What use can and should ACL researchers make of the Cambridge Grammar of the English Language?

Anonymous ACL submission

Abstract

Huddleston and Pullum (2002) (H&P) provides an 1842 page description of the grammar of English. We analysed the top 75[1] citations to H&P from the ACL Anthology. The community has indeed produced work that is strongly influenced by H&P, especially in linguistically challenging areas such as deixis, anaphora and negation. To illustrate the potential of H&P as source material for linguistically informed error analysis in a conceptually complex domain, we extract the examples from chapter 17 of H&P (Stirling and Huddleston, 2002), which deals with deixis and anaphora. We show how a representative modern co-reference engine (Clark and Manning, 2016a,b) handles these examples. Since every example in H&P is chosen to illustrate a point about English, and the authors provide text explaining the importance of the point, the error analyst has immediate access to a good proxy for relevant linguistic expertise.

1 Introduction

Huddleston and Pullum (2002) provides a large and detailed descriptive grammar of English. In this short paper we argue that H&P is a useful resource for NLP and computational linguistics, particularly when dealing with conceptually complex areas of language use.

1.1 H&P

Huddleston and Pullum (2002) provides a large descriptive grammar of English. It has been available to computational linguists since 2002. It makes no claim to being rigorously usage-based[2], but it still contains many naturalistic examples, curated for their relevance to an overall understanding of the grammar of English. (Henceforward, we refer to this work as H&P[3])

Self-evidently, use of a grammar of English is most relevant to English, but there are similar works for other languages that can and should be used in a similar way (Helbig and Buscha, 1998; Abeillé and Godard, 2010).

1.2 The potential of H&P as support for better error analysis

Many NLP researchers evaluate their work solely with respect to progress on standard test data sets, which are dominated by frequently occurring and (often) unchallenging examples. The attitude is completely appropriate when seeking to make progress against a micro-averaged metric, but sometimes misses the mark on advancing science. Fortunately, some researchers are more ambitious. For example, Kolhatkar et al. (2018) aims to broaden the field's understanding of non-nominal anaphora, and Parrish et al. (2021) explicitly challenges the tendency to avoid rare and difficult examples. Researchers whose prior work has used mainly standard test data sets, but who aspire to greater ambition, will need resources for doing more detailed error analysis than they are used to.

Nerbonne (2018) argues that conciseness was a crucial feature of curated test suites such as (Lehmann et al., 1996).

> [The examples were] not taken from corpora [of] authentically occurring speech or text, but instead consisted of minimal examples designed to determine whether grammatical processing systems were assigning correct analyses to sentences. [Nerbonne, 2018]

[1] The Anthology's interface finds the top 100 entries for each search, but not all are relevant.

[2] See Leech (2004) for discussion, and compare Quirk et al. (1985).

[3] Huddleston and Pullum are by no means the only authors. The work is a collaboration with numerous other experts. Chapter 17 (Stirling and Huddleston, 2002), which we treat in detail below, has Stirling as the lead author, supported by Huddleston, who contributes to every chapter.

Why? An integral component of a curated test suite is documentation of *why* each example is of interest. This is an opportunity for computational linguists to collaborate with more theoretically minded colleagues, but many theoreticians will be happier doing it with short, minimal examples than with long newspaper sentences. The skill set of a theoretical linguist includes the habit of attending to fine detail. Newspaper sentences are full to overflowing with distracting fine detail. At best, this slows a theoretician down. At worst, it derails the collaboration.

In H&P, we have close to 2000 pages of detailed content providing precisely the kind of information that a computational linguist could get from a theoretician. We suggest that direct use of examples from H&P is a feasible and useful alternative to the use of curated test suites, and offers the additional benefit of access to detailed descriptions, by expert linguists, of the thinking behind each example.

