

Good Reasons for Noting Bad Grammar: Constructing a Corpus of Ungrammatical Language

Jennifer Foster and Carl Vogel

Computational Linguistics Research Group, Trinity College, Dublin

`jfoster@tcd.ie, vogel@tcd.ie`

We describe the compilation of two written English corpora: the first corpus contains observed ungrammatical sentences, the second corpus contains hand-corrected and hence grammatical versions of the sentences in the first.

The original motivation for compiling the parallel corpora arose when considering the problem of parsing ill-formed language. While it is clear that probabilistic parsers are more successful than traditional non-probabilistic parsers at actually returning an analysis for an ungrammatical sentence, the analysis they return won't necessarily reflect the sentence's meaning if they don't know that sentences can sometimes be ill-formed. A realistic grammar, probabilistic or not, will have a concept of ungrammaticality. Such a concept should be informed by authentic ungrammatical language as opposed to the invented strings often used by linguists. The relationship between the ungrammatical sentences in the first corpus and their grammatical counterparts in the second provides an explicit characterization of the ways in which sentences can become deviant. Not only is this information useful within the practical domain of parsing, it also useful within linguistics, as a form of evidence for theories of grammar.

In order to compile the corpora, we needed to be able to identify sentences as ungrammatical, and in order to do this in a systematic way, we needed a working definition of ungrammaticality. The definition which we used is independent of any particular linguistic theory and defines ungrammaticality in terms of the more practical notion of error: a sentence is deemed ungrammatical if an error occurs in the sentence at a structural level. An error occurs at a structural level if the individual words in the sentence are well-formed. Our decision to classify a sentence as ungrammatical will depend on whether we think that an error has occurred, and as such is dependent on a linguistic judgement. The reliability of grammaticality judgements as a source of evidence has been called into question by many (cf. Schütze (1996)), and hence we err on the side of caution. If we are unsure about whether an error has occurred in a sentence, a note is made of this sentence but it is not included in the corpus. This is, however, relatively

rare because the ungrammatical sentences are encountered *in context* during the course of reading. van Dijk (1976), among others, has argued that grammaticality judgements can only be sensibly applied to sentences appearing in their natural context.

When a sentence is identified as ungrammatical, it is added to the “ungrammatical” corpus. The sentence is then corrected. This correction process involves rewriting the sentence so that it expresses the same meaning but does not contain an error. The corrected sentence is then included in the second “grammatical” corpus. In just over 20% of cases encountered so far, the ungrammatical sentence could be corrected in more than one way while keeping its meaning constant. In such cases all corrected grammatical versions were added to the second corpus.¹ If the meaning of the ungrammatical sentence is unclear and hence it isn’t obvious how to correct it, a note is made of the sentence but it is not included in the corpus. The context in which a sentence occurs reduces ambiguity about how to correct an ill-formed sentence. Consider *The following statements declares and creates an array of five dimensions*. Taken out of context in this way, it is not obvious whether it should be corrected by changing the number of the noun *statements* or the number of the verbs *declares* and *creates*. We might be tempted to suggest the former correction since it involves less change to the sentence but the actual context makes it clear that the latter correction is the appropriate one. The context in which a sentence occurs will also allow certain sentences to be diagnosed as ill-formed when they would not have been if taken out of context, for example, *The error position is indicated by the edges and nodes **build** upon the words*. On its own, this sentence is well-formed but having understood everything which preceded it in the manual in which it appeared, it becomes apparent that it is ill-formed and that the word *build* should be replaced by *built*.

The authors’ reading material is the source for the ungrammatical data. This consists of academic material, emails, newspapers and magazines, websites and discussion forums, drafts of own writing, student assignments, technical manuals, novels, lecture handouts, album sleeve notes, letters, text messages, teletext and signs. The ungrammatical corpus contains just over 20,000 words (1000 sentences). Along with the actual sentences, the string re-write operation which is applied to each grammatical sentence in order to produce its grammatical counterpart is also recorded.

References

- Schütze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. The University of Chicago Press.
- van Dijk, T. A. (1976). Acceptability in context. In S. Greenbaum, ed., *Acceptability in Language*, pp. 39–62. Mouton Publishers, The Hague.

¹The reliability of the corrections could be verified through a psycholinguistic experiment in which the task carried out by an experimental subject mirrors, as far as is possible, the correction task.