Non-canonical language is not harder to annotate than canonical language

Barbara Plank, Héctor Martínez Alonso, Anders Søgaard
Center for Language Technology, University of Copenhagen
Njalsgade 140, Copenhagen, Denmark
bplank@cst.dk,alonso@hum.ku.dk,soegaard@hum.ku.dk

Abstract

As researchers developing robust NLP for a wide range of text types, we are often confronted with the prejudice that annotation of *non-canonical language* (whatever that means) is somehow more arbitrary than annotation of canonical language. To investigate this, we present a small annotation study where annotators were asked, with minimal guidelines, to identify main predicates and arguments in sentences across five different domains, ranging from newswire to Twitter. Our study indicates that (at least such) annotation of non-canonical language is *not* harder. However, we also observe that agreements in social media domains correlate less with model confidence, suggesting that maybe annotators disagree for different reasons when annotating social media data.

1 Introduction

Recently, our research group received the reviews of a paper we submitted to a major, influential journal. The paper included a description of in-house linguistic annotation of Twitter data. One reviewer complained that "the use of Twitter as a corpus might be problematic because of the characteristic use of non-standard/typical language." What the reviewer presumably meant is that linguistic annotation of Twitter data is more arbitrary than annotation of standard or canonical language, e.g., newswire. We believe this premise, or prejudice, is false. "Standard language", as found in newswire and textbooks, for example, is a very biased sample of the linguistic productions in a language community, and the vast majority of the language we process and produce through the course of a day is very different from newswire and textbooks, be it spoken language, literature, or social media text.

Why, then, is newswire considered more standard or more *canonical* than other text types? Obviously, this may simply be because journalists are trained writers and produce fewer errors. But think, for a minute, about languages in which no newspapers are written. What, then, is canonical language? Can spoken language be canonical? Or is newswire called canonical, because, historically, it is what corpora are made of, and the only data that was available to the NLP community for a long time?

This discussion is more than a fight of words. The use of the word 'canonical' alludes to the fact that non-canonical language presents a challenge to the NLP community, but a lot of the reason for NLP tools performing poorly on social media texts and the like seems to be a historical coincidence. Most resources, e.g., syntactic and semantic treebanks, are human-annotated subsets of newswire corpora, simply because most electronic text corpora were newswire corpora when the NLP community began building treebanks. The question is whether annotating non-canonical language, say social media text, is inherently harder than annotating more canonical language, say newswire.

We believe some types of non-canonical language pose interesting *processing* challenges, e.g., with more mixed language, more *ad hoc* spelling conventions, and more texts directed at smaller audiences with more knowledge required during interpretation. However, newswire also comes with its

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 148–151,
Denver, Colorado, June 5, 2015.

complexities (headlinese, creative language use, citations, etc.), and if it was not for the skewed distribution of linguistic resources, we do not see why processing social media should be harder than processing newswire.

The skewed distribution underlines the need for new resources, and consequently, raises the important question whether *annotating* non-canonical language, e.g., social media text, is inherently harder than annotating canonical language. There is no priori reason why this should be the case. A full investigation of this question would take a lot of annotation studies, controlling for task, annotator groups, languages, etc.; something which is out of the scope of this squib. Instead, we present a pilot study of a single, specific linguistic annotation task (identifying main verbs and arguments) with two annotators and 50 sentences for each of five different domains (250 annotated sentences in total). Obviously, this is but a toy experiment, and our results should be taken with a grain of salt. However, our design is replicable, the annotated data available,[1] and we hope that others will take up replicating these experiments on a larger scale. Meanwhile, we leave the world with what our toy experiment suggests.

Note that we cannot just compare reported inter-annotator agreement scores across existing projects. Such scores are affected by sample biases, training of annotators, and the completeness of annotation guidelines. Thus, in this position paper we present an annotation study where we asked the *same* annotators to annotate canonical and non-canonical language (over five domains, ranging from newswire to Twitter) with minimal guidelines.

2 Annotating main verbs and arguments

We introduce the simple annotation task of identifying main predicates and arguments in sentences across five different domains.

Annotation Two expert annotators were asked to provide the following three labels:

1. MAINVERB (MV), the main lexical verb of the predicate, e.g., "he was **eating** apples".[2]
2. A0, the subject.
3. A1, which corresponds to two different syntactic functions. A1 is the direct object if there is a MV in the annotated sentence (i.e., "he had been eating **apples**") or the attribute in a copula construction ("he is **happy**").

