This is a repository copy of *The use of corpora for automatic evaluation of grammar inference systems*.

White Rose Research Online URL for this paper: http://eprints.whiterose.ac.uk/81222/

---

**Proceedings Paper:**

---

**Reuse**
Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher’s website.

**Takedown**
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.
The evaluation of grammar inference systems is clearly a non-trivial task, as it is possible to have more than one correct grammar for a given language. The 'looks good to me' approach, carried out by computational linguists analysing their own grammar inference system results, has prevailed for many years. This paper explores why this method has been so popular, in terms of its strengths, and also why it is no longer adequate as a reliable means to measuring performance. Corpus based methods, that can be performed automatically, are investigated to see how they can meet the needs of this difficult problem.
2 Looks good to me

The success of this approach is essentially its apparent simplicity. A system performs its GI procedures on a piece of unstructured text, and its resulting grammar can be analysed by a computational linguist, generally the computational linguist who built the Grammar Inference system under scrutiny. A qualitative evaluation takes place based on the linguistic intuitions of the evaluator, by highlighting features or structures in the learner output which look "good", or reminiscent of structures in a recognised linguistic theory. Of course, the apparent simplicity is due to the fact that a linguist possesses the required skills and experience to easily deduce whether the grammar in question contains a plausible structure. Researchers who came into the field of GI from a computing/AI/machine learning background are less likely to be as proficient at evaluating grammars using this approach.

‘Looks good to me’ is an important method of evaluation, and certainly should not be discredited just due to its lack of automation. Verification of a system carried out independently by one or more linguists would provide – in most cases – a more reliable measure of performance. It is also arguably the most resource efficient, since it can be evaluated on different languages without the need of structured corpora (van Zaanen 2001). Examples of this method in use can be found in. (Finch and Chater 1992; Losee 1996; Vervoort 2000; Henrichsen 2002).

However, the method has many disadvantages. Evaluation of this nature is mainly conducted by the developer of the system rather than someone independent to the work. Thus, there is a high chance of bias whereby the systems successes are highlighted and shortcomings are glossed over, making it almost impossible to gain an accurate picture of system performance. Even without any bias, the process is time consuming and would rely on the subjectivity of the expert(s), and may also be prone to unknown external factors that can affect humans. And finally, it does not offer the facility for comparing different systems which would be of great benefit.

3 Automatic evaluation

Ideally, the process of evaluation should be performed automatically which will save on time and on the necessity of an expert linguist in the chosen language. This section focuses on methods that harness corpora for this purpose.

3.1 ‘Gold standard’ treebank

This method is currently the most common to be adopted. It works by extracting the original natural language sentences from an existing treebank (the ‘gold standard’), and using it as input to a given GI system. The structured sentences produced by the GI system are then compared to the structure found in the original treebank. Common metrics include recall (measures the completeness of the learned grammar) and precision (measures the correctness of the learned grammar) that can be calculated to give an objective measure of system performance. This method has been used in (Brill 1993; Déjean 2000; van Zaanen 2001)

Although on the surface, this method appears sound, and simple to implement, there are in fact many issues that cause it to be insufficient for it to be reliable and accurate. Treebanks are expensive (both in time and money) to create, and so they are not in plentiful supply. The material within also tends to be exclusive to a particular domain, e.g., the ATIS treebank consists of sentences on questions and imperatives on air traffic, or the Penn Treebank is taken mostly from articles in the Wall Street Journal. If a GI system is not designed to learn from raw text of the same genre or subject,, then it is not an ideal candidate for a ‘gold standard’. The way in which the treebank has been structured poses a problem. Just because it has been labelled the ‘gold standard’ does not necessarily mean that anything that differs is wrong – it may simply be different, but equally correct. However, such differences will clearly affect the measure of the systems’ performance. It does also beg the question about whether certain treebanks are credible enough to be used for the purposes of evaluation. One must be confident that the quality and validity of an annotated is sufficiently high, otherwise, it is pointless evaluating systems with corpora that contain errors or are not well structured.
On a slightly more technical level, a large step to overcome is the likely discrepancy between the actual annotation schemes of the GI system to be evaluated and the ‘gold standard’. For example, an increasingly common grammar being employed within GI is the *categorial grammar* (Ajdukiewicz 1935). However, there is no corresponding treebank whose annotation formalism uses CG, which means the only useful measurements that can be acquired from this mismatch are metrics such as *crossing-brackets*, but so much information is wasted. Therefore, a system of translation is required to either convert the original treebank to the annotation scheme used by the GI system, or the other way around, in order for a more precise comparison to commence. This was the tactic taken by Watkinson and Manandhar (2001b) in their evaluation of CLL, whereby they set about establishing a ‘gold standard’ by translating the Penn Treebank annotation scheme into one using categorial grammar markup.

### 3.2 Multi-annotated corpora

To overcome some of the main disadvantages of the ‘gold standard’ approach, the next logical step is to develop a multi-annotated corpus, as exemplified in Atwell *et al.* (2000). The premise is that the same corpus is parsed by a variety of systems, rather than merely one as all common treebanks are. Providing all parses are trusted to be correct, then upon evaluation of a GI system, it is less likely to be penalised just because it produces different but equally correct parses, because it is being compared to a variety of such parses.

