All that glitters …
Interannotator agreement in natural language processing

Lars Borin
University of Gothenburg

Abstract

Evaluation has emerged as a central concern in natural language processing (NLP) over the last few decades. Evaluation is done against a *gold standard*, a manually linguistically annotated dataset, which is assumed to provide the ground truth against which the accuracy of the NLP system can be assessed automatically. In this article, some methodological questions in connection with the creation of gold standard datasets are discussed, in particular (non-)expectations of linguistic expertise in annotators and the interannotator agreement measure standardly but unreflectedly used as a kind of quality index of NLP gold standards.

Keywords: Evaluation, natural language processing, interannotator agreement, annotation

1. Introduction

Church and Hestness (2019) present "[a] survey of 25 years of evaluation" in speech and language processing.[1] The *evaluation* that they refer to – and this is how this notion is most commonly understood in present-day NLP – involves the following two elements:[2] (1) an NLP system/application for automatically annotating text data for some linguistic features (part of speech, syntactic structure, sentiment polarity, etc.); and (2) a set of text data manually or semi-manually annotated for the same linguistic features, which has not been used in the development of the NLP system to be evaluated. Such a dataset is referred to as a *gold standard* in the literature.

The title of Church and Hestness (2019) refers to a development that the field of NLP has undergone over the last three decades. Trond Trosterud and I both started out in NLP in the previous millennium, and we also have in common an academic background in linguistics rather than in computer science, which has informed the professional trajectories of both of us in the field. At that time, linguistics played a larger role in NLP than at present and rule-based ("symbolic") solutions explicitly realizing linguistic formalizations formed the bulk of NLP systems.

Much has changed since then, and the field looks very different now. Over the last two decades or so, NLP has become increasingly disassociated from the concerns of linguistics (see, e.g. Reiter 2007, Wintner 2009, Manning 2015), and at present, data-driven machine learning approaches hold sway in our field, in particular so-called deep learning (Manning 2015). This development has happened in parallel with the increasing emphasis on evaluation, with noticeable effects on the conception of how gold standard preparation should be carried out.

The same gold standard datasets that are employed for evaluation are also frequently used in order to build NLP systems, in particular systems based on machine learning, such that part of the data is used for training, another part for testing and yet another part for evaluation.[3] Here, we are concerned only with evaluation, however, since this is something that in fact impacts all kinds of NLP systems, regardless of their underlying architecture. In this way the introduction of systematic evaluation of systems is a more profound change in NLP than the "non-symbolic revolution" whereby rule-based applications have been replaced by systems based on machine learning. Thus, the rule-based systems described by Antonsen et al. (2010) and Harrigan et al. (2017) are also formally evaluated against gold standards.

[1] Here I will be concerned only with computational processing of text (not speech processing), a field which goes under several names: *computational linguistics* (CL), *natural language processing* (NLP), *language technology* (LT), and *(natural) language engineering* ([N]LE). In this article, I will refer to it as "NLP". Note that in popular media, much of what is today referred to as "artificial intelligence" (AI) or even "algorithms" is in fact NLP.

[2] In this article, I will discuss only "intrinsic" evaluation, where the output of an NLP system is evaluated directly against an annotated dataset, and not the "extrinsic" kind where such a system is evaluated for its contribution to solving some external, "downstream", task, e.g. a lemmatizer used as a component in an information-retrieval application. For lack of space, I will also need to forgo human evaluation of NLP-system output, which presents plenty of interesting methodological challenges of its own (cf. Hämäläinen and Alnajjar 2021). Further, my interest here is in annotations conforming to best-practice linguistic analysis, rather than to, e.g., popular (mis)conceptions of language. The latter is of course a perfectly valid and interesting research topic, but not the one in focus here.

[3] This presupposes very large datasets, however, and there are many gold standards which can only be used for evaluation because of their size.

