An Empirical Approach to the Interpretation of Superlatives

Johan Bos
Laboratory for Computational Linguistics
Department of Computer Science
University of Rome "La Sapienza"
bos@di.uniroma1.it

Malvina Nissim
Laboratory for Applied Ontology
Institute for Cognitive Science and Technology
National Research Council (CNR), Rome
malvina.nissim@loa-cnr.it

Abstract

In this paper we introduce an empirical approach to the semantic interpretation of superlative adjectives. We present a corpus annotated for superlatives and propose an interpretation algorithm that uses a wide-coverage parser and produces semantic representations. We achieve F-scores between 0.84 and 0.91 for detecting attributive superlatives and an accuracy in the range of 0.69–0.84 for determining the correct comparison set. As far as we are aware, this is the first automated approach to superlatives for open-domain texts and questions.

1 Introduction

Although superlative noun phrases (*the nation's largest milk producer, the most complex arms-control talks ever attempted*, etc.) received considerable attention in formal linguistics (Szabolcsi, 1986; Gawron, 1995; Heim, 1999; Farkas and Kiss, 2000), this interest is not mirrored in computational linguistics and NLP. On the one hand, this seems remarkable, since superlatives are fairly frequently found in natural language. On the other hand, this is probably not that surprising, given that their semantic complexity requires deep linguistic analysis that most wide-coverage NLP systems do not provide.

But even if NLP systems incorporated linguistic insights for the automatic processing of superlatives, it might not be of help: the formal semantics literature on superlatives focuses on linguistically challenging examples (many of them artificially constructed) which might however rarely occur in real data and would therefore have little impact on the performance of NLP systems. Indeed, no corpus-based studies have been conducted to get a comprehensive picture of the variety of configurations superlatives exhibit, and their distribution in real occurring data.

In this paper we describe our work on the analysis of superlative adjectives, which is empirically grounded and is implemented into an existing wide-coverage text understanding system. To get an overview of the behaviour of superlatives in text, we annotated newswire data, as well as queries obtained from search engines logs. On the basis of this corpus study, we propose, implement and evaluate a syntactic and semantic analysis for superlatives. To the best of our knowledge, this is the first automated approach to the interpretation of superlatives for open-domain texts that is grounded on actual corpus-evidence and thoroughly evaluated. Some obvious applications that would benefit from this work are question answering, recognition of entailment, and more generally relation extraction systems.

2 Syntax and Semantics of Superlatives

2.1 Surface Forms

In English, superlative adjectives appear in a large variety of syntactic and morphological forms. One-syllable adjectives and some two-syllable adjectives are directly inflected with the suffix "-est". Some words of two syllables and all words of three or more syllables are instead introduced by "most" (or "least"). Superlatives can be modified by ordinals, cardinals or adverbs, such as intensifiers or modals, and are normally preceeded by the definite article or a possessive. The examples below illustrate the wide variety and uses of superlative adjectives.

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 9–17,
Sydney, July 2006.

the **tallest** woman

AS Roma's **quickest** player

the Big Board's **most respected** floor traders

France's **third-largest** chemical group

the **most-recent** wave of friendly takeovers

the **two largest** competitors

the **the southern-most** tip of England

its **lowest possible** prices

Superlative adjectives can manifest themselves in predicative ("Mia is the tallest.") or attributive form ("the tallest woman"). Furthermore, there are superlative adverbs, such as "most recently", and idiomatic usages.

2.2 The Comparison Set

It is well known that superlatives can be analysed in terms of comparative constructions (Szabolcsi, 1986; Alshawi, 1992; Gawron, 1995; Heim, 1999; Farkas and Kiss, 2000). Accordingly, "the oldest character" can be interpreted as the character such that there is no older character, in the given context. Therefore, a correct semantic interpretation of the superlative depends on the correct characterisation of the *comparison set*. The comparison set denotes the set of entities that are compared to each other with respect to a certain dimension (see Section 2.3). In "the oldest character in the book", the members of the comparison set are characters in the book, and the dimension of comparison is age.