This step should make the evaluation fairer, however, it does add a new set of complexities. Not least due to the aforementioned annotation translation problem, which multiplies for each unique scheme used by the various parsers. Ideally, the same annotation would be used for all parses within the treebank – the ‘gold standard’ annotation scheme. But this too is fraught with difficulties, especially as it would be much simpler to accomplish a multi-annotated corpus by using off-the-shelf parsers.

Another issue is how to compare a given parsed sentence to be evaluated with a set of known correct parses. The best case is that a perfect match will be found. The worse case is no match. However, for many cases, it will be somewhere in between, where a number of partial matches are likely to exist. A best match approach seems logical, i.e., of the candidate parses, the one that is the most similar is the one that is then considered for the actual metric calculations.

### 3.3 Multi-corpora

In a similar vein to the above approach, the idea for using more than one corpus is primarily to avoid the domain specific issue that was addressed earlier (see section 3.1). Unlike other corpora, such as the Brown or LOB, which span many subjects in different contexts (newspaper articles, novels etc), well known treebanks fail to compare in size and variety. Therefore, this method aims to take advantage of smaller individual treebanks and combine into a single, large one. This would certainly make it fairer for general purpose learning systems, although perhaps there is less of a need for more focused systems.

Having said that, an alternative tactic is to treat each of the contributing corpora as individual and each as with the normal ‘gold standard’ method. You would then have a more detailed type of benchmark style scoring system, where not only is there an overall performance measure against all the corpora on test, it can also be seen on which type of corpora the GI system works best (or worse) on.

Of course, the most thorough approach would be if the smaller treebanks were also multi-annotated as described in the previous section. A ‘multi-multi-annotated corpus’ – if you like – would be a magnificent resource. This subtly shifts the aim of evaluation: the result or output is not a single overall “accuracy score”, but an exploration of a range of parsing schemes and genres to highlight similarities and differences between GI output and a variety of accepted linguistic analysis schemes. Ideally such an analysis could be facilitated by a parsed-corpus exploration toolkit: a tool to compare parses in GI output and linguist-annotated multi-multi-corpus, and automatically highlight differences.

### 4 Discussion
There does not seem to be a perfect solution to the problem of reliable evaluation, whether it be manual or automatic. GI and parsing is very subjective, and it is unlikely that a method will be created that will satisfy everyone. However, what should be unanimous is that a system of evaluation needs to exist. It may not be perfect, but at least a fair and consistent scheme can be created.

Unfortunately, it is clear that the most thorough automated evaluation approaches could be such a burden to developers to implement that they will opt for a simpler approach. Which is why it may be worthwhile to outsource the task of creating the ultimate ‘gold standard’ into a research project within its own right. The creation of a collection of multi-annotated corpora as the central resource, as well as a variety of translation interfaces for commonly used annotation schemes, that allow easy comparison between output of a GI system and the ‘gold standard’ corpus. It could be packaged like an evaluation toolkit – a black box that is publicly available, easily accessible, and simple to add on as the final phase within the pipeline of the GI system.

Alternatively, a multi-multi-annotated corpus could be achieved by getting the GI community – or at least a group of GI researchers – to work together, each contributing their preferred “target” analysis schemes. Sutcliffe et al (1996) describes an analogous evaluation exercise for a range of (non-ML) parsers: each researcher was asked to “bring along” their parser analyses of an agreed set of test sentences to a joint workshop, and then highlight and discuss successes (and failings) of their systems. Perhaps in a future GI workshop, GI researchers could be invited to bring along their preferred evaluation treebanks to challenge each other, and to each present both “looks good to me” and quantitative evaluations of systems for comparison.

5 Conclusion

The ‘looks good to me’ approach, despite the critical slant within this paper, is not an inferior one. However, it will not meet the demands for robust NLL evaluation. Which is why corpus based approaches have been presented as the most feasible and reliable way of essentially trying to emulate ‘looks good to me’ whilst eliminating bias and providing consistency.

There is a great need for researchers within the field to begin addressing accurate evaluation techniques. The current ‘gold standard’ method (as described in section 3.1) is too basic and will simply favour GI systems that behave similar to the way the ‘gold standard’ treebank was parsed. Greater flexibility is required, and a method like the multi-multi-annotated corpus (see section 3.3) will be a beneficial innovation to do just that. Instead of replacing “looks good to me”, the two approaches should be combined, so that evaluation is no longer a simple quest for a single “score”, but instead an exploration of strengths (and weaknesses) of GI systems.

References


Roberts A 2002 Automatic acquisition of word classification using distributional analysis of content words with respect to function words. Technical report, School of Computing, University of Leeds.


Watkinson S, Manandhar S 2001a A psychologically plausible and computationally effective approach to learning syntax. In CoNLL '01: the workshop on computational natural language learning, ACL/EACL.