Reducing Subjectivity of Natural Language Processing System Evaluation

Menno van Zaanen
ILK, Tilburg University
P.O. Box 90153,
5000 LE Tilburg,
The Netherlands
mvzaanen@uvt.nl
Tel: +31 13 466 8260,
Fax: +31 13 466 3110

Robert John Freeman
Chaotic Language
9 Ocean View Terrace,
Christchurch 8008,
New Zealand
rob@chaoticlanguage.com
Tel: +64 3 326 6406

Abstract

In this article we investigate problems of the current means of evaluation of natural language systems. We find that apart from the practical problems, there is a more fundamental problem: the evaluation standards we measure against may not be objectively defined. In a sense, the very evaluation problems we set ourselves may not be well posed. We speculate on reasons for this, on ways to contain it, on evaluation standards which may more accurately reflect the underlying nature of language, and indeed on the appropriateness of a narrow focus on evaluation alone at our current stage of understanding the language process.

1 Introduction

Recently, there has been an increasing interest in the evaluation of various types of natural language processing (NLP) systems, e.g. (Carroll et al., 1998a; Resnik and Yarowsky, 1999; Daelemans and Hoste, 2002). There may be many reasons for this. One reason may be that researches find that the performance of the current state-of-the-art NLP systems is hard to compare. On the one hand, the results of systems can be quite similar (and very close together), but on the other hand, results obtained on different datasets may vary greatly and the metrics used do not generate consistent results when parameters are changed.

In this article we will speculate that all these difficulties stem from a fundamental misconception which underlies all current evaluation methods.

The underlying assumption of the current evaluation methods is that there is one (and only one) correct analysis of the linguistic data. This analysis can be, for example, POS tags, parses of sentences in the form of tree structures or dependency relations or some other annotation. The correct analysis is called the *gold standard* and the results of the system under evaluation are compared against it.

What we suggest in this article is that, in contrast to this assumption, there is no single correct analysis in many NLP tasks. People assign different analyses to linguistic data. This may have several reasons, but a discussion on this is beyond the scope of this article. We explore the evidence and suggest measures which may be taken to contain the resulting ambiguity of abstract structural characterizations of language. We can also consider the broader implications of this for language theory.

The broad line is that grammatical characterizations are intermediate representations of language function, arbitrary in some way, and interpreted according to arbitrary bias of human observers. Then any evaluation should be performed with an understanding of this leniency or flexibility on the part of humans. Only then can we get a measure of how well a given NLP system could be expected to perform in human terms.

What perhaps most distinguishes our ideas from others that consider flaws in gold standard linguis-

tic evaluations e.g. (Kilgarriff, 1998) is the extent to which we embrace that subjectivity. Language, and perhaps all cognitive functions, we suggest, are inherently subjective activities. When we study language we try to model a subjective process. Perhaps the subjectivity of language should be the very core of what we seek to understand, model, and evaluate.

We will start with a description of the evaluation methods that are currently in use. Each of these methods has certain problems. Next, our conception of the fundamental problem with current, structurally based, evaluation methods will be presented. Finally, we make suggestions for alternative approaches, essentially seeking to parameterize, reduce, or by-pass the influence of subjective evaluation criteria.

2 Overview of Existing Evaluation Methods

Current evaluation methods can be subdivided into groups in many ways, depending on the point of view. Here, we will divide the methods in two large classes based on the point in time when the reference standard is set.[1] In the case of the looks-good-to-me approaches, the reference is set after the system has been applied to the data, whereas with blind testing, the gold standard is set beforehand.[2]

2.1 Looks-good-to-me

When the looks-good-to-me evaluation approach is taken, the NLP system is applied to unstructured data. The results of the system are then given to (human) evaluators. These evaluators decide whether the assigned structure is correct.

In this paradigm, the output of the NLP system is judged by evaluators when processing is complete. This makes it possible to specifically select (unstructured) input data that will test whether certain analyses are actually generated by the NLP system. The input data can be tuned to research certain abilities of the system. For example, if you are interested in how a system handles PP attachment, you can use input data that mainly contains sentences with PP attachment problems.