The computation of the comparison set is complicated by complex syntactic structure involving the superlative. The presence of possessives for example, as in "AS Roma's quickest player", extends the comparison set to players of AS Roma. Prepositional phrases (PPs), gerunds, and relative clauses introduce additional complexity. PPs that are attached to the head noun of the superlative are part of the comparison set — those that modify the entire NP are not. Similarly, restrictive relative clause are included in the comparison set, non-restrictive aren't.

We illustrate this complexity in the following examples, taken from the Wall Street Journal, where the comparison set is underlined:

The **oldest** designer got to work on the dashboard, she recalls. (WSJ02)

A spokesman for Borden Inc., the nation's **largest** milk producer, concedes Goya may be on to something. (WSJ02)

Right now, the **largest** loan the FHA can insure in high-cost housing markets is $101,250. (WSJ03)

With newspapers being the **largest** single component of solid waste in our landfills ... (WSJ02)

... questions being raised by what generally are considered the **most complex** arms-control talks ever attempted. (WSJ02)

Besides syntactic ambiguities, the determination of the comparison set can be further complicated by semantic ambiguities. Some occurrences of superlatives licence a so-called "comparitive" reading, as in the following example discussed in the formal semantics literature (Heim, 1999; Szabolcsi, 1986):

John climbed the **highest** mountain.

Here, in the standard interpretion, the mountain referred to is the highest available in the context. However, another interpretation might arise in a situation where several people climbed several mountains, and John climbed a mountain higher than anyone else did, but not necessarily the highest of all mountains in the context. Our corpus study reveals that these readings are rare, although they tend to be more frequent in questions than in newspaper texts.

2.3 Dimension

Part of the task of semantically interpretating superlative adjectives is the selection of the *dimension* on which entities are compared. In "the highest mountain" we compare mountains with respect to the dimension height, in "the best paper" we compare papers with respect to the dimension quality, and so on. A well-known problem is that some adjectives can be ambiguous or vague in choosing their dimension. Detecting the appropriate dimension is not covered in this paper, but is orthogonal to the analysis we provide.

2.4 Superlatives and Entailment

Superlatives exhibit a non-trivial semantics. Some examples of textual entailment make this very evident. Consider the contrasts in the following entailment tests with indefinite and universally quantified noun phrases:

I bought a blue car $\models$ I bought a car
I bought a car $\not\models$ I bought a blue car

I bought every blue car $\not\models$ I bought every car
I bought every car $\models$ I bought every blue car

Observe that the directions of entailments are mirrorred. Now consider a similar test with superlatives, where the entailments fail in both directions:

I bought the cheapest blue car $\not\models$ I bought the cheapest car
I bought the cheapest car $\not\models$ I bought the cheapest blue car.

These entailment tests underline the point that the meaning of superlatives is rather complicated, and that a shallow semantic representation, say λx.[cheapest(x) $\wedge$ car(x)] for "cheapest car", simply won't suffice. A semantic represention capturing the meaning of a superlative requires a more sophisticated analysis. In particular, it is important to explicitly represent the comparison set of a superlative. In "the cheapest car", the comparison set is formed by the set of cars, whereas in "the cheapest blue car", the comparison set is the set of blue cars. Semantically, we can represent "cheapest blue car" as follows, where the comparison set is made explicit in the antecedent of the conditional:

λx.[car(x) $\wedge$ blue(x) $\wedge$
$\forall$y((car(y) $\wedge$ blue(y) $\wedge$ x$\neq$y) $\rightarrow$ cheaper(x,y))]

Paraphrased in English, this stipulates that some blue car is cheaper than any other blue car. A meaning representation like this will logically predict the correct entailment relations for superlatives.

3 Annotated Corpus of Superlatives

In order to develop and evaluate our system we manually annotated a collection of newspaper article and questions with occurrences of superlatives. The design of the corpus and its characteristics are described in this section.

