Simultaneous error detection at two levels of syntactic annotation

Adam Przepiórkowski
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warszawa, Poland
adamp@ipipan.waw.pl

Michał Lenart
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warszawa, Poland
michal.lenart@ipipan.waw.pl

Abstract

The paper describes a method for measuring compatibility between two levels of manual corpus annotation: shallow and deep. The proposed measures translate into a procedure for finding annotation errors at either level.

1 Introduction

Syntactic parsers are typically evaluated against manually or semi-automatically developed treebanks. Although, in evaluation tasks, such hand-produced resources are treated as if they were error-free, it is well known that even the most carefully annotated corpora contain errors. Some attention has been given to this problem within the last decade, and statistical techniques have been proposed to locate untypical – and, hence, possibly erroneous – annotations.

In this paper we examine a related issue, namely, the possibility of finding annotation errors by comparing two independently annotated levels of syntactic annotation: shallow (roughly: chunking) and deep (fully connected syntactic trees spanning the whole sentence).

2 Related Work

There are two strands of work relevant to the current enterprise. First, there is a line of work on discovering errors in manually annotated corpora (van Halteren 2000, Eskin 2000, Dickinson and Meurers 2003a), including treebanks (Dickinson and Meurers 2003b, Boyd et al. 2008, Dickinson and Lee 2008, Kato and Matsubara 2010). These methods concentrate on finding inconsistencies in linguistic annotations: if similar (in some well-defined way) inputs receive different annotations, the less frequent of these annotations is suspected of being erroneous. Experiments (reported elsewhere) performed on a Polish treebank show that such methods reach reasonable precision but lack in recall.

The second relevant line of research is concerned with the evaluation of syntactic parsers. The standard measure is the so-called Parseval measure (Black et al. 1991), used in the eponymous series of competitions. It calculates precision and recall on the set of (perhaps labelled, Magerman 1995) spans of words, i.e., on brackets identified in parse results and in the gold standard. Unfortunately, this measure – regardless of the fact that it has been repeatedly criticised on various grounds (Briscoe and Carroll 1996, Sampson and Babarczy 2003, Rehbein and van Genabith 2007, Kübler et al. 2008) – is not applicable to the current problem, as spans of discovered constituents are very different *by design*.

A more promising measure, older than Parseval (cf. Sampson et al. 1989), but gaining prominence only recently, is Leaf-Ancestor (LA; Sampson 2000, Sampson and Babarczy 2003), which compares trees word-by-word. For each word, the similarity of the path from this word to the root of the tree in both trees is calculated as a number in $\langle 0, 1 \rangle$, and the mean of these similarities over all words in a sentence is the score for this sentence.[1] While also not

[1] The very lenient IOB (Ramshaw and Marcus 1995, Tjong Kim Sang and Veenstra 1999) accuracy measure, used sometimes in chunking, can be considered as an extreme case of the LA measure.

Proceedings of the 6th Linguistic Annotation Workshop, pages 118–123,
Jeju, Republic of Korea, 12-13 July 2012.

directly applicable to the current scenario, this measure is much more flexible, as path similarity may be defined in various ways. The method proposed in section 4 has been inspired by this measure. Another general source of inspiration have been evaluation measures used in dependency parsing, where the notion of head is of paramount importance.

3 Levels of Syntactic Annotation

Among the various levels of linguistic annotation in the National Corpus of Polish (`http://nkjp.pl/`; NKJP; Przepiórkowski et al. 2010), two are immediately relevant here: morphosyntax (roughly, parts of speech and values of grammatical categories such as case or gender) and shallow syntactic groups. A 1-million-word subcorpus of NKJP was semi-automatically annotated at these levels: first relevant tools (morphological analyser, shallow grammar) were used to automatically add mark-up and then human annotators carefully (2 annotators per sentence plus a referee) selected the right interpretation, often correcting the automatic outcome.

In a related project (Woliński et al. 2011), the morphosyntactic level was used as a basis for constructing the level of deep syntax. Again, sentences were run through a deep parser and human annotators carefully selected the right parse.