2 Evidence from citations

We[4] analyzed references to H&P in the ACL Anthology (Bird et al., 2008) and sorted them into three coarse categories: **substantive**, **incidental** and **common knowledge**[5].. The goal is not a definitive categorization, but rather an impression of the influence that H&P has had on each study. A reference is **substantive** when it seems to us unlikely that the study in question could have been undertaken without the specific concepts and background provided by H&P. A reference is **incidental** when material from H&P is used to motivate choices made in the study, but the study could perfectly well have been done without its influence. The label **common knowledge** applies if a citation serves to document common knowledge: for example, Alkorta et al. (2018) uses H&P to document a generally accepted notion of the scope of negation. This citation is arguably useful for clarifying that the paper's authors did not invent this notion, but does not shape the work. WE judge there were just 5 clear cases (6%) of citation purely to document common knowledge[6]. It is logically possible for a paper to include more than one of these kinds of citation to H&P. but in our sample this never happened.

Formally, the sample frame for the citation analysis is the set of entries returned by the search function of the ACL anthology when the query is 'Huddleston', restricted to those that are indeed references to H&P, and which have at least some computational linguistic content. The anthology search page, accessed using the term 'Huddleston' estimates that there are 230 matching entries, and shows 100 of them. A single researcher, with over 30 years of experience in computational linguistics, accessed each of the corresponding papers, searched for the reference, and read enough of the surrounding material to arrive at a categorization. Entries in the anthology that are not references to H&P but to other work by Huddleston were excluded, as were those entries that describe work that clearly has no computational content, and to results that were duplicative.This leaves 75 references to H&P that need to be categorized[7].

Examples of the **substantive** category include:

- Mosbach et al. (2020) notes that English grammar requires a relativizer in subject relative clauses, but licenses its omission in object relative clauses. H&P is cited to introduce that claim, and the study proceeds to investigate whether transformer-based masked language models can capture this aspect of the grammar.

- Ficler and Goldberg (2016) studies argument cluster coordination, and adopts a conceptualization and representational proposal from H&P.

- Kazantseva and Szpakowicz (2010) uses an aspectual classification based on H&P's notion of situation type.

35 papers were judged to be substantively influenced by H&P (47%).

Examples of the **incidental category** include the following:

- Zeldes (2018) cites H&P's discussion of singular 'they' as part of a delineation of the idea of notional anaphora: the latter being the focus of the study.

[4]This paper has just one author, but stylistically 'we' seems to read better

[5]We also scanned some references in Semantic Scholar (Ammar et al., 2018), but found no computational linguistic work that was not already referenced in the ACL anthology

[6]A spreadsheet containing our judgments about papers will be made available for research use. As befits an opinion paper, these are opinions.

[7]The supplementary material includes a spreadsheet with pointers to the relevant papers

- Kolhatkar et al. (2018), in a review of anaphora with non-nominal antecedents, quotes H&P to establish a baseline definition of anaphora.

There were 35 such incidental citations (47%)

2.1 Subcommunities

H&P is particularly influential on studies of negation (Morante and Blanco, 2012; Sarabi and Blanco, 2017; Banjade and Rus, 2016) and anaphora (Kolhatkar et al., 2018; Zeldes, 2018). These sub-fields of computational linguistics are conceptually complex, so it is not surprising that support from expert theoretical linguists has particular value. Further citation analysis, beyond the scope of the current study, could measure the possibility that H&P has indirect influence on work that does not cite it, but does cite the work that is directly shaped by H&P[8].

2.2 Conclusions of the citation study

On the basis of this citation analysis we can conclude that that H&P is cited on a systematic and regular basis by an influential part of the community that publishes in the ACL Anthology, realizing some of its potential.

3 Anaphora and deixis: using H&P for error analysis

Error analysis is highlighted as an opportunity in a review of H&P written for the CL journal (Brew, 2003). We demonstrate this opportunity by extracting[9] and using all the examples in chapter 17 of the grammar, which treats deixis and anaphora. These sub-fields are conceptually complex and critical for practical applications. We take error analysis broadly. One kind of error is when a modelling failure leads to a mis-classification of a single example; another is when a whole class of potentially valuable behavior is excluded, presumably because the test suites and the first widely available corpora focus on a narrow subset of the problem. Specifically, we see that most CL work on co-reference focuses only on nominal anaphora. Kolhatkar et al. (2018); Zeldes (2018) are notable exceptions. This is an error of the second kind.

