Improved statistical measures to assess natural language parser performance across domains

Barbara Plank

Alfa informatica, Faculty of Arts
University of Groningen, The Netherlands
b.plank@rug.nl

Abstract

We examine the performance of three dependency parsing systems, in particular, their performance variation across Wikipedia domains. We assess the performance variation of (i) Alpino, a deep grammar-based system coupled with a statistical disambiguation versus (ii) MST and Malt, two purely data-driven statistical dependency parsing systems. The question is how the performance of each parser correlates with simple statistical measures of the text (e.g. sentence length, unknown word rate, etc.). This would give us an idea of how sensitive the different systems are to domain shifts, i.e. which system is more in need for domain adaptation techniques. To this end, we extend the statistical measures used by Zhang and Wang (2009) for English and evaluate the systems on several Wikipedia domains by focusing on a freer word-order language, Dutch. The results confirm the general findings of Zhang and Wang (2009), i.e. different parsing systems have different sensitivity against various statistical measure of the text, where the highest correlation to parsing accuracy was found for the measure we added, sentence perplexity.

1. Introduction

Natural Language Parsing has become an essential part for many Natural Language Processing task. For instance, Question Answering or Machine Translation. Yet, parsing system are very sensitive to the domain they were trained on, as their performance might drop dramatically when the system gets input from another text domain (Gildea, 2001). This is the problem of *domain adaptation*.

Although the problem exists ever since the emerge of supervised Machine Learning, it has started to get attention only in recent years. Studies on *supervised domain adaptation* (where there are limited amounts of annotated resources in the new domain) have shown that straightforward baselines (e.g. models based on source only, target only, or the union of the data) achieve a relatively high performance level and are "surprisingly difficult to beat" (Daumé III, 2007).

In contrast, *semi-supervised adaptation* (i.e. no annotated resources in the new domain) is a much more realistic situation but is clearly also considerably more difficult. Current studies on semi-supervised approaches show very mixed results. Dredze et al. (2007) report on "frustrating" results on the CoNLL 2007 semi-supervised adaptation task for dependency parsing, i.e. "no team was able to improve target domain performance substantially over a state-of-the-art baseline". On the other hand, there have been positive results as well. For instance, McClosky et al. (2006) improved a statistical consituency parser by self-training. Structural Correspondence Learning (Blitzer et al., 2006) was effective for PoS tagging and Sentiment Classification (Blitzer et al., 2006; Blitzer et al., 2007), while modest gains were obtained for structured output tasks like Parsing.

The question addressed in this study is: How does parser performance for Dutch correlate to simple statistical measures of the text? We assess the performance variation of two different kinds of parsing systems on various Wikipedia domains. This can be seen as a first step towards examining the question of how sensitive a parsing system is to the text domain, i.e. which parsing system (hand-crafted versus purely statistical) is more affected by domains shifts, and thus more in need for adaptation techniques.

2. Related Work

Zhang and Wang (2009) examined several state-of-the-art parsing systems for English, and showed that parsing models correlate on different levels to the three statistical measures examined (average sentence length, unknown word ratio and unknown part-of-speech trigram ratio) when tested on the Brown corpus. We start here from their work and examine dependency parser performance for a freer word-order language, Dutch, by examining parsing performance on various Wikipedia subdomains. A related study is Ravi et al. (2008), who build a parser performance predictor system for English constituency parsing.

3. Parsing Systems

We examine two kinds of parsing systems for Dutch: a grammar-based system coupled with a statistical disambiguation system (Alpino) and two data-driven system (MST and Malt). Details about the parsers are given in the sequel.

(1) *Alpino* (van Noord, 2006) is a deep-grammar based parser for Dutch that outputs dependency structure. The system consists of approximately 800 grammar rules in the tradition of HPSG, and a large hand-crafted lexicon, that together with a left-corner parser constitutes the generation component. For words that are not in the lexicon, the system applies a large variety of unknown word heuristics (van Noord, 2006), which among others attempt to deal with number-like expressions, compounds and proper names. The second stage of Alpino is a statistical disambiguation component based on Maximum Entropy. Thus, training the parser requires estimating parameters for the disambiguation component.

(2) *MST Parser* (McDonald et al., 2005) is a language-independent graph-based dependency parser. The system couples a minimum spanning tree search procedure with

a separate second stage classifier to label the dependency edges.