©2022 Lars Borin. *Nordlyd* 46.1: 19-26, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6348

Since evaluation is *de rigueur* in today's NLP, the quality of gold standard datasets should be a high priority in the NLP community. Regrettably, much remains to be desired in this regard, and arguably, this is largely due to how the annotation of gold-standard datasets is carried out. This was pointed out already a decade and a half ago by Zaenen (2006), who specifically pointed to the lack of linguistic analysis in NLP annotation practices. Evaluation has taken center stage in NLP since then, but the development of machine learning has partly redefined the role of linguistic analysis in this context, as we will see below. We should also not forget that datasets with even quite basic linguistic annotation are still lacking for the vast majority of the world's some 7,000 languages (Trosterud 2006; 2012), making linguistic annotation a high-priority, long-term concern.

The point about annotation quality raised above was brought home to me serendipitously about ten years ago, when I happened to be present at a presentation by Anthony Kroch, a historical linguist at the University of Pennsylvania, who had been a pioneer in using diachronic treebanks for studying English syntactic change. He had been involved in the compilation of several such treebanks using the phrase structure formalism of the Penn Treebank (PTB; Marcus et al. 1993). However, he also mentioned in passing that (approximately cited from my memory), much as he would have liked to use PTB as representing Present-Day English, in his view the annotation quality of the corpus was not good enough for his purposes. Relevant in this context is that PTB at that time was widely considered to be *the* gold standard treebank for English NLP, against which in particular phrase-structure parser accuracy was routinely measured (cf. Zaenen 2006). Kroch's picture is confirmed by a number of publications describing efforts to develop automatic or semi-automatic methods for finding annotation errors in corpora (e.g. Dickinson and Meurers 2003), although it is somewhat unfair to single out PTB in this way, when the truth is that it was used in these experiments primarily because of its wide availability and meticulous documentation, not because its annotation quality was believed to be inferior to that of other treebanks.

2. Annotation for gold standard NLP data

The fundamental assumption licensing the use of gold standards for evaluation in NLP is of course that the annotations in the gold standard are correct – that they constitute the *ground truth* in the domain in question – or at least that their degree of correctness is known.

This is where the – often invoked but frequently misunderstood – measure of *interannotator agreement* enters the picture.

2.1. Interannotator agreement in NLP

Interannotator agreement (IAA or ITA)[4] is a measure of annotation reliability (see Artstein and Poesio 2008). The measurement of IAA as originally formulated presupposes that certain preconditions are fulfilled (Artstein and Poesio 2008:574):

- There is more than one annotator
- There must be a detailed and clear annotation manual
- There must be clear explicit criteria for selecting annotators
- The annotators must work independently of each other

In NLP, much effort has been spent on devising fair measures of IAA. Most accounts of gold standard annotation found in the literature do not discuss annotation quality or accuracy. Instead, IAA is reported without comment as if it were such a measure.[5] That this is a mistaken belief is easily seen if we imagine a thought experiment where naive annotators are supplied with an explicit and clear, but false annotation manual.

IAA as practiced in NLP was originally developed for content coding (see Carletta 1996), and because of this annotation is seen as analogous to conducting a scientific experiment, with concomitant requirements of replicability, etc., hence the insistence on the preconditions listed above (Artstein and Poesio 2008:574). Notably, the use of expert annotators is sometimes explicitly eschewed, or at least frowned upon, since experts may make annotation decisions

[4] Also called *inter-coder agreement*.

[5] A notable exception to this is the first widely used NLP gold standard: The relationship of IAA to the quality of the part of speech annotation produced by the Penn Treebank annotators was estimated using the POS-tagged version of the Brown Corpus (Kucera and Francis 1967) as a gold standard (Marcus et al. 1993:318–320).

based on their domain expertise rather than on anything written in the annotation manual, thereby potentially compromising reproducibility. However, Artstein and Poesio (2008) and Artstein (2017) do note that annotation for NLP purposes often deviates from this methodology.

Some kinds of linguistic annotation are indeed similar to content coding (see further below), but "low-level" linguistic annotation such as for parts of speech, syntactic structure, discourse segments, coreference, and word senses, are arguably not among them. Instead, this kind of annotation is more akin to e.g. medical diagnosis, i.e., the remit of experts – highly trained professionals with long practical experience – rather than an activity which laypersons can engage in successfully on the basis of the contents of even very detailed annotation guidelines.

