Basic statistical tests

The mean (cumulative) frequency plots obtained in Figure 3 show a distribution where class frequency is skewed towards positive existential classes: $\{\exists, \wedge\}$, $\{\exists, \forall, \wedge\}$ and positive existential $\{\exists, \forall, \wedge, \vee\}$ MRs occur quite frequently, whereas the opposite holds for negation (low frequency overall). The question is whether this substantiates our hypothesis (H). We ran some basic statistical tests to understand how random or significant this phenomenon is. Table 4 summarizes the test results, which we explain below.

Power Law Behavior. A *power law distribution* is a kind of exponential, non-normal and skewed distribution where the topmost (i.e., most frequent) 20% outcomes of a variable concentrate 80% of the probability mass.

Power law distributions are widespread in natural language data (Baroni, 2009; Newman, 2005). It makes sense to understand whether the relationship stated by (H) can be restated and described as a power law relation between *class frequency* $fr(c)$ and *class rank* $rk(c)$, viz.,

$$fr(c) = \frac{a}{rk(c)^b}. \quad (1)$$

To approximate the parameters a and b it is customary to run a least squares linear regression, since (1) is equivalent to a linear model on the log-log scale:

$$\log_{10}(fr(c)) = \log_{10}(a) - b \cdot \log_{10}(rk(c)). \quad (2)$$

	$H(C)$	$H_{rel}(C)$	Skewness	χ^2	p-value	df.	Power law	R^2
Boxer	1.53	0.51	1.93	293731473.0	0.0	7	$fr(c) = \frac{47.86}{rk(c)^{1.94}}$	0.92
Patterns	1.50	0.50	1.21	906727332.0	0.0	7	$fr(c) = \frac{5128.61}{rk(c)^{4.09}}$	0.73

Table 4: Summary of test results.

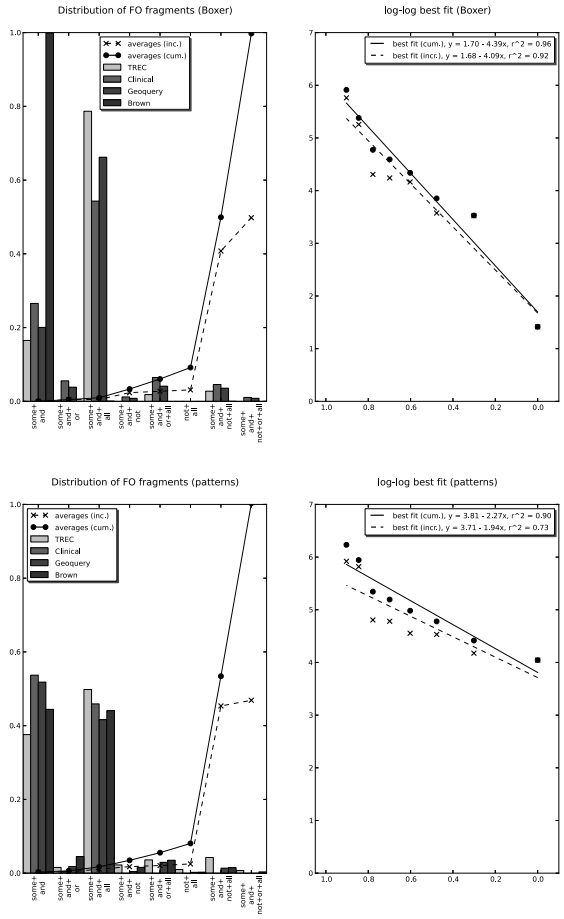

Figure 3: Fragment distribution with Boxer and the pattern-based method, together with their log-log regression plot. We plot class cumulative and mean frequencies (the latter in increasing order).

Entropy, Skewness and χ^2 Tests. Following mainly (Gries, 2010), we conducted the following tests. We computed class *entropy* $H(C)$, where C denotes $\{\forall, \exists, \vee, \wedge, \neg\}$. This number tries to measure the degree of randomness of a distribution:

$$H(C) = -\sum_{c \in C} fr(c) \cdot \log_2(fr(c)). \qquad (3)$$

A low number indicates a low degree of randomness. Entropy can be complemented with its *relative entropy* $H_{rel}(C)$:

$$H_{rel}(C) = \frac{H(C)}{\log_2(\#(C))}. \qquad (4)$$

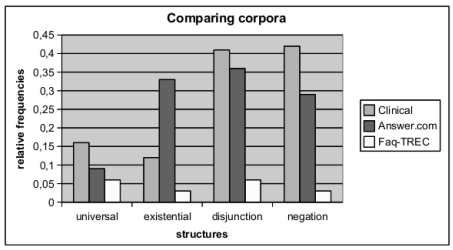

Figure 4: Relative frequency of FO operators in question corpora (Bernardi et al., 2007).