The two syntactic annotation layers, illustrated in Figure 1, are described in more detail below.

3.1 Shallow Syntax

By shallow syntactic annotation we understand here a little more than chunking (Abney 1991): various types of basic groups are found (nominal, prepositional, adverbial, sentential), each marked with a syntactic head and a semantic head, and some hierarchical structure is allowed to the extent that sentential groups may contain smaller groups (including sentential ones). On the other hand, the general chunking principle of not resolving attachment ambiguities is preserved, so, e.g., instead of the nested structure $[\mathrm{P}\ [\mathrm{NP}\ [\mathrm{P}\ \mathrm{NP}]_{\mathrm{PP}}]_{\mathrm{NP}}]_{\mathrm{PP}}$ for *w kolejce do kasy* in the right-hand tree in Fig. 1, two smaller $[\mathrm{P}\ \mathrm{N}]_{\mathrm{PP}}$ constituents are marked at the shallow level (cf. the tree on the left).[2]

3.2 Deep Syntax

Complete constituent trees are assigned to sentences at the deep syntactic level. Labels of pre-terminals reflect parts of speech (e.g., *przyimek* 'preposition' or *formarzecz* 'nominal form'), higher non-terminal labels mostly correspond to standard labels such as PP (*fpm*), NP (*fno*), VP (*fwe*, understood here rather as a verbal group) or S (*zdanie*), with an additional level containing information about argument (*fw*) or non-argument (*fl*) status of phrases. No further dependency-like information is provided, i.e., there is no special marking of subjects, direct objects, etc.

4 Comparing Annotation Levels

Let us first note that all measures mentioned above are symmetrical in the sense that the evaluation of tree T_1 against tree T_2 gives the same results – perhaps after swapping precision and recall – as the evaluation of T_2 against T_1. In the current scenario, the two annotation schemata are rather different, with the shallow level containing – by design – fewer and smaller constituents. Hence, two different measures of precision are needed for the two levels (each measure having the dual role of measuring recall of the other level).

Second, since both annotation schemata assume the existence of syntactic heads for all constituents (see the thick lines in Fig. 1), and – together with dependency grammarians, practitioners of HPSG (Pollard and Sag 1994), etc. – we take headedness to be a crucial property of constituents, the proposed measures will build on this notion.

Let us first start with the types of shallow groups that cannot be nested, i.e., nominal, prepositional, etc., but not sentential. We define shallow precision, P_s, as the percentage of those segments contained in such groups which are annotated consistently with deep syntax:

$$P_s = \frac{|\{w : \exists G\ w \in yield(G) \wedge c(w, G)\}|}{|\{w : \exists G\ w \in yield(G)\}|}, \quad (1)$$

where w ranges over words, G ranges over (non-sentential) groups, and $c(w, G)$ is the compatibility predicate, which is true if and only if the annotation

[2] Note that non-terminal labels used in the figure differ from the ones used in text, and that in particular the deep tree uses Polish mnemonic names such as *fno* (*fraza nominalna*, nominal phrase). We hope that – given explanations in text – this does not lead to much confusion.