3.1 Classification and Annotation Scheme

Instances of superlatives are identified in text and classified into one of four possible classes: *attributive*, *predicative*, *adverbial*, or *idiomatic*:

its rates will be among the **highest** (predicative)

the **strongest** dividend growth (attributive)

free to do the task **most quickly** (adverbial)

who won the TONY for **best featured** actor? (idiom)

For all cases, we annotate the span of the superlative adjective in terms of the position of the tokens in the sentence. For instance, in "its$_1$ rates$_2$ will$_3$ be$_4$ among$_5$ the$_6$ highest$_7$", the superlative span would be 7–7.

Additional information is encoded for the attributive case: type of determiner (possessive, definite, bare, demonstrative, quantifier), number (sg, pl, mass), cardinality (yes, no), modification (adjective, ordinal, intensifier, none). Table 1 shows some examples from the WSJ with annotation values.

Not included in this study are adjectives such as "next", "past", "last", nor the ordinal "first", although they somewhat resemble superlatives in their semantics. Also excluded are adjectives that lexicalise a superlative meaning but are not superlatives morphologically, like "main", "principal", and the like. For etymological reasons we however include "foremost" and "uttermost."

3.2 Data and Annotation

Our corpus consists of a collection of newswire articles from the Wall Street Journal (Sections 00, 01, 02, 03, 04, 10, and 15) and the Glasgow Herald (GH950110 from the CLEF evaluation forum), and a large set of questions from the TREC QA evaluation exercise (years 2002 and 2003) and natural language queries submitted to the Excite search engine (Jansen and Spink, 2000). The data was automatically tokenised, but all typos and extra-grammaticalities were preserved. The corpus was split into a development set used for tuning the system and a test set for evaluation. The size of each sub-corpus is shown in Table 2.

Table 2: Size of each data source (in number of sentences/questions)

source	dev	test	total
WSJ	8,058	6,468	14,526
GH	—	2,553	2,553
TREC	1,025	—	1,025
Excite	—	67,140	67,140
total	9,083	76,161	85,244

The annotation was performed by two trained linguists. One section of the WSJ was annotated by both annotators independently to calculate inter-annotator agreement. All other documents were first annotated by one judge and then checked by the second, in order to ensure maximum correctness. All disagreements were discussed and resolved for the creation of a gold standard corpus.

Inter-annotator agreement was assessed mainly using f-score and percentage agreement as well as

Table 1: Annotation examples of superlative adjectives

example	sup span	det	num	car	mod	comp set
The **third-largest** thrift institution in Puerto Rico also [...]	2–2	def	sg	no	ord	3–7
The Agriculture Department reported that feedlots in the **13 biggest** ranch states held [...]	9–10	def	pl	yes	no	11–12
The failed takeover would have given UAL employees 75 % voting control of the nation 's **second-largest** airline [...]	17–17	pos	sg	no	ord	14–18

the kappa statistics (K), where applicable (Carletta, 1996). In using f-score, we arbitrarily take one of the annotators' decisions (A) as gold standard and compare them with the other annotator's decisions (B). Note that here f-score is symmetric, since precision(A,B) = recall(B,A), and (balanced) f-score is the harmonic mean of precision and recall (Tjong Kim Sang, 2002; Hachey et al., 2005, see also Section 5).

We evaluated three levels of agreement on a sample of 1967 sentences (one full WSJ section). The first level concerns superlative detection: to what extent different human judges can agree on what constitutes a superlative. For this task, f-score was measured at 0.963 with a total of 79 superlative phrases agreed upon.

The second level of agreement is relative to type identification (attributive, predicative, adverbial, idiomatic), and is only calculated on the subset of cases both annotators recognised as superlatives (79 instances, as mentioned). The overall f-score for the classification task is 0.974, with 77 cases where both annotators assigned the same type to a superlative phrase. We also assessed agreement for each class, and the attributive type resulted the most reliable with an f-score of 1 (total agreement on 64 cases), whereas there was some disagreement in classifying predicative and adverbial cases (0.9 and 0.8 f-score, respectively). Idiomatic uses where not detected in this portion of the data. To assess this classification task we also used the kappa statistics which yielded K_{Co}=0.922 (following (Eugenio and Glass, 2004) we report K as K_{Co}, indicating that we calculate K à la Cohen (Cohen, 1960). K_{Co} over 0.9 is considered to signal very good agreement (Krippendorff, 1980).