There are 803 examples in the chapter where at least one candidate is underlined in H&P. The full list of examples is provided in the supplementary material. For 433 of these the Stanford NLU pipeline, used with its default settings, finds no co-reference chain. Noun-phrases and verb-phrases were extracted from the output of the Stanford pipeline using TREGEX expressions[10]. Over half of the examples correspond to something that the standard pipeline does not do. There are 520 distinct underlined targets in the corpus of examples.

(1) Frequent targets

a. herself it himself that she her they Jill he you (*Noun phrases*)

b. will do have to "do so" did doing "invite Kim as well" had be (*Verb phrases*)

c. so here one now then other otherwise their his have (*not identified as phrases*) .

195 of these targets (37.5%) are identified as noun-phrases by the Stanford analyser. (1-a) contains the most frequent of these: they are mainly pronominal antecedents. The full list is available in the supplementary material. see such examples in the test suite for any traditional co-reference analyzer.

There are 88 examples of targets (16.9%) that the Stanford analyzers identify as verb phrases. An example is in (2-a)

(2) Full examples.

a. A: You should phone her and ask if she has finished. B: I will , but I'm pretty sure she hasn't.

b. The idea was preposterous, but no one dared say so.

H&P comments that in (2-a):

> the anaphoric links cut across each other: the missing complement of 'will' is linked to 'phone her' and that of 'hasn't' to 'finished'.

Examples of frequent targets identified as verb phrases are in (1-b) and the full list is in the supplementary material. This points to opportunities

[8] Thanks to an anonymous reviewer for reminding us of this possibility, and suggesting ways in which the additional citation analysis might be done.

[9] The authors of the chapter provided it as a Microsoft Word document with highly regular structure and formatting. The author converted this to similarly regular HTML5 and a Python script used DOM operations to extract the examples. We can provide the software on request, but are not able to share the H&P source text.

[10] software available on request.

available by broadening the scope of mention extraction.

There are 273 targets (52.5%) that were not identified as noun-phrases or verb-phrases by the pipeline. The most frequent of these are shown in (1-c), and the full list is in the supplementary material. H&P comments that in (2-b) *so* [is] interpreted as *"that the idea was preposterous"*. This is part of the meaning, not part of the text.

4 Recommendations for the use of H&P

- The most important thing for an applied NLP researcher to do with this corpus is to systematically consider all the types of expression in the examples, and form a reasoned opinion about which ones they want to cover. At this point, informed choices can be made about the scope of training data creation and/or linguistic effort.
- The second task is to decide on an implementation. For a neural approach this would include the choice of an architecture and loss function. If the chosen method is not neural, corresponding choices can now be informed by a data-driven understanding of the extent of the desired coverage.

H&P can be of great help in shaping the first step, which is frequently decisive on the quality of the outcome.

4.1 Deeper examples from H&P

In discussing 'Anaphors that are not NPs' H&P provides the following examples:

(3) Anaphors that are not NPs

a. I asked for a green shirt, but he gave me a white one.

b. If you want me to stay on I will do so. *Link is to verb phrase, no pronoun*

c. Liz will complain, or at least I think she will _. *Link is to verb phrase, from embedded sentence with gap*

and provide the following commentary on (3-a):

> ...the anaphoric relation is not between the NPs *a green shirt* and *a white one*, but between the nouns *shirt* and *one*. These are not referring expressions: we understand the antecedent and anaphor here to have the same denotation, not the same reference.

In (3-a) H&P distinguishes between a **referential** link: two expressions are linked to the same object; and a **denotational** link: two expressions are linked to the same concept (here, *shirt*), but not to the same object. Such links are beyond the scope of current co-reference tasks and engines. (3-b) is a clear example of reference to a verb-phrase, and (3-c) shows that links can have complicated phrases at both ends. From a linguist's perspective, examples like this are routine and important. Using them and understanding them is a key part of being a competent language user.

It is obvious that (3-a) should matter to an NLP researcher working in a shopping application. The problem that the customer has described is simple, even though it is beyond the scope of current co-reference engines. Supposedly intelligent systems do not even try to handle this case. Insiders will know that the primary reason for this is a historical accident, but naive users will be surprised, and rightly so.