Importantly in this context, it is generally agreed that a key component of expertise is *intuition* (Hetmański 2018), a catch-all label summarizing a kind of Gestalt knowledge formed by long formal training and extensive practical experience. Thus, it may not be possible to have explicit and exhaustive annotation guidelines instead of expert annotators. In fact, even when employing non-experts for linguistic annotation, it seems to me that a lot of background knowledge about language description is assumed ("school grammar"), and not explicitly provided in an annotation manual.

2.2. Interannotator agreement and annotation quality

Given that gold standards have such pride of place in present-day NLP, it is somewhat surprising that studies into the methodology of gold standard compilation are few and far between, the exception being works dedicated to the formal aspects of interannotator agreement (see Artstein and Poesio 2008, Artstein 2017, and references provided there). Similarly, since practical experience teaches us that failure to report IAA may be seen as sufficient grounds for rejection of conference papers, we would expect more and more varied investigations of how IAA relates to gold annotation quality than we actually see in the literature.

There are some studies – not of annotation quality directly – but of how various characteristics of annotation tasks influence IAA. Bayerl and Paul (2011) present a meta-study where they attempt to tease out contributions to IAA from factors such as *domain*, *complexity of annotation scheme*, *language* (of the text), *number of annotators*, and notably *annotator training* and *domain expertise*. There are also some studies which have looked specifically at differences between non-expert and expert annotators (e.g. Snow et al. 2008, Gillick and Liu 2010, Munro et al. 2010, Plank et al. 2014), as well as some studies of other factors influencing annotation quality (e.g. Babarczy et al. 2006, Sampson and Babarczy 2008, Brown et al. 2010).

For the factors annotator training and domain expertise, the meta-study by Bayerl and Paul (2011) is hampered by an easily observable fact about accounts of gold standard annotation efforts, viz. that the relevant background of annotators is rarely specified and also subject to a "Chinese whispers/Telephone" game effect when cited in other sources. For example, Snow et al. (2008) refer to five kinds of annotations made by "experts", without clarifying the credentials of these experts, however. The publications cited by Snow et al. (2008) where these datasets are presented describe their annotators as "human annotators" (Dagan et al. 2006), "[undergraduate] students at the State University of New York at Oswego [...] native speakers of English" (Miller and Charles 1991), "linguistics students" (Pradhan et al. 2007), "Double blind annotation by two linguistically trained annotators was performed on corpus instances, with a third linguist adjudicating between inter-annotator differences" (Palmer et al. 2004:51), "annotators" (Strapparava and Mihalcea 2007), and "The initial stage was carried out by 5 annotators of remarkably different profiles with regards to their linguistic background. All of them however had participated in the development of the TimeML annotation scheme. The group of annotators for the second stage comprised 45 computer science undergraduate and graduate students" (Pustejovsky et al. 2003:652). Similarly, Hovy et al. (2014) refer to the annotators of the dataset described by Gimpel et al. (2011) as "experts" and "professional", while their original characterization by Gimpel et al. (2011) is simply as "researchers", where the inference is that these are researchers in a computer science department. They may of course have linguistic training, but we are not given any information about this.

2.3. The role of expertise in linguistic annotation for NLP

Some of the findings of the methodological studies cited above are, in brief summary:

- Teams of "experts" tend to show higher IAA than teams of "non-experts",
- but the differences are generally small.
- IAA is lower in mixed groups of experts and non-experts than in homogeneous groups (of both kinds).

- IAA is dependent on the task.
- IAA is dependent on the complexity of the annotation scheme.
- IAA is dependent on the number of annotators.

Most of these findings come from the meta-study by Bayerl and Paul (2011), where we learn many interesting things about linguistic annotation projects. However, new questions also arise in this connection which Bayerl and Paul (2011) do not address (presumably because of insufficient data). In particular, it would be useful to find out if training and expertise interact with other factors, for example, if the impact of annotation scheme complexity is different with experts and non-experts, if an increase in the number of annotators impacts IAA negatively to the same extent with experts and non-experts, etc.