The third and last level of agreement deals with the span of the comparison set and only concerns attributive cases (64 out of 79). Percentage agreement was used since this is not a classification task and was measured at 95.31%.

The agreement results show that the task appears quite easy to perform for linguists. Despite the limited number of instances compared, this has also emerged from the annotators' perception of the difficulty of the task for humans.

3.3 Distribution

The gold standard corpus comprises a total of 3,045 superlatives, which roughly amounts to one superlative in every 25 sentences/questions. The overwhelming majority of superlatives are attributive (89.1%), and only a few are used in a predicative way (6.9%), adverbially (3.0%), or in idiomatic expressions (0.9%).[1] Table 3 shows the detailed distribution according to data source and experimental sets. Although the corpus also includes annotation about determination, modification, grammatical number, and cardinality of attributive superlatives (see Section 3.1), this information is not used by the system described in this paper.

Table 3: Distribution of superlative types in the development and evaluation sets.

type	dev		test			total
	WSJ	TREC	WSJ	GH	Excite	
att	240	43	218	68	2,145	2,714
pre	40	3	26	17	125	211
adv	17	2	22	9	41	91
idi	6	5	1	2	15	29
total	303	53	267	96	2,326	3,045

4 Automatic Analysis of Superlatives

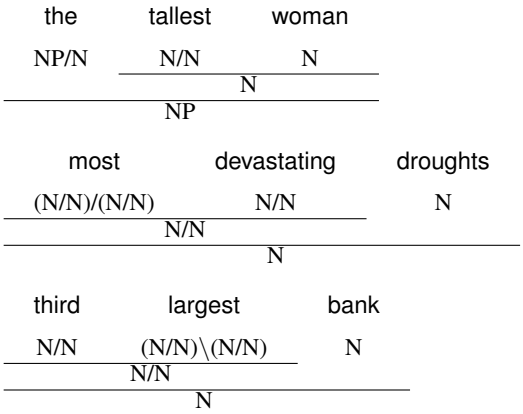
The system that we use to analyse superlatives is based on two linguistic formalisms: Combinatory Categorial Grammar (CCG), for a theory of syntax; and Discourse Representation Theory (DRT)

[1] Percentages are rounded to the first decimal and do not necessarily sum up to 100%.

for a theory of semantics. In this section we will illustrate how we extend these theories to deal with superlatives and how we implemented this into a working system.

4.1 Combinatory Categorial Grammar (CCG)

CCG is a lexicalised theory of grammar (Steedman, 2001). We used Clark & Curran's wide-coverage statistical parser (Clark and Curran, 2004) trained on CCG-bank, which in turn is derived from the Penn-Treebank (Hockenmaier and Steedman, 2002). In CCG-bank, the majority of superlative adjective of cases are analysed as follows:



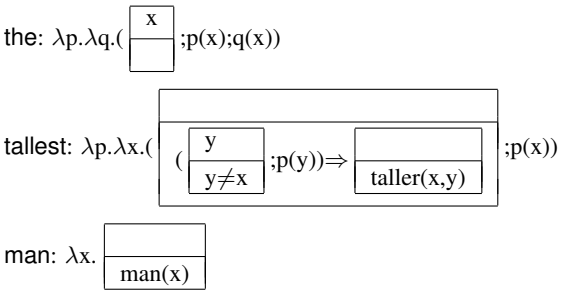
Clark & Curran's parser outputs besides a CCG derivation of the input sentence also a part-of-speech (POS) tag and a lemmatised form for each input token. To recognise attributive superlatives in the output of the parser, we look both at the POS tag and the CCG-category assigned to a word. Words with POS-tag JJS and CCG-category N/N, (N/N)/(N/N), or (N/N)\(N/N) are considered attributive superlatives adjectives, and so are the words "most" and "least" with CCG category (N/N)/(N/N).