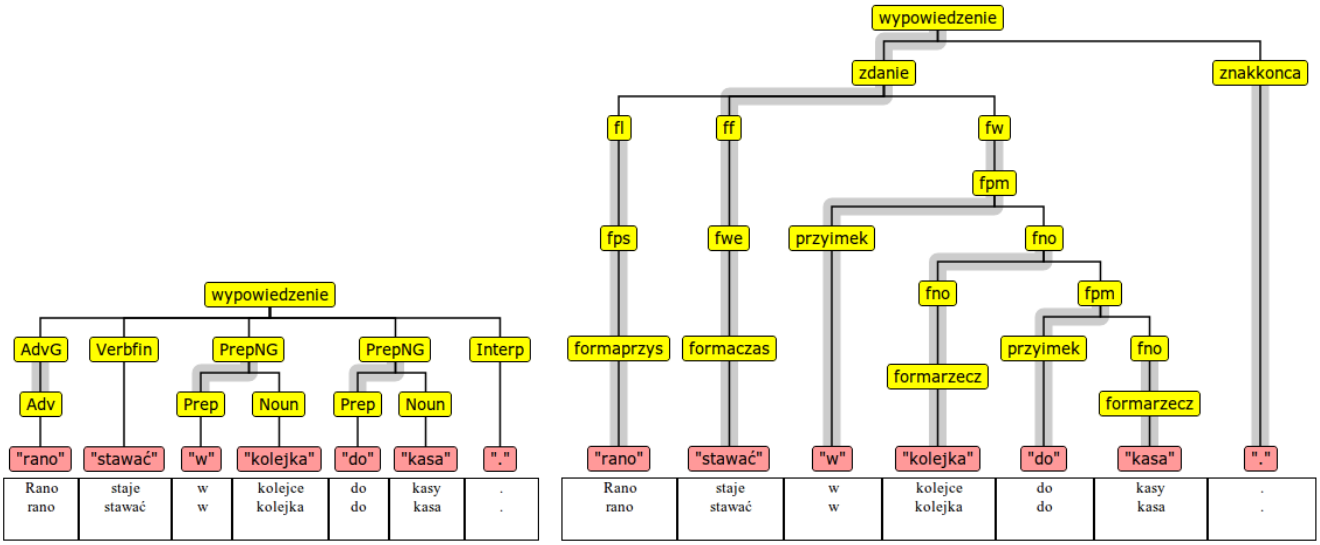

Figure 1: An example of shallow (on the left) and deep (on the right) syntactic annotation of *Rano staje w kolejce do kasy.* 'In the morning, (s)he queues to the cash desk.', lit. 'morning stands in queue to cash-desk'. In the shallow annotation, an artificial root (*wypowiedzenie* 'utterance') is added to connect all words and groups.

of w is compatible across the two levels. More precisely, $c(w, G)$ is true iff there exists a phrase F at the deep annotation of the same sentence such that $w \in yield(F)$, and also G and F have the same lexical heads. These conditions imply that w has the same headedness status with respect to G and F, i.e., it is either the head of both or of neither.

A labelled version of P_s, marked as lP_s, additionally requires that labels of G and F are compatible, in the sense of a manually defined mapping that relates – to give examples based on Fig. 1 – *PrepNG* to *fpm*, *AdvG* to *fps*, etc.

Applying this measure to Fig. 1 we note that there are 5 words belonging to some shallow group (*Rano*, *w*, *kolejce*, *do*, *kasy*). All these words, together with their respective groups, satisfy $c(w, G)$ and the condition on labels, so both P_s and lP_s are 1.0. For example, for w = *kolejce*, G is the *PrepNG* yielding *w kolejce*, whose head is the preposition *w*. Consequently, F is the *fpm* yielding *w kolejce do kasy*.

Deep precision, P_d, is defined in a similar way, but we are only interested in words w which are *more or less directly* contained in a phrase of a type corresponding to the types of groups considered here (i.e., nominal, prepositional, etc.). We say that w is *more or less directly* contained in F iff the path from w to F does not contain any sentential labels.[3] For every such word w we require that for one of its *more or less directly* dominating phrases, F, there is a corresponding shallow group G with the same head as F and also containing w; in case of labelled deep precision, lP_d, the labels of F and G should also match. For the deep annotation in Fig. 1, both unlabelled and labelled precision is again 1.0. This means that the two trees in this figure match perfectly, given the differing annotation schemata.

Recall that above measures do not take into account sentential constituents. This is due to the fact that finding clauses is not typically part of shallow parsing, and also in the current setup it is limited to complementiser clauses (*CG*) and embedded questions (*KG*). Although, given these constraints, it is not clear how to measure recall in this task, we can measure precision by checking that for each constituent *CG* and *KG* there is a corresponding sentential node at deep syntax. However, aware of the criticisms directed at Parseval, we do not want to excessively punish annotations for having slightly different spans of clauses, so we define the proximity of a clause in shallow syntax to a sentential constituent

[3]The reason for this requirement is that we cannot expect shallow nominal, prepositional, etc., groups to contain sentential clauses.

in the deep syntax as the F-measure over the words they contain.[4] The final clausal precision of the shallow level is defined as the mean over all clauses.

