Doing Corpus Pragmatics in DART 2.0 – New and Improved Ways

Martin Weisser (Guangdong University of Foreign Studies, China)

Corpus Pragmatics and corpus-based discourse studies are becoming increasingly important sub-disciplines of Corpus Linguistics, as e.g. evidenced through publications like Aijmer & Rühlemann (2014a), the Yearbook of Corpus Linguistics and Pragmatics series edited by Romero-Trillo, or Baker & McEnery (2015). Yet, in order to investigate meaning in context, most of the research described in these publications still relies on more traditional and basic corpus linguistics methodology – i.e. what Aijmer & Rühlemann (2014b: 8) refer to as the "vertical reading" of concordance lines – which is only suited for very limited, small-scale pragmatic analysis, or on largely unsuitable or imprecise techniques, such as simple keyword or frequency analyses. At least part of the reason for this shortcoming probably lies in what Searle (1963: 136ff.) refers to as "[t]he speech act fallacy", i.e. the mistaken belief that single words may allow us to characterise and/or identify meaning adequately. How-ever, prior attempts to resolve this issue in order to be able to identify contextual meaning, such as basic collocation analysis or even the identification of "functional profiles" (Adolphs 2008: 10) still have not advanced the field of corpus pragmatics enough to make the large-scale analysis of pragmatic meaning possible, not only be-cause they are too limited in scope, but also because they do not leave a clear re-cord of the facts since they do not – as yet – "[…] make explicit the relationship between individual speech act expressions and their distribution across different con-texts", as Adolphs (ibid.) stipulates needs to be done. The only way in which such an endeavour can be realised is to drive forward the creation of pragmatically annotated corpora, and with it, the methodology required for achieving this.

As illustrated in Weisser 2014 & 2017, the first version of the Dialogue Annotation and Research Tool (DART) already presented a major novel way of enriching dialogue data largely automatically with pragmatics-relevant annotations on a number of different levels, thereby taking the potential for genuine corpus-based approaches to the field of pragmatics one step further. The distinct levels covered there comprise syntax (both traditional and extended 'sentence' types), semantics ('topics'), semantico-pragmatics ('IFIDs'; Searle 1969: 16), surface polarity, and pragmatics (in the form of speech acts). The number of potential individual speech acts the first version was able to recognise with a high degree of precision (cf. Weisser 2016a) was 57, some of which could occur in combination. This number already exceeded that of the speech acts employed in most traditional taxonomies, such as those established by Austin (5), Searle (5), as well as those derived from the latter for the annotation of the SPICE Ireland (9; Kallen & Kirk 2012), by far. Even in comparison to the more practice-oriented taxonomies employed in recent NLP-oriented projects, such as the Maptask Corpus (12; Kowtko et al. 1993), DAMSL (31; Allen & Core 1997), or Switchboard DAMSL's "approximately 60 basic tags" (Jurafsky et al. 1997: 1), DART 1.0 already performed rather well, too, especially as the taxonomies implemented there mostly still needed

to be applied manually before allowing computational linguists to devise more or less successful algorithms based on machine learning techniques.

Version 2.0 of DART now supports an even more fine-grained basic taxonomy of more than 120 basic categories and their potential combinations, distinguishing between different types of speech acts as realised through and in different c-unit types, the sequencing of units in dialogue, the influence of modality, polarity, etc. In comparison to the first version, it also features a more robust grammar for recognising different syntactic types, a larger inventory of IFIDs, and an improved inferencing mechanism for deducing speech acts, all based on symbolic, rather than probabilistic identification strategies. The annotations produced in DART thus not only make it possible to achieve the aims pointed out by Adolphs, but also make it possible to carry out further investigations into the form–function relationship embodied in, and expressed through, the different levels, potentially leading to far deeper in-sights into the mechanisms that underlie different communicative strategies, as already illustrated to some extent in Weisser (2016b), where the interactional behaviour of one British and one American call-centre agent was profiled one against the other, as well as against that of their respective callers.

In this talk, I first want to present the design of the new version of DART in terms of the enhancements in its interface and corpus handling features compared to the earlier version. This will then be followed by a brief illustration of the annotation process and analysis options, finally pointing forward to how these features can be exploited for various purposes in research into Corpus Pragmatics.

References

Adolphs, Svenja. (2008). *Corpus and Context: Investigating Pragmatic Functions in Spoken Discourse*. Amsterdam: John Benjamins.

Aijmer, K. and Rühlemann, C. (Eds.). (2014a). *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press.

Aijmer, K. and Rühlemann, C. (2014b). Introduction. In Aijmer, K. and Rühlemann, C. (Eds.). *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press.

Allen, J. and Core, M. (1997) *Draft of DAMSL: Dialog Act Markup in Several Layers*. Available from: <ftp://ftp.cs.rochester.edu/pub/packages/dialog-annotation/manual.ps.gz> (accessed 20 December 2016).

Austin, J. (1962) *How to Do Things with Words*. Oxford: Oxford University Press.

Baker, P. & McEnery, T. (2015). *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Basingstoke: Palgrave Macmillan.

Jurafsky, D., Shriberg, E. and Biasca, D. (1997) *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coder Manual*. Available from: <http://www.stanford.edu/~jurafsky/ws97/ics-tr-97-02.ps> (accessed 20 December 2016).

Kallen, J. and Kirk, J. (2012) *SPICE-Ireland: A User's Guide*. Queen's University Belfast, Trinity College Dublin, and Cló Ollscoil na Banríona.

Searle, John. (1969). *Speech Acts: an Essay in the Philosophy of Language*. Cambridge: CUP.

Weisser, Martin. (2014). Speech act annotation. In Aijmer, K. & Rühlemann, C. (Eds.). *Corpus Pragmatics: a Handbook*. Cambridge: Cambridge University Press. Chapter DOI: 10.1017/CBO9781139057493.005.

Weisser, Martin. (2016a). DART – the Dialogue Annotation and Research Tool. *Corpus Linguistics and Linguistic Theory*, 12(2), pp. 355-388. DOI: 10.1515/cllt-2014-0051.

Weisser, Martin. (2016b). Profiling Agents & Callers: a Dual Comparison Across Speaker Roles and British vs. American English. In Pickering, L., Friginal, E., & Staples, S. (Eds.). *Talking at Work: Corpus-based Explorations of Workplace Discourse*. London: Palgrave Macmillan.

Weisser, Martin. (2017). Corpora. In Barron, A., Gu, Y., & Steen, G. (Eds.). *The Routledge Handbook of Pragmatics*. London: Routledge.