However, most hyphenated superlatives are not recognised by the parser as JJ instead of JJS, and are corrected in a post-processing step.[2] Examples that fall in this category are "most-recent wave" and "third-highest".

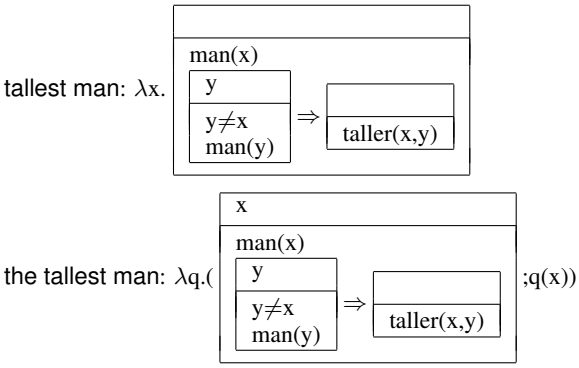
4.2 Discourse Representation Theory (DRT)

The output of the parser, a CCG derivation of the input sentence, is used to construct a Discourse Representation Structure (DRS, the semantic representation proposed by DRT (Kamp and Reyle, 1993)). We follow (Bos et al., 2004; Bos, 2005) in automatically building semantic representation on the basis of CCG derivations in a compositional fashion. We briefly summarise the approach here.

The semantic representation for a word is determined by its CCG category, POS-tag, and lemma. Consider the following lexical entries:



These lexical entries are combined in a compositional fashion following the CCG derivation, using the λ-calculus as a glue language:



In this way DRSs can be produced in a robust way, achieving high-coverage. An example output representation of the complete system is shown in Figure 1.

As is often the case, the output of the parser is not always what one needs to construct a meaningful semantic representation. There are two cases where we alter the CCG derivation output by the parser in order to improve the resulting DRSs. The first case concerns modifiers following a superlative construction, that are attached to the NP node rather than N. A case in point is

> ... the largest toxicology lab in New England ...

where the PP in New England has the CCG category NP\NP rather than N\N. This would result in a comparison set containing of toxicology labs, rather than a set toxicology labs in New England.

The second case are possessive NPs preceding a superlative construction. An example here is

> ... Jaguar's largest shareholder ...

[2] This is due to the fact that the Penn-Treebank annotation guidelines prescribe that all hyphenated adjectives ought to be tagged as JJ.

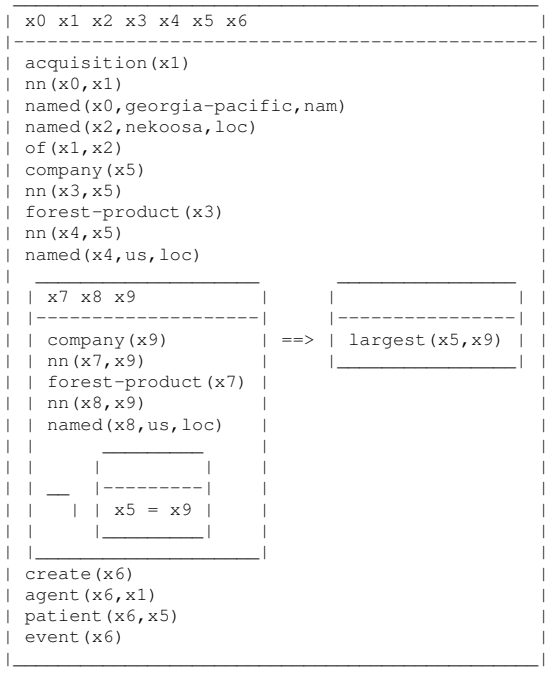
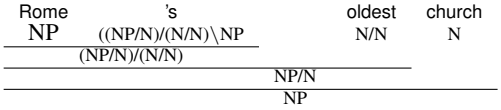