5 Experiments and Evaluation

The measures defined above were applied to a 7600-sentence subcorpus annotated at both syntactic levels. For the whole corpus, the mean (micro-average) unlabelled precisions were: $P_s = 98.7\%$ and $P_d = 93.4\%$. This shows that, while the two levels of annotation are largely compatible, there are differences in the extents of some constituents. Also, the fact that $P_d < P_s$ shows that it is more common for the shallow level to miss (parts of) deep-level constituents, than the other way round.

We manually examined 50 sentences containing words on which the two annotations do not agree according to the unlabelled measures; there were 104 such word-level disagreements.

Discrepancies discovered this way may be divided into those 1. resulting from the insufficient subtlety of the measure, 2. reflecting controversial design decisions at the shallow level, 3. showing real differences, i.e., possible errors.

The biggest subset of class 1. results from the fact that not only syntactic groups are marked at the shallow level, but also some multi-token syntactic words, e.g., some adverbial groups resembling prepositional constructions. If such a syntactic word is the head of a group, a mismatch with the corresponding deep phrase is over-zealously reported. Around 35% of all differences belong to this group. Additionally, 16% of mismatches reflect differences in the treatment of adjectival participles. Hence, over 50% of reported differencies can be avoided by making the measures sensitive to such special cases.

Another 15% of differences, belonging to class 2., are caused by the controversial design decision to split larger coordinate structures at the shallow level into separate constituents, with only the final two conjuncts forming a coordinated group.

Finally, the remaining 1/3 of mismatches reflect real differences, often corresponding to errors at one of the levels. The most interesting subclass of these are discontinuities, currently handled only at the shallow level, e.g., cases of sentential conjunctions incorporated into NPs or discontinuous numeral phrases. Other differences include: some particles analysed as parts of NPs at one level, but not at the other, some adverbs or participles not analysed as adverbial groups at the shallow level, incorrect analysis of the highly ambiguous *to* as a noun (instead of a particle) at the deep level, etc.

Labelled measures have significantly lower values than the the unlabelled equivalents: $lP_s = 95.1\%$ and $lP_d = 91.1\%$. This is somewhat surprising, as at both levels constituents are marked for their lexical heads and it would seem that the morphosyntactic properties of the head should determine the label of the constituent. It turns out that the two main reasons for label mismatches are different approaches to some relative pronouns, and to some apparently prepositional constructions (analysed as adverbial at the shallow level).

Let us also note that the overall clausal precision of the shallow level is 0.996. Out of 691 sentences containing *CG* and *KG* groups, 670 match the deep level perfectly. In the remaining sentences, the usual problem is that *CG* or *KG* extends too far to the right (in 1 case it is too short), although in some cases it is the deep phrase that is too long or that is wrongly analysed, and in other cases two different spans reflect a genuine semantic ambiguity in the sentence.

6 Conclusion

It is not always easy to ascertain whether a mismatch between two syntactic annotation levels is a real error, but – on the basis of the manual examination of 50 sentences containing such mismatches – we estimate that between 12 and 15 of them contained errors at one or the other level. Since in the whole corpus 1882 non-matching (in the strong sense of unlabelled precision measures) sentences were found, this gives us the estimate of between 450 and 565 sentences containing real errors, thus complementing other methods currently used for Polish, which are estimated to find around 185 mismorfmed trees at the deep syntax level. Once these measures are made more subtle along the lines proposed above, the precision of such error reports should increase twofold from the current 20–30%, making human inspection of these reports worthwhile.

[4] Obviously, for any shallow-level clause we select a deep-level sentential constituent that maximises this F-measure.

References

Steven Abney. Parsing by chunks. In Robert Berwick, Steve Abney, and Carol Tenny, editors, *Principle-Based Parsing*, pages 257–278. Kluwer, 1991.