(3) *Malt Parser* (Nivre et al., 2007) is a language-independent transition-based dependency parser. Malt parser uses SVMs to learn a classifier that predicts the next parsing action. Instances represent parser configurations and the label to predict determines the next parser action.

Both data-driven parsers (MST and Malt) are thus not specific for the Dutch Language, however, they can be trained on a variety of languages given that the training corpus complies with the column-based format introduced in the 2006 CoNLL shared task (Buchholz and Marsi, 2006). Additionally, both parsers implement projective and non-projective parsing algorithms, where the latter will be used in our experiments on the relatively free word-order language Dutch. Despite that, we train the data-driven parsers using their default settings (e.g. first order features for MST, SVM with polynomial kernel for Malt).

4. Datasets and Treebank conversion

We train the MST and Malt Parser as well as the disambiguation component of Alpino on cdb, the standard Alpino Treebank. For our cross-domain evaluation, we consider various Wikipedia articles from the Dutch Wikipedia project.

Cdb The cdb (Alpino Treebank) consists of 7,136 sentences from the Eindhoven corpus (newspaper text). It is a collection of text fragments from 6 Dutch newspapers (het Nieuwsblad van het Noorden, de Telegraaf, de Tijd, Trouw, het Vrije Volk, Nieuwe Rotterdamse Courant). The collection has been annotated according to the guidelines of CGN (Oostdijk, 2000) and stored in XML format. It is the standard Treebank used to train the disambiguation component of the Alpino parser.

Wikipedia We use 95 Dutch Wikipedia articles which were annotated in the course of the LASSY project.[1] They are mostly about Belgium issues, i.e. locations, politics, sports, arts, etc. We have grouped them into ten subdomains, as specified in Table 1, which also gives an overview of the size of these datasets.

CoNLL2006 This is the test file for Dutch that has been used in the CoNLL 2006 shared task on multi-lingual dependency parsing. The file consists of 386 sentences from an institutional brochure (about 'Jeugdgezondheidszorg'/youth healthcare). We use this file to check our data-driven models against state-of-the-art performance.

Alpino to CoNLL format In order to train the MST parser and evaluate it on the various Wikipedia articles, we needed to convert the Alpino Treebank format into the tabular CoNLL format. To this end, we adapted the treebank conversion software developed by Erwin Marsi for the CoNLL 2006 shared task on multi-lingual dependency parsing. Instead of using the PoS tagger and tagset used in the CoNLL shared task (to which we did not have access to), we replaced the PoS tags with more fine-grained tags obtained by parsing the data with the Alpino parser.[2]

Domain	**Wikipedia articles (excerpt)**	**# art.**	**# sents**	**# words**
BUS (business)	Algemeen Belgisch Vakverbond	9	405	4440
COM (comics)	Suske en Wiske	3	380	4000
HIS (history)	Geschiedenis van België	3	468	8396
HOL (holidays)	Feest van de Vlaamse Gemeenschap	4	43	524
KUN (arts)	School van Tervuren	11	998	17073
LOC (location)	België, Brussel (stad)	31	2190	25259
MUS (music)	Sandra Kim, Urbanus (artiest)	3	89	1296
NOB (nobility)	Albert II van België	6	277	4179
POL (politics)	Belgische verkiezingen 2003	16	983	15107
SPO (sports)	Spa-Francorchamps, Kim Clijsters	9	877	9713
Total		**95**	**6710**	**89987**

Table 1: Overview Wikipedia corpus including number of articles (art.), sentences (sents) and words.

5. Features and Evaluation

We follow Zhang and Wang (2009) and look at this stage at simple characteristics of the dataset without looking at syntactic annotation. We are interested how they correlate to parsing performance for the three parsing systems: Alpino, MST and Malt parser. We depart from their feature set (Zhang and Wang, 2009) and add a perplexity feature estimated from a trigram Language Model.

Sentence Length (l) measures the average sentence length. Intuitively, longer sentences should be more difficult to parse than shorter ones.

Simple Unknown Word Rate (sUWR) calculates how many words (tokens) in the dataset have not been observed before, i.e. are not in the cdb corpus. For the Alpino parser, we use the percentage of words that are not in the lexicon (aUWR, Alpino Unknown Word Rate).