Figure 1: Example DRS output

where a correct interpretation of the superlative requires a comparison set of shareholders from Jaguar, rather than just any shareholder. However, the parser outputs a derivation where "largest" is combined with "shareholder", and then with the possessive construction, yielding the wrong semantic interpretation. To deal with this, we analyse possessives that interact with the superlative as follows:



This analysis yields the correct comparison set for superlative that follow a possessive noun phrase, given the following lexical semantics for the genitive:

$$\lambda n.\lambda S.\lambda p.\lambda q.(\boxed{u \mid \;};S(\lambda x.(p(x);n(\lambda y.\boxed{\;\mid \text{of}(y,x)})(u);q(u))))$$

For both cases, we apply some simple post-processing rules to the output of the parser to obtain the required derivations. The effect of these rules is reported in the next section, where we assess the accuracy of the semantic representations produced for superlatives by comparing the automatic analysis with the gold standard.

5 Evaluation

The automatic analysis of superlatives we present in the following experiments consists of two sequential tasks: *superlative detection*, and *comparison set determination*.

The first task is concerned with finding a superlative in text and its exact span ("largest", "most beautiful", "10 biggest"). For a found string to to be judged as correct, its whole span must correspond to the gold standard. The task is evaluated using precision (P), recall (R), and f-score (F), calculated as follows:

$$\text{P} = \frac{correct\ assignments\ of\ c}{total\ assignments\ of\ c}$$

$$\text{R} = \frac{correct\ assignments\ of\ c}{total\ corpus\ instances\ of\ c}$$

$$\text{F} = \frac{2P_cR_c}{P_c+R_c}$$

The second task is conditional on the first: once a superlative is found, its comparison set must also be identified ("rarest flower in New Zealand", "New York's tallest building", see Section 2.2). A selected comparison set is evaluated as correct if it corresponds exactly to the gold standard annotation: partial matches are counted as wrong. Assignments are evaluated using accuracy (number of correct decisions made) only on the subset of previously correctly identified superlatives.

For both tasks we developed simple baseline systems based on part-of-speech tags, and a more sophisticated linguistic analysis based on CCG and DRT (i.e. the system described in Section 4). In the remainder of the paper we refer to the latter system as DLA (Deep Linguistic Analysis).

5.1 Superlative Detection

Baseline system For superlative detection we generated a baseline that solely relies on part-of-speech information. The data was tagged using TnT (Brants, 2000), using a model trained on the Wall Street Journal. In the WSJ tagset, superlatives can be marked in two different ways, depending on whether the adjective is inflected or modified by most/least. So, "largest", for instance, is tagged as JJS, whereas "most beautiful" is a sequence of RBS (most) and JJ (beautiful). We also checked that they are followed by a common or proper noun (NN.*), allowing one word to occur in between. To cover more complex cases, we also considered pre-modification by adjectives (JJ), and cardinals (CD). In summary, we matched on sequences found by the following pattern:

```
[(CD || JJ)* (JJS || (RBS JJ)) * NN.*]
```

This rather simple baseline is capable of detecting superlatives such as "100 biggest banks", "fourth largest investors", and "most important

element", but will fail on expressions such as "fastest growing segments" or "Scotland 's lowest permitted 1995-96 increase".

DLA system For evaluation, we extrapolated superlatives from the DRSs output by the system. Each superlative introduces an implicational DRS condition, but not all implicational DRS conditions are introduced by superlatives. Hence, for the purposes of this experiment superlative DRS conditions were assigned a special mark. While traversing the DRS, we use this mark to retrieve superlative instances. In order to retrieve the original string that gave rise to the superlative interpretation, we exploit the meta information encoded in each DRS about the relation between input tokens and semantic information. The obtained string position can in turn be evaluated against the gold standard.

Table 4 lists the results achieved by the baseline system and the DLA system on the detection task. The DLA system outperforms the baseline system on precision in all sub-corpora. However, the baseline achieves a higher recall on the Excite queries. This is not entirely surprising given that the coverage of the parser is between 90–95% on unseen data. Moreover, Excite queries are often ungrammatical, thus further affecting the performance of parsing.