Ezra Black, Steve Abney, Dan Flickinger, Claudia Gdaniec, Ralph Grishman, Phil Harrison, Don Hindle, Robert Ingria, Frederick Jelinek, Judith L. Klavans, Mark Liberman, Mitchell P. Marcus, Salim Roukos, Beatrice Santorini, and Tomek Strzałkowski. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 306–311, Pacific Grove, 1991.

Adriane Boyd, Markus Dickinson, and Detmar Meurers. On detecting errors in dependency treebanks. *Research on Language and Computation*, 6(2):113–137, 2008. URL `http://jones.ling.indiana.edu/~mdickinson/papers/boyd-et-al-08.html`.

Ted Briscoe and John Carroll. A probabilistic LR parser of part-of-speech and punctuation labels. In Jenny Thomas and Mick Short, editors, *Using Corpora for Language Research, London*, pages 135–150. Longman, London, 1996.

Markus Dickinson and Chong Min Lee. Detecting errors in semantic annotation. In LREC. URL `{http://jones.ling.indiana.edu/~mdickinson/papers/dickinson-lee08.html}`.

Markus Dickinson and W. Detmar Meurers. Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 107–114, Budapest, 2003a.

Markus Dickinson and W. Detmar Meurers. Detecting inconsistencies in treebanks. In Joakim Nivre and Erhard Hinrichs, editors, *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pages 45–56, Växjö, Norway, 2003b. URL `http://jones.ling.indiana.edu/~mdickinson/papers/dickinson-meurers-tlt03.html`.

LREC. *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, 2008. ELRA.

Eleazar Eskin. Automatic corpus correction with anomaly detection. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, pages 148–153, Seattle, WA, 2000.

Yoshihide Kato and Shigeki Matsubara. Correcting errors in a treebank based on synchronous tree substitution grammar. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 74–79, Stroudsburg, PA, 2010. URL `http://dl.acm.org/citation.cfm?id=1858842.1858856`.

Sandra Kübler, Wolfgang Maier, Ines Rehbein, and Yannick Versley. How to compare treebanks. In LREC.

David M. Magerman. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 276–283, Cambridge, MA, 1995. doi: 10.3115/981658.981695. URL `http://www.aclweb.org/anthology/P95-1037`.

Carl Pollard and Ivan A. Sag. *Head-driven Phrase Structure Grammar*. Chicago University Press / CSLI Publications, Chicago, IL, 1994.

Adam Przepiórkowski, Rafał L. Górski, Marek Łaziński, and Piotr Pęzik. Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta, 2010. ELRA.

Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Cambridge, MA, 1995. ACL.

Ines Rehbein and Josef van Genabith. Treebank annotation schemes and parser evaluation for German. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 630–639, 2007. URL `http://www.aclweb.org/anthology/D/D07/D07-1066`.

Geoffrey Sampson. A proposal for improving

the measurement of parse accuracy. *International Journal of Corpus Linguistics*, 5:53–68, 2000. URL http://www.grsampson.net/APfi.html.

Geoffrey Sampson and Anna Babarczy. A test of the leaf-ancestor metric for parse accuracy. *Natural Language Engineering*, 9:365–380, 2003. URL http://www.grsampson.net/ATot.html.

Geoffrey Sampson, Robin Haigh, and Eric S. Atwell. Natural language analysis by stochastic optimization: a progress report on Project APRIL. *Journal of Experimental and Theoretical Artificial Intelligence*, 1:271–287, 1989.

Erik F. Tjong Kim Sang and Jorn Veenstra. Representing text chunks. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999)*, pages 173–179, Bergen, 1999.

Hans van Halteren. The detection of inconsistency in manually tagged text. In *Proceedings of the 2nd Workshop on Linguistically Interpreted Corpora (LINC 2000)*, 2000.

Marcin Woliński, Katarzyna Głowińska, and Marek Świdziński. A preliminary version of Składnica—a treebank of Polish. In Zygmunt Vetulani, editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 299–303, Poznań, Poland, 2011.