Another advantage is that only unstructured data is needed for evaluation. The structured output of the system is analyzed by experts, the evaluators. This means that it is possible to evaluate NLP systems on data for which little or no annotation is available. This includes the evaluation using data from, for example, minority languages, but also the evaluation of systems providing new annotation schemes.

The main disadvantage of this approach is that in practice, often only one evaluator is used to classify the output of the NLP system. This, of course, makes the final result depend greatly on the quality of the evaluator. To make matters worse, the evaluator is often the designer of the system being assessed. The reason for this is that evaluators are often expensive and because each repetition of an evaluation like this again requires evaluators. Overall, the method is expensive in time and resources.

This means of evaluation, however, is accepted standard practice in certain sub-fields of NLP. Mainly in the field of machine translation, where multiple translations are equally valid, this is the case. Note that for a correct evaluation multiple independent evaluators should be used. (Elliott et al., 2003)

2.2 Blind testing

With blind testing, the data used to testing is structured beforehand. An unstructured version of the data is handed to the NLP system, and the output of the system is compared against the structured version. The work of the evaluators in the looks-good-to-me approach has been moved from the moment following the application of the system on the data to before the system is evaluated on the data.

This approach has several advantages. The setup of blind testing requires structured data to be created beforehand. This structuring only needs to be done once, reducing the costs of evaluation (with respect to the looks-good-to-me approach). The data can be reused for different evaluations and it can also be made publically available and be used to evaluate other, similar NLP systems. This

[1] Both approaches assume a gold standard.

[2] A similar division was made in (van Zaanen et al., 2004). That article discussed the evaluation of grammatical inference systems in particular.

again, can result in de facto standard test datasets, which allow researchers to examine whether their systems perform better than previous ones.

A major difference between the looks-good-to-me and blind testing approaches is that with the blind testing method the evaluators have to make their own decisions on how to structure the data, instead of deciding whether a structuring is correct yes or no as is the case with the looks-good-to-me approach. Because the evaluators have to make the decisions beforehand, the objectivity of the evaluation increases. Sometimes there are several ways to structure the data, but one structuring is preferred. The evaluators need to select this structure beforehand with blind testing, but they might be tempted to accept a less preferred structuring with the looks-good-to-me method.

3 General Problems of Current Evaluation Methods

In this section, we will look at some of the problems of current evaluation methods. We think that these are all related to an underlying problem which, as far as we know, has not been recognized as such.[3] By discussing the observed problems we hope that the nature of the underlying problem will become clear.

3.1 Metrics

So far, we have talked about evaluation in a very generic way. Of course, comparison of the output of a system is performed using metrics. It is the metrics which give the actual figure describing the quality of the output of the system.

Within the field of NLP, many different metrics have been proposed and used. For example, if we look at parsing, the PARSEVAL metrics (Black et al., 1991) are well known. However, these metrics have certain disadvantages. This has resulted in other work that tries to eliminate or at least reduce these problems (Carroll et al., 1998a; Carroll et al., 1998b; Sampson, 2000).[4]

[3]Or if it has been recognized then it has not been understood in the same way, i.e. as a solution rather than a problem, cf. (Kilgarriff, 1998).

[4]Similar metrics have been used in the field of grammatical inference. These metrics also create problems. (van Zaanen et al., 2004)

Gold standard	John	sees	(	the	man	)
Matching	John	sees	(	the	man	)
Non-crossing	John	(	sees	the	man	)
Crossing	(	John	sees	the	)	man

Figure 1: Correct and incorrect brackets

Most evaluation metrics are quite strict. If something is wrong, the final score goes down. Most of the time, this is correct; if the output of the system is worse than the output of another system, the score should be lower.

However, often humans do not seem to agree (completely) on certain structures in the data. If a system generates another valid structure that is different from the structure in the gold standard, it is penalized even though the output is correct. The main point here is that "correct" is defined in terms of the gold standard.

There exist metrics that are not as strict. For example, non-crossing brackets in the field of parsing. The idea is to base the final results not on the number of correct pieces of structure, but on pieces that are not incorrect. Unfortunately, this does not solve the problem. Non-crossing brackets metrics do not allow completely incorrect structures, but they do allow certain structures that would not considered correct according to humans.