Table 4: Detection of Attributive Superlatives, reporting P (precision), R (Recall) and F-score, for WSJ sections, extracts of the Glasgow Herald, TREC questions, and Excite queries. D indicates development data, T test data.

	Baseline			DLA		
Corpus	**P**	**R**	**F**	**P**	**R**	**F**
WSJ (D)	0.93	0.86	0.89	0.96	0.90	0.93
WSJ (T)	0.91	0.83	0.87	0.95	0.87	0.91
GH (T)	0.80	0.76	0.78	0.87	0.81	0.84
TREC (D)	0.76	0.91	0.83	0.85	0.91	0.88
Excite (T)	0.92	0.92	0.92	0.97	0.84	0.90

5.2 Comparison Set Determination

Baseline For comparison set determination we developed two baseline systems. Both use the same match on sequences of part-of-speech tags described above. For Baseline 1, the beginning of the comparison set is the first word following the superlative. The end of the comparison set is the first word tagged as NN.* in that sequence (the same word could be the beginning and end of the comparison set, as it often happens).

The second baseline takes the first word after the superlative as the beginning of the comparison set, and the end of the sentence (or question) as the end (excluding the final punctuation mark). We expect this strategy to perform well on questions, as the following examples show.

Where is the oldest synagogue in the United States?
What was the largest crowd to ever come see Michael Jordan?

This approach is obviously likely to generate comparison sets much wider than required.

More complex examples that neither baseline can tackle involve possessives, since on the surface the comparison set lies at both ends of the superlative adjective:

The nation's largest pension fund
the world's most corrupt organizations

DLA 1 We first extrapolate superlatives from the DRS output by the system (see procedure above). Then, we exploit the semantic representation to select the comparison set: it is determined by the information encoded in the antecedent of the DRS-conditional introduced by the superlative. Again, we exploit meta information to reconstruct the original span, and we match it against the gold standard for evaluation.

DLA 2 DLA 2 builds on DLA 1, to which it adds post-processing rules to the CCG derivation, i.e. before the DRSs are constructed. This set of rules deal with NP post-modification of the superlative (see Section 4).

DLA 3 In this version we include a set of post-processing rules that apply to the CCG derivation to deal with possessives preceding the superlative (see Section 4).

DLA 4 This is a combination of DLA 2 and DLA 3. This system is clearly expected to perform best.

Results for both baseline systems and all versions of DLA are shown in Table 5

On text documents, DLA 2/3/4 outperform the baseline systems. DLA 4 achieves the best performance, with an accuracy of 69–83%. On questions, however, DLA 4 competes with the baseline: whereas it is better on TREC questions, it performs worse on Excite questions. One of the obvious reasons for this is that the parser's model

Table 5: Determination of Comparison Set of Attributive Superlatives (Accuracy) for WSJ sections, extracts of the Glasgow Herald, TREC and Excite questions. D indicates development data, T test data.

Corpus	Base 1	Base 2	DLA 1	DLA 2	DLA3	DLA 4
WSJ (D)	0.29	0.17	0.29	0.52	0.53	0.78
WSJ (T)	0.31	0.22	0.32	0.59	0.53	0.83
GH (T)	0.23	0.31	0.22	0.51	0.38	0.69
TREC (D)	0.10	0.69	0.13	0.69	0.23	0.82
Excite (T)	0.23	0.90	0.32	0.82	0.33	0.84

for questions was trained on TREC data. Additionally, as noted earlier, Excite questions are often ungrammatical and make parsing less likely to succeed. However, the baseline system, by definition, does not output semantic representations, so that its outcome is of little use for further reasoning, as required by question answering or general information extraction systems.

6 Conclusions

We have presented the first empirically grounded study of superlatives, and shown the feasibility of their semantic interpretation in an automatic fashion. Using Combinatory Categorial Grammar and Discourse Representation Theory we have implemented a system that is able to recognise a superlative expression and its comparison set with high accuracy.

