Balancing the Existing and the New in the Context of Annotating Non-Canonical Language

Ann Bies
Linguistic Data Consortium
University of Pennsylvania
3600 Market Street
Philadelphia, PA 19104, USA
bies@ldc.upenn.edu

Abstract

The importance of balancing linguistic considerations, annotation practicalities, and end user needs in developing language annotation guidelines is discussed. Maintaining a clear view of the various goals and fostering collaboration and feedback across levels of annotation and between corpus creators and corpus users is helpful in determining this balance. Annotating non-canonical language brings additional challenges that serve to highlight the necessity of keeping these goals in mind when creating corpora.

Introduction

Context is important – both the linguistic context of a specific annotation and also the external context of the project as a whole affect what type of annotation scheme can be developed, what kind of annotation can be done, and what the balance of existing and new will need to be in an annotation scheme. Non-canonical language can make the usual linguistic and situational context considerations for annotation even more relevant: how broad the context is (word, sentence, document, conversation, world knowledge), how much that context affects the feature that is being annotated, and whether it is possible for an annotator to take that context into account. In addition, particularly when developing large corpora as part of projects with a short timeline and restricted funding, which is often the case at the Linguistic Data Consortium (LDC), a necessary part of choosing or designing an annotation scheme is considering who the end users of the annotated data will be, what the annotations will be used for, what level of detail is important for the project, and what level of accuracy or consistency is desired.

What are the factors that lead to the adoption of a totally new annotation scheme rather than using an existing annotation scheme?

Since the development of entirely new annotation guidelines is a time-consuming endeavor, it is worth considering whether totally new development is necessary. It may be necessary, if the annotation task is entirely new, or if the goals for using the annotation are entirely new, and neither can take advantage of existing resources.

However, in addition to the potential cost and time to develop entirely new guidelines, several factors could lead positively to the choice of using or adapting existing annotation guidelines for a new task:

- The existence of a large volume of annotated data in an existing annotation scheme that is closely related
- The goal or need to combine existing annotated data with the newly annotated data for statistical, training, or evaluation purposes

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 140–143,
Denver, Colorado, June 5, 2015.

- A team of annotators already well trained in an existing annotation scheme
- The feasibility of adapting existing annotation guidelines to meet the goals of a new task
- The existence of a well-designed annotation GUI for an existing task

The non-canonical language that LDC has had experience with includes informal genres (such as SMS/Chat data and speech data) and also dialectal data in languages other than English (such as Egyptian Arabic, which does not have a standardized written form).

When LDC began a project to create English treebank annotations on web text data, we chose to use the existing Penn Treebank guidelines (Bies et al., 1995), but to make additions and adaptations to account for the non-canonical language that appears in internet communication. The existing guidelines addressed most of the syntactic structures that were likely to come up, and the existing annotation tool could handle most of them as well. However, the novel constructions that were present in the data required new guidelines, and some new features also had to be added to the annotation tool. In this case, developing entirely new annotation guidelines and tools would have been prohibitively expensive in both time and effort, and the combination of existing and new worked well for the project (Bies et al., 2012).

Similarly, LDC developed Entities, Relations, and Events (ERE) annotation to support requirements in the DEFT program, including informal genres, and based that development on adapting existing ACE guidelines (Doddington et al., 2004). LDC first defined Light ERE as a simplified form of ACE annotation, with the goal of being able to rapidly produce consistently labeled data in multiple languages (Aguilar et al., 2014), taking advantage of the taxonomy and distinctions developed for ACE. In a second phase of development, Rich ERE expanded entity, relation and event ontologies and also expanded the notion of what is taggable, to provide better support for evaluation tasks in the program. Rich ERE also introduced expanded event coreference with the notion of event hoppers, particularly with respect to event mention and event argument granularity variation (Song et al., 2015).

Treebank and ERE guidelines that have been completed for English have been later adapted for other languages as well – for example, Modern Standard Arabic and also dialectal Arabic treebanks (Maamouri and Bies, 2004; Maamouri et al, 2014; Maamouri et al., 2006; Eskander et al., 2013), as well as Chinese and Spanish ERE (Song et al., 2015). Clearly, new guidelines are necessary to account for language-specific constructions for each language and annotation task, but developing them based on existing guidelines for another language is a considerable head start.

How do you decide on the granularity of the distinctions you choose to annotate? Give examples.

We aim for a level of granularity in annotation distinctions that is

- Consistent with goals of the annotation task and the guidelines
- Useful for downstream users of the data or additional downstream annotation
- Possible for annotators to distinguish reliably

For example, in part-of-speech tagging English web and SMS/Chat text, we make a distinction between emoticons and other decorative uses of punctuation. End users of the data have suggested that the distinction could be useful, since there could be a semantic difference between the two uses, and annotators are able to make the distinction reliably.

In a more structural example from the same data, the syntactic annotation of internet initialisms (such as lol, icymi, rofl, etc.) requires a decision about how much internal structure to give them. Since not every word of the spelled out version is necessarily part of the initials, and since in any case there is often disagreement about what the full spelled out version should be, we do not spell out internet initialisms as part of the annotation. They are left as written and annotated by function in the tree, even if the spelled out version could have internal structure. For example, "atm" for *at the moment* is annotated simply as a one-word temporal adverbial phrase (although the fully spelled out *at the moment* would be a more complex prepositional phrase that includes a noun phrase complement):

```
(ADVP-TMP atm)
```

However, if an initialism takes additional arguments, such as clausal arguments of "idk" for *I*

don't know, the argument structure is shown in the tree, so that it is as consistent as possible with other similar structures. The initialism is not spelled out, but at the same time its clausal complement is also annotated:

```
(S (NP-SBJ *PRO*)
   (VP idk
      (SBAR (WHADVP-1 where)
         (S (NP-SBJ I)
            (VP can
               (VP go
                  (ADVP-DIR-1 T*)
   ))))))
```