Differences between expert and non-expert annotators have been noted by a number of authors (Kilgarriff 1999, Wilks 2000, Snow et al. 2008, Gillick and Liu 2010, Artstein 2017), and more generally for linguistic judgements by Dąbrowska (2010). Despite this, a general impression gleaned from accounts of NLP gold standard creation is that linguistic expertise acquired through formal academic training is undervalued.[6] It is difficult to understand the frequent omission of annotator qualifications in any other way, especially since we sometimes are told explicitly that the annotators are e.g. computer science students.

This much-discussed difference between expert and non-expert annotators in fact largely mirrors another, in my view more fundamental dichotomy in the kinds of annotations found in NLP gold standard datasets, to which we now turn.[7]

3. Baby and bathwater

A recent article by Uma et al. (2021) presents a useful summary of many issues in connection with IAA, and in particular in connection to the low IAA scores often reported for various kinds of NLP annotation tasks (e.g., by Lindahl et al. 2019 for argumentation annotation).[8]

They single out "inherently subjective judgments" as particularly amenable to variation in annotation results (and hence low IAA), but note that even annotation of "objective and 'simple' aspects of language" is not free from such problems (Uma et al. 2021:1387).

I believe that some of the discussion in the literature around annotation quality is confused by a mixup of two quite different kinds of annotation. By and large, the "objective and 'simple' aspects of language" mentioned above are facets of *linguistic analysis*, while the "inherently subjective judgments" are exactly this: judgements of (aspects of) language expressions made by language users. Filing both these activities under the heading "annotation" serves to obscure the fact that they are in reality quite different phenomena. The former requires (highly trained) experts, and the latter "only" ordinary language users. On a charitable interpretation, their conflation may be due to a belief that the native speaker's status as "expert" wielder of their language automatically also implies their expertise in formal linguistic analysis of the language, somewhat analogous to a belief that if you happen to inhabit and operate a human body, you will also automatically possess complete medical knowledge. This is a fallacy similar to the one discussed by Santana (2018) with regard to what should and should not count as scientific evidence.

Another reason for this ambiguity of the term "annotation" is surely to be found in the recent history of the NLP field. From the point of view of linguistics the current focus on deep-learning systems arguably represents a return to behaviorism (see, e.g., Alkon 1959, Passos and Matos 2007); annotation is framed as a black-box solution to a black-box problem. Instead of (scientific) analysis, annotation encodes (observational) data on which analyses are based.

This fits well with an observation made by Öhman (2021), that comparing lexicon-based and data-driven sentiment analysis is not comparing like with like. Word polarity values retrieved from a sentiment lexicon form a component in a suggested explanation of why a text or text passage will be perceived to carry a particular sentiment polarity, where the central explanatory device is tried-and-true linguistic compositionality. This may or may not be correct – this is an empirical matter – but it is emphatically not the same thing as attaching judgements about their sentiment to whole texts or text passages.

[6] How else are we to interpret statements such as "using experts is very expensive, prohibitively so for large-scale projects" (Uma et al. 2021:1386), with its implication that the experts are not really needed anyway.

[7] I am indebted to the two anonymous reviewers who both made remarks which steered my thinking in a – hopefully – fruitful direction.

[8] Uma et al. (2021) also propose and test concrete ways of compensating for low IAA in a machine learning setting, a topic which we will not be able to discuss further here.

From the point of view of somebody who still likes to believe that linguistics has something to contribute to annotation methodology in NLP, this wholesale return to behaviorism seems somewhat defeatist: the blanket label "inherently subjective judgments" is wielded in too cavalier a fashion, with sentiment analysis or detection of offensive language implicitly being put on a par with fashion preferences or individual (but in many cases shared) aversions to some words (see Liberman 2012). I suspect that similarly to how NLP research for a very long time largely ignored the quite mature linguistic subarea of language typology (Bender 2011; 2016), it seems that the field still remains unaware of the potential contributions by conversation analysis and text linguistics towards more objective analyses in these cases. This is not to deny that there are clearly more objective (or better: intersubjective) and more subjective language phenomena,[9] but I also believe that there is still a largely untapped resource (by NLP researchers) in the form of highly refined linguistic analysis of the "inherently subjective" phenomena (e.g. Klein 2018).