For developing and testing our system, we have created a collection of over 3,000 instances of superlatives, both in newswire text and in natural language questions. This very first corpus of superlatives allows us to get a comprehensive picture of the behaviour and distribution of superlatives in real occurring data. Thanks to such broad view of the phenomenon, we were able discover issues previously unnoted in the formal semantics literature, such as the interaction of prenominal possessives and superlatives, which cause problems at the syntax-semantics interface in the determination of the comparison set. Similarly problematic are hyphenated superlatives, which are tagged as normal adjectives in the Penn Treebank.

Moreover, this work provides a concrete way of evaluating the output of a stochastic wide-coverage parser trained on the CCGBank (Hockenmaier and Steedman, 2002). With respect to superlatives, our experiments show that the quality of the raw output is not entirely satisfactory. However, we have also shown that some simple post-processing rules can increase the performance considerably. This might indicate that the way superlatives are annotated in the CCGbank, although consistent, is not fully adequate for the purpose of generating meaningful semantic representations, but probably easy to amend.

7 Future Work

Given the syntactic and semantic complexity of superlative expressions, there is still wide scope for improving the coverage and accuracy of our system. One obvious improvement is to amend CCGbank in order to avoid the need for postprocessing rules, thereby also allowing the creation of more accurate language models. Another aspect which we have neglected in this study but want to consider in future work is the interaction between superlatives and focus (Heim, 1999; Gawron, 1995). Also, only one of the possible types of superlative was considered, namely the attributive case. In future work we will consider the interpretation of predicative and adverbial superlatives, as well as comparative expressions. Finally, we would like to investigate the extent to which existing NLP systems (such as open-domain QA systems) can benefit from a detailed analysis of superlatives.

Acknowledgements

We would like to thank Steve Pulman (for information on the analysis of superlatives in the Core Language Engine), Mark Steedman (for useful suggestions on an earlier draft of this paper), and Jean Carletta (for helpful comments on annotation agreement issues), as well as three anonymous reviewers for their comments. We are extremely grateful to Stephen Clark and James Curran for making their parser available to us. Johan Bos is supported by a "Rientro dei Cervelli" grant (Italian Ministry for Research); Malvina Nissim is supported by the EU FP6 NeOn project.

References

Hiyan Alshawi, editor. 1992. *The Core Language Engine*. The MIT Press, Cambridge, Massachusetts.

J. Bos, S. Clark, M. Steedman, J.R. Curran, and Hockenmaier J. 2004. Wide-Coverage Semantic Representations from a CCG Parser. In *Proceedings of*

the 20th International Conference on Computational Linguistics (COLING '04), Geneva, Switzerland.

Johan Bos. 2005. Towards wide-coverage semantic interpretation. In *Proceedings of Sixth International Workshop on Computational Semantics IWCS-6*, pages 42–53.

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA*.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

S. Clark and J.R. Curran. 2004. Parsing the WSJ using CCG and Log-Linear Models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*, Barcelona, Spain.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*, 20:37–46.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: a second look. *Computational Linguistics*, 30(1).

Donka F. Farkas and Katalin È. Kiss. 2000. On the comparative and absolute readings of superlatives. *Natural Language and Linguistic Theory*, 18:417–455.

Jean Mark Gawron. 1995. Comparatives, superlatives, and resolution. *Linguistics and Philosophy*, 18:333–380.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the 9th Conference on Computational Natural Language Learning, Ann Arbor, Michigan, USA*.

Irene Heim. 1999. Notes on superlatives. MIT.

J. Hockenmaier and M. Steedman. 2002. Generative Models for Statistical Parsing with Combinatory Categorial Grammar. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA.

Bernard J. Jansen and Amanda Spink. 2000. The excite research project: A study of searching by web users. *Bulletin of the American Society for Information Science and Technology*, 27(1):5–17.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

M. Steedman. 2001. *The Syntactic Process*. The MIT Press.

Anna Szabolcsi. 1986. Comparative superlatives. In N. Fukui et al., editor, *Papers in Theoretical Linguistics, MITWPL*, volume 8. MIT.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.