Unknown PoS Trigram Ration (UPTR) calculates the number of unknown PoS trigrams with respect to the original cdb training data.

Perplexity (ppl) is the perplexity score assigned by a word-trigram language model estimated from the original cdb training data. This feature, also used by (Ravi et al., 2008), is intended as a refinement of the unknown word rate feature.

Evaluation In contrast to Zhang and Wang (2009), we evaluate each parser with the same evaluation metric: Labeled Attachment Score (LAS). That is, performance is determined by the percentage of tokens with the correct dependency edge and label. To compute LAS, we use the CoNLL 2007 evaluation script[3] with punctuation tokens excluded from scoring (as was the default setting in CoNLL 2006). Note that the standard metric for Alpino would be a variant of LAS, which allows for a discrepancy between expected and returned dependencies. Such a discrepancy can occur, for instance, because the syntactic annotation of Alpino allows words to be dependent on more than a single head ('secondary edges') (van Noord, 2006). However, such edges are ignored in the CoNLL format; just a single head per token is allowed. Furthermore, there is another simplification. As the Dutch tagger used in the CoNLL

[1] LASSY (Large Scale Syntactic Annotation of written Dutch), ongoing project. Corpus version 17905, obtained from `http://www.let.rug.nl/vannoord/Lassy/corpus/`

[2] The datasets in retagged CoNLL format are available at `http://www.let.rug.nl/bplank/alpino2conll`.

[3] `http://nextens.uvt.nl/depparse-wiki/SoftwarePage`

2006 shared task did not have the concept of multiwords, the organizers chose to treat them as a single token (Buchholz and Marsi, 2006). We here follow the CoNLL 2006 task setup.

6. Experimental Results

First of all, we performed a sanity check and trained the MST and Malt parser on the cdb corpus converted into the retagged CoNLL format, and tested on CoNLL 2006 test data (also retagged). As seen in table 2, the performance level corresponds to state-of-the-art performance for statistical parsing on Dutch, and is actually even higher. We believe this increase in performance can be attributed to two sources: (a) the more fine-grained PoS tagset obtained by parsing the data with the deep grammar; (b) improvements in the Alpino treebank itself over the course of the years.

Model	**LAS**
MST (train data: cdb retagged)	82.14
Malt (train data: cdb retagged)	80.64
MST (Nivre and McDonald, 2008)	79.19
Malt (Nivre and McDonald, 2008)	78.59

Table 2: Performance of the data-driven parsers versus state-of-the-art performance on the CoNLL 2006 test set.

We now turn to the various statistical measures. The parsers were all evaluated on the 95 Wikipedia articles. Figure 1 plots the correlation between each parser's performance and the four measures: average sentence length (l), simple unknown word rate (sUWR, as well as aUWR for Alpino), unknown pos trigram rate (UPTR) and perplexity (ppl). The first row shows the Alpino parser, the second row shows the MST parser and the third row the Malt parser.

Pre-result Three datasets immediately catch our eyes (the red crossed dots; cf. the graphs about sentence length or perplexity in Figure 1): these are three sports (SPO) articles about bike races. By inspecting them we see that they contain a long list of winners from the various race years (on average 86% of the articles constitute this 'winner list'). Thus, despite the average short sentence length (6.03 words per sentence; in contrast to an average sentence length over all Wikipedia articles of 13.68 words), the parsers exhibit very different performance levels on these datasets. Alpino, who includes various unknown word heuristics and a named entity tagger, is rather robust against the very high unknown word rate and reaches a very high accuracy level on these datasets. The Malt parser also reaches a high performance level on this special datasets. In contrast, the MST parser is more influenced by unknown words, and the performance on these articles drops actually to its lowest level. These three sports articles thus form 'outliers' and we exclude them from the remaining experiments.

Results Figure 2 depicts parser performance against the four statistical measures of the text on the Wikipedia data with the three aforementioned sports articles removed. All parsers are robust to average sentence length (leftmost graphs in Figure 2). They basically do not show any correlation with this measure. This is in line with the results of Zhang and Wang (2009) for MST and Malt. It is different for the grammar-based parsing system. Their grammar-based parser (ERG) is highly sensitive to average sentence length (correlation coefficient of -0.61 on their datasets), as longer sentences "lead to a sharp drop in parsing coverage of ERG" (Zhang and Wang, 2009). This is not the case for the Alpino parser. The system suffers less from coverage problems and is thus not so sensitive against increasing sentence length.