The only guideline was not to mark auxiliaries, and that the first word in a coordination or multiword unit is the head.

DOMAIN	TOK	TTR	$\overline{SL}$	OOV
WSJ	743	0.56	14.86±2.93	4.2%
Twitter	657	0.67	13.14±3.30	**38.9%**
Answers	674	0.54	13.48±3.00	9.4%
Spoken	646	**0.35**	**12.92±3.05**	6.6%
Fiction	691	0.51	13.82±3.26	8.2%

Table 1: Data characteristics (50 sentences each).

Corpora We selected five different corpora constituting different degrees of perceived canonicity.

1. Wall Street Journal (WSJ): Section 23 from the Ontonotes distribution of the Wall Street Journal dependency treebank (Bies et al., 2012; Petrov and McDonald, 2012).
2. Answers: The Yahoo! Answers test section from the English Web Treebank (Bies et al., 2012; Petrov and McDonald, 2012).
3. Spoken: The Switchboard corpus section of the MASC corpus (Ide et al., 2008).
4. Fiction: The literature subset of the test section of the Brown test set from CoNLL 2008 (Surdeanu et al., 2008), which encompasses the *fiction*, *mystery*, *science-fiction*, *romance* and *humor* categories of the Brown corpus.
5. Twitter: The test section of the Tweebank dependency treebank (Kong et al., 2014).

WSJ is the perceived-of-as-canonical dataset. Answers and Twitter are datasets of social media texts from two different social media. We include Switchboard as an example of spoken language (transcriptions of telephone conversations), and Fiction to incorporate carefully edited (i.e., not user-generated) text that is lexically and syntactically different to newswire. From each corpus, we randomly selected 50 sentences and doubly-annotated them.

[1] `https://bitbucket.org/bplank/predicates`

[2] We follow the Stanford dependency convention in that copulative verbs are not treated as main verbs, and are dependents of the attribute. Thus, here the copula is not marked as MV.

DOMAIN	A0	A1	MV
WSJ	99	76	72
Twitter	88	72	56
Answers	92	79	63
Spoken	100	86	81
Fiction	96	76	78

Table 2: Frequency counts for arguments in the annotated data (50 sentences per domain, two annotators each).

	MATCH		F1			
DOMAIN	EXACT	FRAMES	A0	A1	MV	MICRO
WSJ	66%	82%	0.87	0.66	0.83	0.79
Twitter	52%	66%	0.91	0.69	0.79	0.80
Answers	74%	84%	0.98	0.81	0.88	**0.90**
Spoken	43%	74%	0.91	0.56	0.88	0.79
Fiction	64%	78%	0.83	0.75	0.79	0.80

Table 3: Agreement statistics between the two annotators.

DOMAIN	ρ
WSJ	0.8002
Twitter	0.7019
Answers	**0.6489**
Spoken	0.8165
Fiction	0.8406

Table 4: Correlation (Spearman's ρ) between annotator agreement (how many arguments match out of both) and system confidence (average per-edge confidence).

Table 1 provides statistics for all datasets, namely the amount of tokens (TOK), the type-token ratio (TTR), the average sentence length ($\overline{SL}$), and the out-of-vocabulary rate with regards to the WSJ training section (OOV). We use this last metric as an indicator on how much a domain deviates lexically from newswire. No normalization has been performed. Spoken data has the shortest sentences but the lowest TTR, that is, it is the domain with the highest lexical variation. Nevertheless, the domain with by far the highest OOV is Twitter. $\overline{SL}$ is 13–15 words for the five domains, with slightly longer sentences in newswire. Table 2 provides characteristics of the annotations, i.e., counts for the three annotation labels by both annotators without adjudication (i.e., over the union of the data annotated by two annotators). Subject dropping and imperative mood is common in Twitter, which decreases A0, and fully-formed clauses are also less frequent, thus affecting MV and A1. For completeness, we compare the annotations to the gold dependency trees available in the treebanks. We do so by computing labeled attachment scores for strictly the set of annotated words. The results range from 0.85 LAS on WSJ to 0.56 on Switchboard.

Results Table 3 shows label-wise and micro-averaged F1 scores between annotators for each of the domains. Surprisingly, we see among the lowest agreement on newswire, but all five domains seem about equally hard to annotate, except Answers (which is easier). Again, we remind the reader that this is miniature annotation study, but we think this is an interesting observation.