4. After the gold rush

Hopefully, it has become clear from the above, that to me, one of the more problematic aspects of NLP gold standard creation has to do with annotator qualifications and annotation quality.

I admit to being biased by having extensive linguistic training and having been exposed to large volumes of linguistically annotated text where IAA is unavailable for the simple reason that the annotation has invariably been carried out by only one expert (although drawing on native-speaker consultants). This is interlinear glossed text (IGT), resulting from linguistic fieldwork.[10]

Here are some methodological questions and hypotheses concerning the role of expertise in the form of formal training in linguistic analysis:

- Intuitively, we would expect different pairs of expert annotators to differ on more or less the same items, reflecting genuine disagreements of theory or practice (cf. Plank et al. 2014). This may mean that IAA will not be a very meaningful measure with expert annotators, and consequently that only one annotator will be needed in many cases (cf. IGT, mentioned above).
- Given the preceding point, what are the absolute limits to expert annotation in different domains (this is basically the question posed by Babarczy et al. 2006 and Sampson and Babarczy 2008)?
- Can we develop effective tools helping us to "clean up" inconsistent or erroneous annotations (cf. Dickinson and Meurers 2003, Dickinson 2009, Loftsson 2009, Kato and Matsubara 2010, Dickinson 2015, Hollenstein et al. 2016, de Marneffe et al. 2017)?
- The distinction between analysis and judgement is obviously primary when selecting an annotator pool. This will form a basis for when to solicit non-expert "products" and when an expert is required. However, this is not a strict dichotomy, since:
- Different linguistic analysis tasks seem differently tractable to lay annotators (Munro et al. 2010). Generally, it seems that annotation which focuses on analysis of linguistic form causes more difficulty to non-experts – except possibly (non-)existence of a particular form (such as a non-word) – than annotation focusing on (suitably granular semantic or pragmatic) content.
- Which kinds of linguistic training or expertise make a difference, and to which annotation tasks? Can we agree on how to specify annotator qualifications?[11]

5. A conclusion of sorts

Since this article has been written as a kind of methodological opinion piece, a definite conclusion is not so easily formulated. I can offer this, however: In order not to give in completely to the "new behaviorism" characterizing present-day NLP, it could be beneficial to ask annotators not only for judgements, but also for them to indicate what prompted (positive) judgements, similarly to McDonnell et al. (2016), who ask their annotators to provide (free-text) rationales

[9] There are of course also various kinds of linguistic – including idiolectal – variation, which may inform some kinds of annotation, but these are notably *not* subjective, but amenable to linguistic analysis.

[10] To be fair, I do find inconsistencies in such data, so that this form of annotation could benefit from consistency-checking NLP tools.

[11] It is at least questionable whether (even linguistics) student annotators can credibly be referred to as "experts".

for information-retrieval relevance assessment ratings. In a linguistic annotation context, annotation instructions could be, e.g., not "mark all instances of hate speech in these texts", but instead: "mark all instances of hate speech in these texts, together with the features (words, phrases, etc.) that in your view flag them as hate speech". Hence: a structural analysis task rather than a pure classification task.[12]

In addition, I have hopefully managed to convince the reader that there are many exciting methodological avenues to explore in the area of linguistic annotation for NLP, with the potential to contribute to improving the quality of gold standard datasets, or at least to improving our confidence in them.

Acknowledgements

I extend my heartfelt thanks to two anonymous reviewers for sharing useful literature references and for posing some pointed questions which have forced me to sharpen my thinking about the issues discussed here.

References

Alkon, Paul K. 1959. Behaviourism and linguistics: An historical note. *Language and Speech* 2 1: 37–51. https://doi.org/10.1177/002383095900200105.