In developing the concept of event hoppers for Rich ERE, we coreference event mentions at the same level of granularity as ACE (i.e., type and subtype match, and sub-events are treated as separate events), but we allow a greater degree of flexibility in the granularity of the arguments that can be participants in coreferenced event mentions than in ACE (Song et al., 2015). For example,

- Granularity of temporal and spatial expressions (*Attack in Baghdad on Thursday* vs. *Bombing in the Green Zone last week*)
- Trigger granularity (*assaulting 32 people* vs. *wielded a knife*)
- Argument granularity (*18 killed* vs. *dozens killed*)

Relaxing the granularity requirements in this way allows annotators to coreference more event mentions that they know refer to the same event. It more closely matches annotator intuitions, and it gives end users a more complete picture of the annotated events and their participants.

For building new resources for NCLs, is it still worthwhile to invest a huge amount of time and human labour for manual annotation, considering that the annotators spend most of their time making arbitrary decisions, and that the aim of building 'high-quality resources' for NCLs might not be realistic?

Manually annotated resources provide information that may not be possible to determine using automated systems only. High-quality manual annotation of non-canonical language is possible to achieve, given clear annotation guidelines and careful training of annotators.

The premise of the question – that annotators must spend most of their time making arbitrary decisions – seems incorrect to me. It is possible to eliminate or minimize arbitrary decisions in the development of annotation guidelines when that is a priority.

It is also important to keep in mind, however, that different projects and different users may have different requirements regarding quality. "High quality" will not mean the same thing to everybody, and an annotated corpus is valuable if it helps the end users do what they want to do with it. Not all end users require high annotator consistency, and not all end users require a notion of a single right answer.

In addition, not all annotation "improvements" have the same cost, or the same benefit. Some annotation updates may be quite simple or fast but are high value in terms of system performance. Other updates might be difficult or slow and end up not bearing much fruit for the end users. A close feedback loop between corpus creators and corpus users is helpful in terms of selecting what kinds of updates are worthwhile given limited resources. This type of beneficial feedback loop was in place during the development of the Arabic Treebank and Arabic morphological analyzers and parsers (Maamouri et al., 2014; Maamouri et al., 2008; Maamouri et al., 2011; Eskander et al., 2013).

On a related note, what are the considerations when choosing the level of expertise of the annotators? When is crowd sourcing appropriate? When do we need linguistic experts?

The complexity of the annotation task and the required level of consistency for the annotation are the primary considerations in determining the necessary level of linguistic expertise.

Can the concept of "gold annotations" be applied to non-canonical languages where the inherent ambiguity in the data makes it hard to decide on the "ground truth" of an utterance?

For tasks such as syntactic annotation, instances where the inherent ambiguity in the linguistic data makes it impossible to decide on the ground truth of an utterance in context are rare, even in informal genres. If language as it is used were impossibly

ambiguous, human communication could not take place. However, the context of the utterance is important, as is giving annotators access to as much of that context as possible. There are certainly situations where the full context may not be available, or where the full context may include non-linguistic factors such as gesture or world knowledge, and those cases will be difficult.

Ambiguity is certainly present in many forms, in non-canonical (and also canonical) language. It may be that allowing or highlighting that ambiguity as part of the "gold annotation" would be valuable. There are also annotation tasks where various gradient phenomena in the data call into question the reality of a single correct answer. When annotating those phenomena is valuable, multiple correct answers or annotated gradients could also be considered as a part of gold annotation.

Acknowledgments

This material is based upon research supported by the Defense Advanced Research Projects Agency (DARPA) Contract No. HR0011-11-C-0145 and Air Force Research Laboratory agreement number FA8750-13-2-0045. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory and Defense Advanced Research Projects Agency or the U.S. Government. Portions of this work were supported by a gift from Google, Inc.

References

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, Joe Ellis. 2014. A Comparison of the Events and Relations Across ACE, ERE, TAC-KBP, and FrameNet Annotation Standards. *ACL 2014: 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, June 22-27. 2nd Workshop on Events: Definition, Detection, Coreference, and Representation*.

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre. 1995. *Bracketing Guidelines for the Treebank II-style Penn Treebank Project*. University of Pennsylvania, Department of Computer and Information Science Technical Report MS-CIS-95-06.

Ann Bies, Justin Mott, Colin Warner, Seth Kulick. 2012. *English Web Treebank*. Linguistic Data Consortium, LDC Catalog No.: LDC2012T13.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program - tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004), Lisbon, May 24-30.

Ramy Eskander, Nizar Habash, Ann Bies, Seth Kulick, Mohamed Maamouri. 2013. Automatic Correction and Extension of Morphological Annotations. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 1–10, Association for Computational Linguistics, Sofia, Bulgaria, August 8-9, 2013.

Mohamed Maamouri and Ann Bies. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING 2004).

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow, Dalila Tabessi. 2006. Developing and Using a Pilot Dialectal Arabic Treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC 2006).

Mohamed Maamouri, Ann Bies, Seth Kulick. 2008. Enhancing the Arabic Treebank: A Collaborative Effort toward New Annotation Guidelines. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC 2008).

Mohamed Maamouri, Ann Bies, Seth Kulick, Nizar Habash, Reem Faraj, Ryan Roth. 2011. Arabic Treebanking. In Joseph Olive, Caitlin Christianson and John McCary (Eds.), *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 2014).

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, Xiaoyi Ma. 2015. From Light to Rich ERE: Annotation of Entities, Relations, and Events. In *Proceedings of The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, at The 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015).