5 Conclusions

As far as we know, and as far as the authors of H&P know, the present study is the first to systematically extract and use all the examples from a chapter of the grammar. We hope that copyright issues can be resolved and that this work can be extended to the whole of the grammar.

Citation analysis reveals that H&P is influential, especially for sub-communities of ACL working on linguistically complex problems such as negation and anaphora. Error analysis reveals that a representative high-quality co-reference analyzer (Clark and Manning, 2016a) covers less than half of the examples that linguists provided. The main reason for failure is the mention extraction phase, which precedes efforts to disambiguate reference. The main opportunity for progress lies not in varying the way referential links are disambiguated, but in reformulating the problem to account for a wider range of everyday anaphoric relationships. We provide quantitative detail for the Stanford co-reference pipeline, working in English, but, impressionistically, the conclusions would have been similar for any typical co-reference system, such as (Versley et al., 2008; Poesio et al., 2010). In summary, ACL researchers should and can make use of the Cambridge Grammar of the English Language.

6 Societal impacts and Ethical Considerations

The work described in this paper raises few ethical concerns. The authors of H&P have given permission for the use of the data from their chapter, and the quotations in the paper in any case fall within standard definitions of fair use.

References

Anne Abeillé and Danièle Godard. 2010. The grande grammaire du français project. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Jon Alkorta, Koldo Gojenola, and Mikel Iruskieta. 2018. Saying no but meaning yes: negation and sentiment analysis in Basque. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu A. Ha, Rodney Michael Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler C. Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna L. Power, Sam Skjonsberg, Lucy Lu Wang, Christopher Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *NAACL*.

Rajendra Banjade and Vasile Rus. 2016. DT-neg: Tutorial dialogues annotated for negation scope and focus in context. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3768–3771, Portorož, Slovenia. European Language Resources Association (ELRA).

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Chris Brew. 2003. Book review: The Cambridge grammar of the English language by rodney huddleston and geoffrey K. pullum. *Computational Linguistics*, 29(1).

Kevin Clark and Christopher D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2016. Improved parsing for argument-clusters coordination. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 72–76, Berlin, Germany. Association for Computational Linguistics.

Gerhard Helbig and Joachim Buscha. 1998. *Deutsche Grammatik*. Langenscheidt, Leipzig, Deutschland.

Rodney D. Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Anna Kazantseva and Stan Szpakowicz. 2010. Summarizing short stories. *Computational Linguistics*, 36(1):71–109.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Survey: Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3):547–612.

Geoffrey Leech. 2004. A new Gray's Anatomy of English grammar. *English Language and Linguistics*, 8:121 – 147.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Herve Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. TSNLP - test suites for natural language processing. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada. Association for Computational Linguistics.

Marius Mosbach, Stefania Degaetano-Ortlieb, Marie-Pauline Krielke, Badr M. Abdullah, and Dietrich Klakow. 2020. A closer look at linguistic knowledge in masked language models: The case of relative clauses in American English. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 771–787, Barcelona,

Spain (Online). International Committee on Computational Linguistics.

John Nerbonne. 2018. Vaulting ambition. In *Grammar and Corpora 2016*, pages 445–468. Heidelberg University Publishing. Grammar and Corpora, Grammar and Corpora ; Conference date: 09-11-2016 Through 11-11-2016.

Alicia Parrish, William Huang, Omar Agha, Soo hwan Lee, Nikita Nangia, Alex Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve nlu data collection? *ArXiv*, abs/2104.07179.

Massimo Poesio, Olga Uryupina, and Yannick Versley. 2010. Creating a coreference resolution system for Italian. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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

Zahra Sarabi and Eduardo Blanco. 2017. If no media were allowed inside the venue, was anybody allowed? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 860–869, Valencia, Spain. Association for Computational Linguistics.

Lesley Stirling and Rodney Huddleston. 2002. Deixis and anaphora. In *The Cambridge Grammar of the English Language*.

Yannick Versley, Simone Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Amir Zeldes. 2018. A predictive model for notional anaphora in English. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 34–43, New Orleans, Louisiana. Association for Computational Linguistics.