To illustrate matching and crossing brackets, see figure 1. The first structure should be considered the gold standard. The second structure contains a pair of brackets that are matching and therefore are also non-crossing. The third structure contains a pair of brackets that is not entirely correct (i.e. the opening bracket doesn't match any of the opening brackets in the gold standard), but it is not crossing any pairs of brackets of the gold standard. It is therefore incorrect with respect to the matching metrics, but correct with respect to the non-crossing brackets metrics. The final structure has a pair of brackets that overlaps with the pair of brackets in the gold standard. This is incorrect even with non-crossing brackets metrics.

Summarizing, using a strict evaluation metric will measure the performance of a system with respect to the gold standard. When "loose" metrics are used, such as the non-crossing brackets, the re-

sults allow for partially incorrect structures to be counted as correct. The idea behind this is that the output of an NLP system does not necessarily conform to the gold standard completely. In fact, allowing less strict metrics, such as the non-crossing metrics, is equivalent to loosening up the strict definitions of the gold standard.

What is most interesting about this, however, is that the need for such "loosening" can be taken as tacit admission not all details of the gold standard are significant, that the gold standard is, to an extent, random.

3.2 Over-training

It is common practice to divide a dataset in three parts. One part is used to train a system. A different part is used for tuning and the final part is used to test the final system. By comparing results obtained on the tuning and test set, it is possible to get an idea of the amount of over-training of the system. If the results of the tuning data are much better than that of the testing data (which is taken from the same original dataset and so it should be similar in many respects), the system has been over-trained on the training (and tuning) data. It indicates that the system will not generalize well over new data.

However, even if the results of the tuning and test sets do not show this, over-training will still occur. This becomes apparent when the (tuned) system is applied to a different data set. Different in this context can be, for example, a different language domain or a different annotation scheme. This influence can be quite large, as is shown in (Entwisle and Powers, 1998).

In effect, an NLP system should find the right bias between fully general and fully specific. A fully general system is not useful, because everything is possible. The system will not get guidance for specific analyses and will not be useful at all. A fully specific system, on the other hand, can only analyze a limited amount of data and only in a fixed way. The analyzes are completely defined (in a fixed way) by the training data. No generalization is done.

One traditional goal when designing an NLP system is to find the right amount of generalization. Analyzing and evaluation on similar tuning and testing data may give information whether the right amount of generalization is found on that specific dataset, but it cannot show the effectiveness of the generalization on language in general.

Once again, however, it is not the ways we optimize this tension between generality and specificity to minimize over-training between datasets which is most interesting. What is most interesting is what the need to perform such optimization tells us about our goals with respect to a single gold standard. What is it which makes us think a single set of generalizations is going to capture all relevant generalizations about all possible sub-sets of data? Is the idea of training to a single gold standard sensible? The persistence of the phenomenon of over-training between different subsets of language suggests it is not.

3.3 Inadequately defined goals

Historically it has proved convenient to describe language in terms of grammars: word classes and rules for the combination of such classes. However, it is unclear what evidence we have for the objective existence of such grammars or other sets of linguistic classes.

(Dagan et al., 1993) described the problem nicely:

> It has been traditionally assumed that semantic information about words should be generalized using word classes. In systems which rely on manual encoding of knowledge, this assumption seems necessary to cope with the high complexity of lexical relationships. However, it was never clearly shown that unrestricted language is indeed structured in accordance with this assumption. Moreover, the high variability in lexical cooccurrence data suggests that rather few generalizations can be performed on safe grounds.

Other researchers have examined the problem from another direction, in a proliferation of annotation and evaluation standards (Atwell et al., 2000). If indeed "few generalizations can be performed on safe grounds", this proliferation of standards may reflect a fundamental reality that the

evaluation problem, as currently conceived, is not well posed.

The impossibility of achieving a perfect evaluation score for tasks such as tagging has also been directly considered before. For instance (Church, 1992) presented results which indicated 97% is an upper bound for inter-tagger agreement, even after negotiation.

Now, other researchers, notably (Voutilainen, 1999), have disputed the universality of this claim for carefully selected annotation schemes. Voutilainen claims 100% agreement to three decimal places (27 unresolvable disputes over 55724 words) for annotation using his EngCG-2 morphological tags.

Voutilainen's results may reflect a particularly distinct tagset (and it remains to be seen whether such results demand a choice between grammatical characterizations which are distinct, and those which are informative), but even this is not enough to result in a completely objective annotation.

