Exploiting parsed corpora in grammar teaching

Sean Wallis
University College London

Parsed corpora, or 'treebanks', have a wide range of applications, from scientific research into language production and variation, to algorithms for automatic natural language processing. In language education, parsed corpora have a great deal of potential for developing teaching resources and curricula, at a wide range of levels.

Corpora have been used in education for a number of years. The principal perspective of 'data driven learning' ('classroom concordancing', Johns and King 1991) hands students the same research tools as academic researchers, and encourages them to explore the data. However, several linguists, including Kaltenböck and Mehlmauer-Larcher (2005) have questioned the effectiveness of this approach. Without a highly knowledgeable teacher to direct the class, students learn only that language use is varied, without grasping the reasons for that variation. Whereas translation studies can benefit from a purely data-driven approach, without a highly-skilled teacher, data driven learning is not well suited to the teaching of grammatical concepts and rules.

The Survey of English Usage at UCL published the million-word *British Component of the International Corpus of English* (ICE-GB, Nelson *et al* 2002) parsed corpus, complete with the ICECUP research software in 1998. At the time, English language school teachers had a mixed reaction. They were enthusiastic at the availability of this rich resource: indeed, a 10,000 word sample corpus plus software, sufficient for many teaching tasks, was freely available. But they realised that they had insufficient knowledge of *what to teach* (to learn the grammar) or *how to teach it*.

This problem reflects a general problem in grammar teaching pedagogy. The missing element in applying corpora to teaching lies at the interface between students and corpus: *the knowledge of school teachers*.

In this respect, British school teachers face a particular historical problem (Crystal 2017). Formal grammar teaching in English was taken out of the UK school curriculum in the mid-1960s, but was reintroduced gradually from the 1990s onwards, becoming a formal requirement in 2000, and being codified further in 2013. The result is a generation of school teachers increasingly required to teach English grammar without ever being formally taught the subject. There is, consequently, a large demand for high-quality resources. But there is also a lot of misunderstandings and anxieties about grammar and its role in English language teaching.

At the same time, there is a long track record of using corpora to create English grammar teaching resources. Indeed, the first 'Survey' corpus was developed at the Survey of English Usage and was used to write the publication of Quirk *et al* (1985), a highly influential grammar book which spawned a Student edition. 1996 saw the first *Internet Grammar of English*, updated and republished as the

interactive Grammar of English for mobile devices in 2011.

However, whether traditional books, websites or mobile apps, these resources had one thing in common. They were academic works aimed at university students learning grammar for linguistic study. They were not resources aimed at school children learning English language skills, nor were they designed for their teachers. At most they contained the 'what to teach' element mentioned above. The language was academic and the problem of how to teach English grammar effectively to children was simply not covered.

In this paper we will discuss a research project begun in 2010, called *Teaching English Grammar in Schools*. This project began as an exploration of employing corpus resources in teaching English grammar to secondary school (high school) students and their teachers. Resources were developed on the website platform, *Englicious* (www.englicious.org). These included lesson plans and interactive assessment exercises, 'CPD' resources aimed at training teachers, project work and five-minute lesson 'starters'. When the UK National Curriculum was published in 2014 we incorporated the entire glossary as a searchable integrated tool (with further commentary to address missing or unclear aspects).

From the commencement of the project, we intended that as far as possible example sentences should be drawn directly from the ICE-GB corpus. This would be attractive on good computational design principles, providing a clean subdivision between teaching script and corpus example. The ICECUP database on the webserver can serve up random examples from hundreds of thousands of clauses and phrases, plus context. ICECUP's grammatical search tool, plus the full parsing in ICE-GB provides a very powerful method for obtaining examples.

However, this exercise raises what we elsewhere term the 'selection problem' (Mehl *et al* 2016): in brief, in the context of a particular exercise or script, *what meta-linguistic principles are required to select an example from a corpus?* We need to ensure that each example is readable, relevant and age-appropriate for the age group targeted, but we also need to ensure that where multiple examples are presented, the overall content is appropriately balanced. We may also need to prune back longer examples for reasons of space.