Antonsen, Lene, Trond Trosterud, and Linda Wiechetek. 2010. Reusing grammatical resources for new languages. In *Proceedings of LREC 2010*, pp. 2782–2789. ELRA, Valletta.

Artstein, Ron. 2017. Inter-annotator agreement. In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, pp. 297–313. Springer, Dordrecht. https://doi.org/10.1007/978-94-024-0881-2_11.

Artstein, Ron and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics* 34 4: 555–596. https://doi.org/10.1162/coli.07-034-R2.

Babarczy, Anna, John Carroll, and Geoffrey Sampson. 2006. Definitional, personal and mechanical constraints on part of speech annotation performance. *Natural Language Engineering* 12 1: 77–90. https://doi.org/10.1017/S1351324905003803.

Bayerl, Petra Saskia and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics* 37 4: 699–725. https://doi.org/10.1162/COLI_a_00074.

Bender, Emily M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6 3.

Bender, Emily M. 2016. Linguistic typology in natural language processing. *Linguistic Typology* 20 3: 645–660. https://doi.org/10.1515/lingty-2016-0035.

Brown, Susan Windisch, Travis Rood, and Martha Palmer. 2010. Number or nuance: Which factors restrict reliable word sense annotation? In *Proceedings of LREC 2010*, pp. 3237–3243. ELRA, Valletta.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics* 22 2: 249–254.

Church, Kenneth Ward and Joel Hestness. 2019. A survey of 25 years of evaluation. *Natural Language Engineering* 25 6: 753–767. https://doi.org/10.1017/S1351324919000275.

Dąbrowska, Ewa. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27: 1–23. https://doi.org/10.1515/tlir.2010.001.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, edited by Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché-Buc, pp. 177–190. Springer, Berlin.

Dickinson, Markus. 2009. Correcting dependency annotation errors. In *Proceedings EACL 2009*, pp. 193–201. ACL, Athens.

Dickinson, Markus. 2015. Detection of annotation errors in corpora. *Language and Linguistics Compass* 9 3: 119–138. https://doi.org/10.1111/lnc3.12129.

Dickinson, Markus and W. Detmar Meurers. 2003. Detecting errors in part-of-speech annotation. In *Proceedings of EACL 2003*, pp. 107–114. ACL, Budapest.

Gillick, Dan and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 148–151. ACL, Los Angeles.

[12]This would also mesh nicely with a growing interest on the part of the NLP community in *explainable AI*, with work on "probing" neural networks for internal structures corresponding to conventionally assumed linguistic information (e.g., Şahin et al. 2020).

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of ACL/HLT 2011*, pp. 42–47. ACL, Portland.

Hämäläinen, Mika and Khalid Alnajjar. 2021. The Great Misalignment Problem in human evaluation of NLP methods. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 69–74. ACL, Online.

Harrigan, Atticus G., Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology* 27 4: 565–598. https://doi.org/10.1007/s11525-017-9315-x.

Hetmański, Marek. 2018. Expert knowledge: Its structure, functions and limits. *Studia Humana* 7 3: 11–20. https://doi.org/10.2478/sh-2018-0014.

Hollenstein, Nora, Nathan Schneider, and Bonnie Webber. 2016. Inconsistency detection in semantic annotation. In *Proceedings of LREC 2016*, pp. 3986–3990. ELRA, Portorož.

Hovy, Dirk, Barbara Plank, and Anders Søgaard. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proceedings of ACL 2014*, pp. 377–382. ACL, Baltimore. https://doi.org/10.3115/v1/P14-2062.

Kato, Yoshihide and Shigeki Matsubara. 2010. Correcting errors in a treebank based on synchronous tree substitution grammar. In *Proceedings of ACL 2010*, pp. 74–79. ACL, Uppsala.

Kilgarriff, Adam. 1999. 95% replicability for manual word sense tagging. In *Proceedings of EACL 1999*, pp. 277–278. ACL, Bergen.

Klein, Gabriella B. 2018. Applied linguistics to identify and contrast racist 'hate speech': Cases from the English and Italian language. *Applied Linguistics Research Journal* 2 3: 1–16. https://doi.org/10.14744/alrj.2018.36855.