Similar disagreement is observed for other grammatical tasks. For instance there is an observed disagreement in human evaluations for basic structural tasks like word segmentation for languages which are not traditionally segmented, like Chinese (Fung and Wu, 1994).

And while deciding that the idea of a gold standard is not "fool's gold", (Kilgarriff, 1998) admits that consensus as high as that claimed by Voutilainen is an unrealistic goal for the related task of word sense disambiguation, where he cites where inter-tagger agreement (ITA) norms hovering around 70-80%.

Whatever we believe about linguistic structures such as grammars, surely the least that can be said is that we do not have any direct access to them. This is reflected in the wide variety of grammar formalisms available. Grammars only exist by virtue of subjective evaluations made by human observers, and these observations palpably vary.

3.4 Human ambiguity

As we have seen before, humans do not always agree on annotations of linguistic data. There may be several reasons for this. Likely some of these reasons relate to random factors, things unrelated to the nature of the language system. A human "bias" if you will.

Conceivably it is just our observations which vary, and the underlying system we seek to describe is objectively defined. Perhaps it is possible to define a linguistic evaluation task, even a grammatical evaluation task, which is beyond the subjectivity of human perception. However, language is a quintessentially human activity. It is hard to take humans out of the loop, and humans are conceivably always subjective in their opinions. Indeed, perhaps subjectivity is the true essence of cognition, and what we really should be trying to model, not any particular structural abstractions which make up the content of our subjective evaluations.

4 Problem of Subjectivity

We have discussed problems in current methods of evaluation, speculated on objective reality of grammatical perceptions. Finally, we have introduced the idea that the subjectivity of human perception, of which language is our most essential expression, might be the very thing we are trying to describe when we model language.

Our suggestion, then, for the underlying difficulty with all current evaluation methods, is just that they all assume a single, universal, and importantly a global (to the language), gold standard. We suggest that, on the contrary, all human evaluation criteria (and it seems likely that an evaluation of language must be at least human-like) are subjective, that this should be at the core of what we seek to evaluate.

This subjectivity is apparent in evaluation metrics, in the "over-training" which is evident from sub-corpus to sub-corpus in trained systems, in the very annotation schemes themselves, and indeed it seems to be present in all systems where human judgment is required.

There is subjectivity from person to person, but perhaps it is the subjectivity from sub-corpus to sub-corpus, and sentence-to-sentence (leading to under-specification and thus ambiguity in any global labeling scheme) which is most interesting (though it is not obvious that it is possible to distinguish the two in the final analysis).

To discuss the possible causes and nature of the subjectivity we observe in evaluation, and evalu-

ation standards, goes beyond the scope of this article. We wish to limit ourselves here, firstly, to noting the fact, and discussing its implications for evaluation.

5 Evaluation of Subjectivity

In this section we will try to return to the concrete again, and suggest two possible approaches to handling the subjectivity which seems to be present in the evaluation of NLP systems. The first approach is to directly address this subjectivity, to make it the focus of evaluation rather than an impediment to evaluation. The second approach is to speculate on the possibility of evaluation with respect to linguistic tasks which are beyond subjectivity. Though the existence of such remains to be seen.

5.1 Modeling subjectivity in objective evaluation standards

Clearly we do not want an evaluation criterion which leaves us with nothing but a subjective evaluation (which it seems our systems of evaluation, with their subjectively defined, global, gold standards, do today by default). Arguably what we need is a system which gives us an (objective) evaluation of the subjectivity which seems to be an inherent part of our evaluations and evaluation standards.

Perhaps we can accept subjectivity in evaluation standards, but tame it to some extent by embracing it rather than ignoring it. The subjectivity of an observation then becomes part of the evaluation rather than a bug in the evaluation problem. By embracing the problem we can try to contain it. This is better than the alternative of allowing it to affect our results randomly, as it must if we simply ignore it.

To do this, we probably require a model for the subjectivity of a grammar. When this model is incorporated in the evaluation procedure, we will not, strictly speaking, be evaluating the same things any more, there is a new variable to be evaluated.

For instance, we could require of our models that they produce not only a grammatical analysis, but also a confidence value for that analysis. Many data-based models, such as stochastic context-free grammars and the like, are of exactly the type to produce such a confidence value.