For Unknown Word Rate (UWR), the data-driven parsers show a high correlation with this measure (correlation of -0.39 and -0.28), which is in line with previous findings (Zhang and Wang, 2009). This is not the case for Alpino: again, its very good handling of unknown words make the system robust to UWR. Note that for Alpino the unknown word rate is measured in a slightly different way (i.e. words not in the lexicon). However, if we would apply the same simple unknown word rate (sUWR) measure to Alpino, it would also result in a weak negative correlation only ($sUWR = -0.07$). Thus, Alpino does not seem to be sensitive to this measure.

No parser does show any correlation with the third measure, Unknown Part-of-Speech Trigram Rate (UPTR). This is contrary to previous results (Zhang and Wang, 2009), most probably due to the usage of a different tagset and the freer word-order language.

Our last measure, sentence perplexity, exhibits the highest correlation to parsing performance: all parsers show the highest sensitivity against this measure, with the data-driven parsers being slightly more sensitive ($cor = -0.67$ and $cor = -0.57$) than the grammar-driven parser Alpino ($cor = -0.33$). Note that this still holds if we would remove two other possible 'outliers', the turquoise diamond and grey star on the right bottom of Figure 2 (right-most graphs), resulting in a correlation coefficient of: Alpino $cor = -0.12$, MST $cor = -0.57$ and Malt $cor = -0.34$. Moreover, also on another corpus (DPC, the Dutch Parallel Corpus[4]) sentence perplexity gave us the highest correlation to parsing performance.

Finally, because evaluation metrics are directly comparable, the figures show that the Alpino parser, tailored to the language, reaches an overall higher performance level (between 80 and 100% LAS) than the data-driven counterparts (between 50 and 95% LAS).

7. Conclusion and Future work

We evaluated a deep grammar-based system coupled with a statistical disambiguation system (Alpino) and two data-driven parsers (MST and Malt) for dependency parsing of Dutch. The empirical evaluation was performed on Wikipedia domains.

By looking at four simple statistical measure of the text and their correlation to parsing performance, we could confirm the general result found by Zhang and Wang (2009): different parsing systems have different sensitivity against statistical measures of the text. While they evaluated parsing systems for English, we here looked at dependency parsing for a freer word-order language as Dutch.