In practice, for secondary school students (11-16 years old) we found that this type of direct sampling could be effective for the more advanced, older students. However, the simplified grammar children were to be taught was insufficient when faced with the grammatical productivity of naturally-occurring examples. To take a simple example: to carry out a 'Spot the Verb' exercise with examples drawn from the corpus, it becomes necessary to explain that *mean* in *I mean it's working extremely well* is considered a discourse marker rather than a literal verb.

For the rest of the resource, examples were manually selected from the corpus and, where necessary, supplemented with invented or modified examples. Linguists carried out the selection task.

In 2013, *Englicious* was extended to the teaching of primary students (5-11). With younger children, the adult language of ICE-GB was simply too advanced for direct corpus sampling to be

usable. However, some of the interactive tools and resources could be appropriate, once they were simplified and directed to the pedagogical needs of primary children and teachers.

A single platform for all ages teaching the UK national curriculum in English grammar has the advantage of consistency. The grammar glossary and terminology is consistent across all ages. In practice, the single platform also helps students and teachers make the transition from primary to secondary education. The more advanced secondary school resources use the same grammatical terms as at primary, while the primary resources are still available for revision and reinforcement.

References

Crystal, David (2017). English grammar in the UK: a political history. Supplementary material *to Making Sense: the Glamorous Story of Engish Grammar.* London: Profile. Available from http://www.davidcrystal.com/?fileid=-5222

Johns, Tim and Philip King (eds.) (1991). *Classroom concordancing. English Language research Journal 4 (New Series)*, Birmingham University.

Kaltenböck, Gunther and Barbara Mehlmauer-Larcher (2005). Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching. *ReCALL* **17**.1: 65-84

Mehl, Seth, Sean Wallis and Bas Aarts (2016). Language learning at your fingertips: deploying corpora in mobile teaching apps. In Corrigan, K., Mearns, A. (eds.) *Creating and digitizing language corpora. Volume 3: Databases for Public Engagement*. Palgrave, Basingstoke. 211-239.

Nelson, Gerald, Sean Wallis, and Bas Aarts (2002). *Exploring Natural Language: working with the British component of the International Corpus of English*. Amsterdam: Benjamins.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Completing Parsed Corpora. From correction to evolution. Authors.Â Teaching and Learning English as a Global Language: Native and Non-Native Perspectives, TÃ¼bingen: Stauffenberg Verlag, p. 273â€“290.Google Scholar. BÃ¶hmovÃ¡, Alena, J. HajiÄ, E. HajiÄovÃ¡, B. HladkÃ¡ (2003), â€˜The Prague Dependency Treebank: a three-level annotation scenarioâ€™. This volume.Google Scholar. Carter, David (1997). The TreeBanker: a Tool for Supervised Training of Parsed Corpora.Â Exploiting fuzzy tree fragments in the investigation of parsed corpora, Literary and Linguistic Computing, 15: 251â€“263.CrossRefGoogle Scholar. Wallis, Sean, Gerald Nelson (2001). Knowledge discovery in grammatically analysed corpora, Data Mining and Knowledge Discovery.Google Scholar. Copyright information. Corpora are becoming well established in teaching in Universities. Corpora also have a role to play in secondary education, in that they can help decide how and what to teach, as well as changing the way in which puplis learn and providing the possibility of open-ended machine-aided tuition. Corpora also seem to provide what UK goverment sponsored reports on teaching grammar have called for â€“ a data-driven approach to the subject. Export citation Request permission. Copyright. Corpus ID: 11375700. Parsing Word-Aligned Parallel Corpora in a Grammar Induction Context. @inproceedings{Kuhn2005ParsingWP, title={Parsing Word-Aligned Parallel Corpora in a Grammar Induction Context}, author={J. Kuhn}, booktitle={ParallelText@ACL}, year={2005} }. J. Kuhn. Published in ParallelText@ACL 2005. Computer Science. We present an Earley-style dynamic programming algorithm for parsing sentence pairs from a parallel corpus simultaneously, building up two phrase structure trees and a correspondence mapping between the nodes. The intended use of the algorithm is in bootstrapping grammar...