Newswire may be harder to understand because it is more complex language. For example, we observed that average sentence length was slightly longer for newswire. We measured the correlation between sentence length and sentence-wise agreement for all 250 annotated sentences, however, found the correlation to be low (0.1364). Consequently, it seems unlikely that sentence length had a major effect on our annotations.

We may speculate that annotation disagreements can be due to rare linguistic phenomena and linguistic outliers. In Table 4 we show the correlation per domain between sentence-wise agreement and dependency parsing confidence. We have obtained this confidence from the edge-wise confidence scores provided by an instance of the MST parser (McDonald et al., 2005) trained on WSJ. The parsing confidence for a sentence is obtained from the average of the edges that have received a label (A0, MV, A1) by the annotators, averaged between the two annotators. The correlation for newswire is high, but not the highest, because despite high parsing confidence, annotation agreement is rather low. On the other end, the lowest correlation between parser confidence and agreement is for Answers, which has the highest inter-annotator agreement.

These results, in our view, indicate that what makes annotating social media text hard (at times) is not what makes annotating newswire hard. We leave it for now to validate this finding on a larger scale, as well as to try to understand what makes annotating social media (relatively) hard.

	DOMAIN	FRAME	EXAMPLE
1	Twitter		@user he/A0 better/A1 !! we/A0 buy/MV his stuff/A1 ! haha
2	Spoken	x	those/A1A0 are the ones/A0A1 that I really really hate too
3	Spoken		I/A0 agree/MV with you/A1 on that particular subject there
4	Fiction	x	" I/A0 mean/MV , do you/A0 feel/A1MV like seeing/A1 Kate " ? ?
5	Answers		– sigh – not trying/MV to sounds snooty or stuck up but I/A0 mean/MV really !
6	WSJ	x	Fidelty/A0 on Saturday opened/MV its 54 walk/A1 – in investor centers/A1 across the country .
7	WSJ	x	Nevertheless , he/A0 says/MV a depression does n't appear/A1 likely/A1 .

Table 5: Disagreement examples from all domains, annotator1=blue, annotator2=red, matches=black, Frame (cf., §3).

3 Discussion

Table 5 shows examples of different cases of disagreement from different domains. The native tokenization is kept intact. The FRAME column indicates whether the annotators provided the same valency frame, regardless of which words were said to be the arguments.

In Example 1, we can see a characteristic property of Twitter data, namely that there can be more than one sentence per tweet, and it is therefore often hard to decide what the main predicate is. Example 2 shows a copula case where the same frame is chosen by the two annotators, but they disagree which words satisfy which arguments. In Example 3, the annotators disagree on wether the verb "agree" has a valency-bound preposition ("with"), and thus whether it has a direct object or not. In Example 4, annotators disagree on whether "I mean" is the main clause, and thus the main predicate, or an off-clause satellite that roughly has the function of an interjection. In Example 5, annotators disagree what is the main clause. Example 6 shows disagreement caused by the difficulty to annotate already tokenized text, where it is not straightforward that the adjective "walk-in" has been tokenized apart. In Example 7, there is agreement on whether it is "appear" or "likely" that heads the subordinate clause and fulfills the A1 of the verb say. This disagreement stems from the copulative reading of "appear", which makes it a dependent of "likely" instead of its head in one case. To sum up, the main sources for disagreement stem from choice of main predicate and verb valency.

4 Conclusions

This squib presents a bold opinion and a severely underpowered pilot annotation study. The pilot study, in which we had professional annotators annotate main verbs and arguments with minimal guidelines, indicates that what some refer to as non-canonical language is not harder to annotate than canonical language. Our bold opinion is that the notion of canonical language is absurd and harmful, suggesting that some language, say, newswire, is better suited for linguistic resources than other types of language, say, spoken language or social media texts. What is considered non-canonical language is often the language that we use more often, and often commercially and scientifically more interesting. We believe there is no reason to expect that processing this type of text should be harder, with appropriate training data, and the pilot study presented here suggests that annotation is not harder either.

References

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank. *Linguistic Data Consortium, Philadelphia, PA*.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Charles Fillmore. 2008. Masc: The manually annotated sub-corpus of american english. In *LREC*.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A Smith. 2014. A dependency parser for tweets. In *EMNLP*.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *ACL*.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 Shared Task on Parsing the Web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL*.