[4] http://www.kuleuven-kortrijk.be/dpc

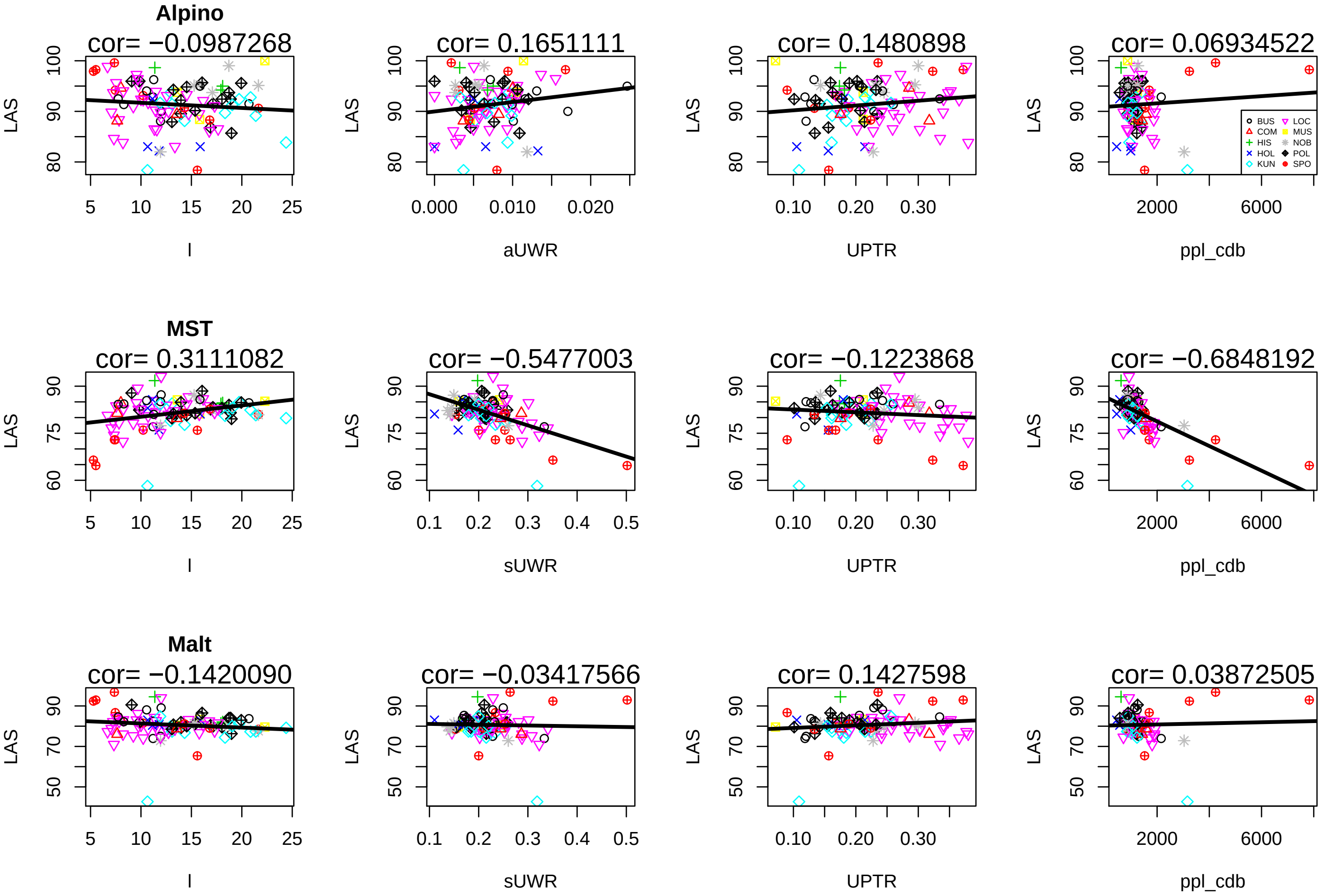

Figure 1: Pre-results (on 95 Wikipedia articles): Parser performance on a per-article basis against statistical measure on the text, including correlation coefficient.

Both data-driven parsers show a high correlation to unknown word rate, while this is not the case for the grammar-based system. The highest correlation with parsing accuracy was found for the measure we added, sentence perplexity. This is true for both kinds of parsing systems, grammar-based and data-driven, but especially for the statistical parsers MST and Malt. This might first seem counterintuitive, as a grammar-based system usually suffers more from coverage problems. However, Alpino successfully implements a set of unknown word heuristics to achieve robustness. For instance, on the 'bike winners list' sports domain, which we could identify through these simple statistical measures, Alpino and MST indeed exhibit a very different performance level, showing that the grammar-based system suffered less from the peculiarities of that domain.

In future, we would like to extend this line of work. The most immediate step is to integrate more statistical measures of the text and go towards building a 'parse performance predictor'. We might see that as a proxy for domain difference, i.e. to give us a rough estimate on how far or difficult a given text is for a parsing system.

8. References

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, Prague, Czech Republic.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *In Proc. of CoNLL*, pages 149–164.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Mark Dredze, John Blitzer, Pratha Pratim Talukdar, Kuzman Ganchev, Joao Graca, and Fernando Pereira. 2007. Frustratingly hard domain adaptation for parsing. In *Proceedings of the CoNLL Shared Task Session - Conference on Natural Language Learning*, Prague, Czech