In addition to measuring $H(C)$ and $H_{rel}(C)$, we also run a χ^2 test (since distributions in natural language data are often non-parametric) and measured the overall skewness of the distribution.

3.3 Discussion and Related Work

Table 4 shows that the distributions observed under both methods possess a relatively low entropy (relatively to a peak entropy of 3.0), and thus appear not to be so random. The χ^2 statistic, moreover, entails that such distributions differ from uniform or random distributions (the null hypothesis rejected by the test), since $p < 0.01$. They also show a high measure of skewness. Lest, but not least, the cumulative and non-cumulative distributions seem to follow, to some extent a power-law model. The reader will find on the second (right) column of Figure 3 the plots of the log-log regressions, which show a strong positive correlation (the R^2 index), stronger for Boxer (0.92) than for the patterns (0.73)[6].

This analysis can be compared to the more linguistics-based methodology followed in (Bernardi et al., 2007), in which was analyzed the distribution, in (solely) interrogative corpora, of classes of logical word patterns (but not of their co-occurence), e.g., "all", "both", "each", "every", "everybody", "everyone", "any", "none", "nothing". See Figure 4.

This may suggest that, while users use negation or disjunction words as frequently as conjunction

[6]Quite strong for both cumulative distributions: 0.96 and 0.90 resp., in the plot.

and existential words, and all these more than universal words, when combining them *within* sentences, "non-Boolean-closed" combinations are preferred.

Moreover (Szymanik, 2009) reports results that may seem to imply that natural reasoning *accuracy* (and not only their distribution in corpora) may be inversely correlated to expressiveness and semantic complexity, where by accuracy is meant the ability of speakers to correctly infer logical consequences from texts describing logical arguments (in the experiments, arguments regarding FO and HO generalized quantifiers). Users make more mistakes when the underlying logic (or logical MRs) are NPTIME-hard than when they are PTIME, and take more time to understand and infer such consequences.

This said, the corpora considered in our study were small, and the two methods (Boxer and the patterns) inaccurate (indeed, the pattern-based method remains quite simple). The results reported here, while encouraging, cannot be regarded yet as fully conclusive for (H).

4 Conclusions

We have presented a preliminary study on how to apply deep semantic annotation techniques to understand the distribution of specific fragments of English in English corpora, and specifically to understand if is is possible to infer relationships between their distribution and their semantic complexity (i.e., the computational complexity of their logic MRs).

We have experimented with a methodology based on the Boxer semantic parser, and applied some basic statistical tests on the distribution obtained that may seem to indicate that "non-Boolean-closed" (tractable) fragments might occur more often than "Boolean-closed" (intractable) fragments, although the results obtained thus far remain still inconclusive.

To counter these shortcomings we would like in the future to *(i)* run the experiment with other deep semantic annotation methods and parsers (such as, e.g., those based on minimal recursion semantics (Copestake, 2007)), *(ii)* consider larger corpora, in particular declarative corpora (over which the performance of Boxer is higher) *(iii)* consider more involved statistical tests, to try to understand how the fragments are distributed. We believe however that the methodology proposed is interesting and promising, and all the more due to the current advances in semantic annotation, which may yield better results once points *(i)–(iii)* are addressed.

References

Marco Baroni. 2009. Distributions in text. In Anke Lüdeling and Merja Kytö, editors, *Corpus linguistics: An International Handbook*, volume 2, pages 803–821.

Raffaella Bernardi, Francesca Bonin, Domenico Carbotta, Diego Calvanese, and Camilo Thorne. 2007. English querying over ontologies: E-QuOnto. In *Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence (AI*IA 2007)*.

Johan Bos. 2008. Wide-coverage semantic analysis with Boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing (STEP 2008)*.

Ann Copestake. 2007. Semantic composition with (robust) minimal recursion semantics. In *Proceedings of the ACL-07 workshop on Deep Linguistic Processing*.

Stefan Th. Gries. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez and Moisés Almela, editors, *A mosaic of corpus linguistics: selected approaches*, pages 269–291. Peter Lang.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing (RTE 2007)*.

Lawrence S. Moss. 2010. Natural logic and semantics. In *Proceedings of the 2009 Amsterdam Colloquium (AC 2009)*.

Reinhard Muskens. 2010. An analytic tableau system for natural logic. In *Proceedings of the 2009 Amsterdam Colloquium (AC 2009)*.

M. E. J. Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

Ian Pratt-Hartmann and Allan Third. 2006. More fragments of language. *Notre Dame Journal of Formal Logic*, 47(2):151–177.

Jakub Szymanik. 2009. *Quantifiers in Time and Space*. Institute for Logic, Language and Computation.

Camilo Thorne. 2010. *Query Answering over Ontologies Using Controlled Natural Languages*. KRDB Centre for Knowldge and Data.

Jan van Eijck. 2005. Natural logic for natural language. In *Proceedings of the 6th International Tbilisi Symposium on Logic, Language, and Computation TbiLLC 2005*.