Another option would be to evaluate a whole range of observation standards, a whole range of treebanks for a parser evaluation, for instance. The variation in agreement between the different standards would provide an automatic measure of confidence to be assigned to different evaluation values. Essentially, we would be filtering out the values against which we should be willing to accept a variation of opinion, finding the values which a model should predict to be subjective, but finding it independently of the model.

Models which model subjectivity accurately should perform better against such a proliferation of standards because they should perform more reliably on the decisions about which reliable evaluations can be made.

5.2 Task-oriented evaluation

Another way of handling the subjectivity in conventional evaluation would be to attempt something rather more like the classic Turing test, where we by-pass the idea of abstract structure entirely and attempt to measure the effectiveness of a system directly in terms of concrete language processing tasks.

In this context, task-oriented evaluation means evaluation with respect to something, a task, about which humans can reach some degree of agreement. More agreement than they reach over the task of grammatical annotation, in any case.

Indeed, systems that generate any kind of grammatical annotation might be specifically excluded, since there is no real objective evidence that a grammar (in the form of the current computational grammars) exists. People may argue over the form of the grammar (or the analysis) without even having to agree or disagree on the actual outcome of the system. Likely we would need to concentrate on other tasks. Tasks for which the format of the outcome is unarguably correct. Among candidates might be translation, information retrieval, or summarization.

However, it remains to be seen whether such objectively verifiable tasks can be found. Quite possibly all language processing tasks (and perhaps all cognitive tasks) are essentially equivocal. Pos-

sibly the subjectivity of grammar is not an exception, but is general to any linguistic task.

6 Future Work

The problem of the current ways of evaluation, we suggest, is that they do not take into account the subjectivity which seems to be an integral part of every linguistic evaluation decision.

We recognize two major directions for future work. The first is to continue to look at ways of containing the subjectivity of current grammatical evaluations.

One possibility is to incorporate subjectivity (in an objective evaluation) by adjusting the metrics used. Something similar has been done using the kappa metric (Cohen, 1960; Siegel and Castellan Jr., 1988), that tries to reduce the influence of the data by taking the complexity of the data into account. Similarly, this measure is used to compute inter-annotator agreement. (Kilgarriff, 1998)

Another way is to evaluate on multiple, complete different datasets. Initial work on this has already been done. For example, (Roberts and Atwell, 2003) argues that evaluation on parallel-parsed corpora result in more stable results. Similarly, (van Zaanen et al., 2004) describe an extension of this work for evaluation on grammatical inference systems.

The second major direction is to look beyond a narrow focus on evaluation. Perhaps it is fair to argue that the realization that language is inherently subjective should lead us to rethink our current models of language form and function in their entirety. Assessment can be a valuable tool for the improvement of technology, but in the "tallest tree" sense it does not always lead to the best solution. Perhaps what is most appropriate now in language processing is not a greater focus on assessment, but a broader consideration of what might be the underlying nature of language process.

One direction to explore would be to find a model which recognizes the importance of subjectivity (i.e. the grammatical subjectivity on sub-text observed in over-training) in the analysis and annotation of data. Current evaluation criteria are assumed to be global, gold standards, which do not take into account variation from sub-corpus to sub-corpus. Possibly the constraints of generality are what force our global evaluation criteria to be under-specified, and being under-specified, subject to random bias (and individually subjective judgments) on the part of the human annotators who abstract them.

In any case, a "natural" evaluation standard almost certainly requires a model of what language actually does, which is different from the conventional. More research needs to go into finding real objective evaluation problems (if they exist).

So while it makes sense to examine ways we could contain the subjectivity of evaluation against grammatical standards it is also worth keeping in mind that an important direction for future work is to study the nature of the language models which will predict subjectivity, and against which the new subjectivity containing evaluation measures might be compared.

7 Conclusion

In this article we have shown that the current approaches to evaluation of NLP systems are problematic. We think the underlying problem is that language is inherently subjective from person to person and most importantly from linguistic unit to linguistic unit. Current evaluation methods do not recognize this inherent subjectivity; evaluation is done against a fixed gold standard.

We have speculated on possible reasons for this, and suggested ways to tame this subjectivity and obtain a truly objective evaluation criterion for NLP systems.