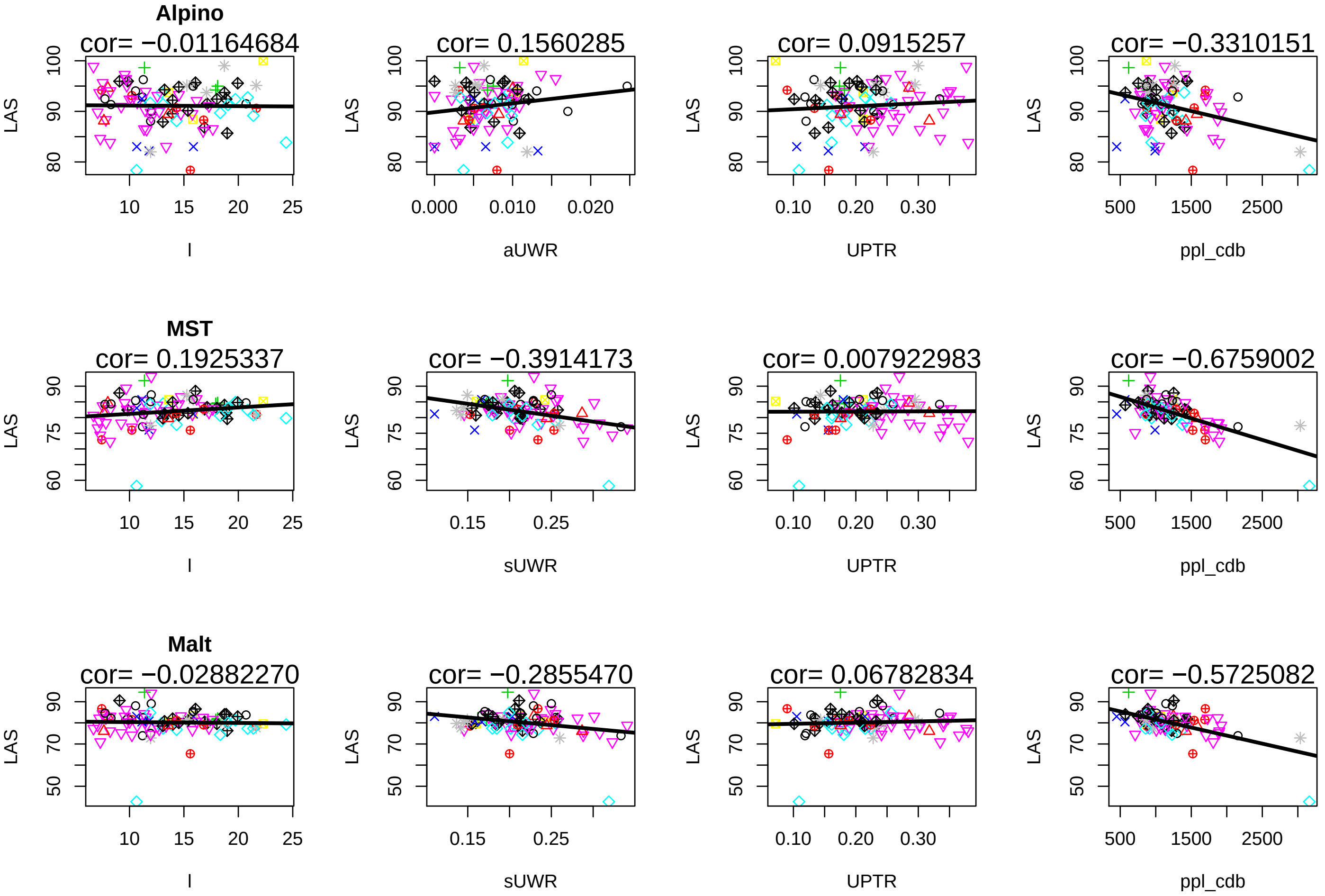

Figure 2: Results: Parser performance on a per-article basis against statistical measure on the text (93 Wikipedia articles - 3 sports articles removed).

Republic.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530.

Joakim Nivre and Ryan McDonald. 2008. Integrating graph-based and transition-based dependency parsers. In *Proceedings of ACL-08: HLT*, pages 950–958, Columbus, Ohio, June. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13:95–135.

Nelleke Oostdijk. 2000. The Spoken Dutch Corpus: Overview and first evaluation. In *Proceedings of Second International Conference on Language Resources and Evaluation (LREC)*, pages 887–894.

Sujith Ravi, Kevin Knight, and Radu Soricut. 2008. Automatic prediction of parser accuracy. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Morristown, NJ, USA. Association for Computational Linguistics.

Gertjan van Noord. 2006. **A**t **L**ast **P**arsing **I**s **N**ow **O**perational. In *TALN 2006 Verbum Ex Machina, Actes De La 13e Conference sur Le Traitement Automatique des Langues naturelles*, pages 20–42, Leuven.

Yi Zhang and Rui Wang. 2009. Correlating natural language parser performance with statistical measures of the text. In *In Proceedings of KI 2009*, Paderborn, Germany.