Kucera, Henry and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence.

Liberman, Mark. 2012. Literary *moist* aversion. `https://languagelog.ldc.upenn.edu/nll/?p=4389`. Language Log post.

Lindahl, Anna, Lars Borin, and Jacobo Rouces. 2019. Towards assessing argumentation annotation – a first step. In *Proceedings of the 6th Workshop on Argument Mining*, pp. 177–186. ACL, Florence. https://doi.org/10.18653/v1/W19-4520.

Loftsson, Hrafn. 2009. Correcting a POS-tagged corpus using three complementary methods. In *Proceedings of EACL 2009*, pp. 523–531. ACL, Athens.

Manning, Christopher D. 2015. Last words: Computational linguistics and deep learning. *Computational Linguistics* 41 4: 701–707. https://doi.org/10.1162/COLI_a_00239.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19 2: 313–330.

de Marneffe, Marie-Catherine, Matias Grioni, Jenna Kanerva, and Filip Ginter. 2017. Assessing the annotation consistency of the Universal Dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 108–115. LiUEP, Pisa.

McDonnell, Tyler, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? Collecting annotator rationales for relevance judgments. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pp. 139–148. AAAI Press, Palo Alto.

Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6 1: 1–28. https://doi.org/10.1080/01690969108406936.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 122–130. ACL, Los Angeles.

Öhman, Emily. 2021. The validity of lexicon-based emotion analysis in interdisciplinary research. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, pp. 7–12. ACL, Online.

Palmer, Martha, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different sense granularities for different applications. In *Proceedings of ScaNaLU 2004 at HLT-NAACL 2004*, pp. 49–56. ACL, Boston.

Passos, Maria de Lourdes R. da F. and Maria Amelia Matos. 2007. The influence of Bloomfield's linguistics on Skinner. *Language and Speech* 30 2: 133–151.

Plank, Barbara, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of ACL 2014*, pp. 507–511. ACL, Baltimore. https://doi.org/10.3115/v1/P14-2083.

Pradhan, Sameer, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proceedings of SemEval 2007*, pp. 87–92. ACL, Prague.

Pustejovsky, James, Patrick Hanks, Roser Saur, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003)*, pp. 647–656. Lancaster University, Lancaster.

Reiter, Ehud. 2007. Last words: The shrinking horizons of computational linguistics. *Computational Linguistics* 33 2: 283–287. https://doi.org/10.1162/coli.2007.33.2.283.

Şahin, Gözde Gül, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics* 46 2: 335–385. https://doi.org/10.1162/coli_a_00376.

Sampson, Geoffrey and Anna Babarczy. 2008. Definitional and structural constraints on structural annotation of English. *Natural Language Engineering* 14 4: 471–494. https://doi.org/10.1017/S1351324908004695.

Santana, Carlos. 2018. Why not all evidence is scientific evidence. *Episteme* 15 2: 209–227. https://doi.org/10.1017/epi.2017.3.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, pp. 254–263. ACL, Honolulu.

Strapparava, Carlo and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of SemEval 2007*, pp. 70–74. ACL, Prague.

Trosterud, Trond. 2006. Grammatically based language technology for minority languages. In *Lesser-Known Languages of South Asia: Status and Policies, Case Studies and Applications of Language Technology*, edited by Anju Saxena and Lars Borin, pp. 293–315. Mouton de Gruyter, Berlin.

Trosterud, Trond. 2012. A restricted freedom of choice: Linguistic diversity in the digital landscape. *Nordlyd* 39 2: 89–104.

Uma, Alexandra N., Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research* 72: 1385–1470.

Wilks, Yorick. 2000. Is word sense disambiguation just one more NLP task? *Computers and the Humanities* 34: 235–243.

Wintner, Shuly. 2009. What science underlies natural language engineering? *Computational Linguistics* 15 4: 641–644. https://doi.org/10.1162/coli.2009.35.4.35409.

Zaenen, Annie. 2006. Last words: Mark-up barking up the wrong tree. *Computational Linguistics* 32 4: 577–580.