This includes extending the current evaluation methods with different metrics, but also including different datasets, reducing the influence of the annotation scheme. Completely different evaluation methods, such as the task-oriented approach will need to be implemented and tested as well.

Underlying all these attempts to reduce the influence of subjective assessments on NLP evaluation, however, there is a broader issue. Perhaps it is time to take a step back from the recent focus on evaluation and re-examine our goals and assumptions when studying language. When do we consider a NLP system successful? The current focus on narrow structural abstraction (in defiance of notable threads of pure linguistic theory) seems to be showing its limitations. When we understand

why our assessments of structure are subjective we should be able to contain that subjectivity, and in the process of understanding we may be able to identify more meaningful evaluation goals.

References

D. Archer, P. Rayson, A. Wilson, and T. McEnery, editors. 2003. *Proceedings of the Corpus Linguistics 2003 conference; Lancaster, UK*.

E. Atwell, G. Demetriou, J. Hughes, A. Schiffrin, C. Souter, and S. Wilcock. 2000. A comparative evaluation of modern english corpus grammatical annotation schemes. *ICAME Journal, International Computer Archive of Modern and medieval English*, 24:7–23.

E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of a Workshop—Speech and Natural Language*, pages 306–311, February 19–22.

John Carroll, Ted Briscoe, and Antonio Sanfilippo. 1998a. Parser evaluation: A survey and a new proposal. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 447–454.

John Carroll, Guido Minnen, and Ted Briscoe. 1998b. Corpus annotation for parser evaluation. In *Proceedings of the EACL-99 Post-Conference Workshop on Linguistically Interpreted Corpora; Bergen, Norway*, pages 35–41. Association for Computational Linguistics (ACL).

K. Church. 1992. Current practice in part of speech tagging and suggestions for the future. In Simmons, editor, *Sbornik praci: In Honor of Henry Kučera*. Michigan Slavic Studies.

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

Walter Daelemans and Véronique Hoste. 2002. Evaluation of machine learning methods for natural language processing tasks. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002); Las Palmas, Gran Canaria*, pages 755–760.

I. Dagan, S. Marcus, and S. Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (ACL); Columbus:OH, USA*, pages 164–171. Association for Computational Linguistics (ACL), June 22–26.

Debbie Elliott, Anthony Hartley, and Eric Atwell. 2003. Rationale for a multilingual aligned corpus for machine translation evaluation. In Archer et al. (Archer et al., 2003), pages 191–200.

Jim Entwisle and David Powers. 1998. The present use of statistics in the evaluation of NLP parsers. In D. M. W. Powers, editor, *New Methods in Language Processing and Computational Natural Language Learning; Sydney, Australia*, pages 215–224, January 22–24.

Pascale Fung and Dekai Wu. 1994. Statistical augmentation of a chinese machine-readable dictionary. In *Proceedings of the Second Annual Workshop on Very Large Corpora; Kyoto, Japan*, pages 69–85.

Adam Kilgarriff. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language*, 12(4):453–472.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Andrew Roberts and Eric Atwell. 2003. The use of corpora for automatic evaluation of grammar inference systems. In Archer et al. (Archer et al., 2003), pages 657–661.

Geoffrey Sampson. 2000. A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics*, 5(1):53–68.

S. Siegel and N. J. Castellan Jr. 1988. *Nonparametric Statistics for the Behavioural Sciences*. McGraw-Hill, New York:NY, USA.

Menno van Zaanen, Andrew Roberts, and Eric Atwell. 2004. A multilingual parallel parsed corpus as gold standard for grammatical inference evaluation. In Lambros Kranias, Nicoletta Calzolari, Gregor Thurmair, Yorick Wilks, Eduard Hovy, Gudrun Magnusdottir, Anna Samiotou, and Khalid Choukri, editors, *Proceedings of the Workshop: The Amazing Utility of Parallel and Comparable Corpora; Lisbon, Portugal*, pages 58–61, May.

Atro Voutilainen. 1999. An experiment on the upper bound of interjudge agreement: The case of tagging. In *Proceedings of the 9th Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL); Bergen, Norway*, pages 204–208. European Chapter of the Association for Computational Linguistics (EACL).