Assessing the Incremental Algorithm: a Response to Krahmer et al.

Kees van Deemter [1] & Albert Gatt [2]
& Ielka van der Sluis [3] & Richard Power [4]

In response to Krahmer et al., we express some reservations regarding the experiment reported in their Letter. Specifically, we observe that their results do not tell us whether the Incremental Algorithm is better or worse than its competitors, and we speculate about implications for reference in complex domains, and for learning from "normal" (i.e., non-semantically balanced) corpora.

In their Letter to the Editor, Krahmer et al. discuss our article "Generation of Referring Expressions: Assessing the Incremental Algorithm" (Krahmer et al. 2012, Van Deemter et al. 2012). In this response, we assess their contribution.

1 Our comparison between algorithms

In a famous article in this journal (Dale and Reiter 1995), Dale and Reiter discussed algorithms that had been proposed for the generation of referring expressions and compared them to a new algorithm, which has come to be known as the Incremental Algorithm (IA). Essentially, their article makes two claims: (1) IA produces referring expressions that are more similar to the ones produced by human speakers than its two main competitors (the Full Brevity and the Greedy algorithm), and (2) IA is computationally more efficient than these competitors. They concluded, essentially, that IA is superior.

Our article details the TUNA experiment, a systematic attempt to assess these two claims. TUNA compared the performance of the IA and its competitors on a corpus of referring expressions that resulted from an experiment in which subjects were asked to refer to objects to an imaginary listener. In a nutshell, we found that IA is not always superior to its competitors. The performance of IA turned out to depend strongly on a parameter not specified by Dale and Reiter, namely the Preference Order (PO) of attributes. By determin-

[1]University of Aberdeen
[2]University of Malta and Tilburg University
[3]Trinity College Dublin
[4]Open University, Milton Keynes

ing in what order the IA examines attributes (e.g., colour, size, etc.), the PO influences the likelihood with which an attribute is included in an expression generated by the IA. Different POs thus lead to different referring expressions. Some POs make IA superior to its competitors, others make it inferior. Crucially, our study suggested that it is difficult to predict whether a given PO will make IA superior to its competitors. And, given that the IA's competitors performed quite reliably, we speculated that, confronted with a new domain for which no corpus is available, practitioners in Natural Language Generation might be wise to use one of these competitors, instead of the IA.

2 Krahmer et al.'s study of small domains

Krahmer and colleagues wondered how difficult it is to find "good" POs for IA. They decided to use one of the methods proposed in our article, namely to determine POs by counting the frequencies of attributes in the TUNA corpus. They did not focus on the corpus as a whole, however, but on tiny parts of the corpus. The TUNA corpus consists of two parts: the *furniture* corpus, consisting of stylised pictures of furniture; and the *people* corpus, consisting of descriptions of photographic portraits. The authors found that, for *furniture*, tiny samples suffice to construct an IA that performs as well as their best-performing IA; for the *people* corpus, the results were more varied, though small samples still tended to perform well.

Their procedure went like this: they computed DICE and PRP scores (metrics that measure how "human-like" the expressions generated by an algorithm are. See our article for definitions) for samples consisting of 1, 5, 10, 20, 30, 40, 50, and (roughly) 150 descriptions, that is, at 8 different levels. Larger samples might be expected to lead to better IAs, because they give a more accurate picture of language use. The authors show, however, that for the *furniture* corpus, a "ceiling" is reached with as few as 5 descriptions; for the *people* corpus, a DICE ceiling is reached at 10 descriptions, whereas a PRP ceiling is reached at 40. The ceiling is defined as the lowest level that did not score significantly worse than the IA associated with the highest level (i.e., level 8).

3 What can one learn from small domains?

We make a number of observations, starting with what we see as the main strength of the Letter.

1 *Where we agree:* To find that, in some domains, a tiny amount of data suffices to construct a well-performing reference production algorithm is genuinely interesting, and reminiscent of calculations showing that complex Bayesian decision making can often be approximated very well through small samples (Vul et al. 2009, and the references contained therein). Let us place these results in context. TUNA's two domains were both tiny (involving just 6 distractors, offered in a neat visual arrangement on a small computer screen). Nonetheless, the *furniture* corpus made it substantially easier to find a version of IA that beats its competitors than the slightly more complex *people* corpus. (Krahmer et al. find analogous differences between the two, with higher "ceilings" for the *people* corpus.) One wonders how difficult it is to find "good" POs for truly large and complex domains, as when we point out a person in a crowd, or when we refer to a place that the speaker and the listener once visited together. We believe that the real challenges for research on reference lies in such "real-world" domains – a position for which we find support in Krahmer and van Deemter (2012), which gives us confidence that we may be able to find common ground with the authors of the Letter.

2 *An important reservation:* Unlike Dale and Reiter (1995), and unlike our own article, the Letter does not compare IA with its competitors. This means that the results of the letter do not speak directly to the question on which our paper focussed, namely whether IA is superior to its competitors. To defend the IA against its competitors, it does not suffice to show that small corpora allow one to choose an IA that is "good": one would have to show that the chosen IA is *better* than these competitors. Although direct comparisons are a bit tricky, because of the way in which the authors decided to apply the algorithms (e.g., treating the TYPE attribute differently from us), it may be illustrative to note that the DICE scores that we found for one of the competitor algorithms, Full Brevity, lie above the relevant figures for IA reported in the Letter.

3 *A quibble about stats:* the definition of Krahmer et al.'s notion of a ceiling rests on non-significance. But non-significance results need to be treated with particular caution. For example, the authors observe that a DICE score of 0.605 (*people* corpus, domains of size 10) was not statistically distinguishable from a DICE score of 0.724 (*people* corpus, domains of size 50), yet an experiment that used a larger number of referring expressions to test the algorithms may well have revealed a difference. Besides, it may

not be justified to treat level 8 as if it was a gold standard; larger data sets may well have yielded even higher scores. These are minor issues.

4 *Scarce resources:* By using the TUNA corpus, Krahmer and colleagues exploit a unique resource, which is semantically balanced. Semantic balance means, for example, that the number of situations in which a given property (e.g., colour) suffices to individuate the referent equals the number of situations where some other property (e.g., size) suffices, and similar for combinations of properties. Semantically balanced corpora tell us something about language use. Non-semantically balanced corpora are much more widely available, and more representative of what humans are exposed to. In the latter, however, frequencies are influenced not just by language use, but also by frequencies in the world. The Letter raises the question whether good POs can also be found on the basis of small corpora that are *not* semantically balanced.

We're still inclined to believe that, confronted with a new domain for which no corpus is available (let alone a semantically balanced one), it might be better to use one of its competitors, instead of the IA. In our article, we argue that factors such as computational complexity do not alter this assessment. Krahmer et al.'s experiment with small data sets is interesting nonetheless, not least because of the questions it raises about human acquisition of reference production.

4 References

Dale and Reiter (1995). R.Dale and E.Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science, 18:233263.

Krahmer and van Deemter (2012) Computational Generation of Referring Expressions: a Survey. Computational Linguistics 38 (1), March 2012, pp.173-218.

Krahmer et al. (2012). E. Krahmer, R. Koolen and M. Theune. Is it that difficult to find a good Preference Order for the Incremental Algorithm? To appear in Cognitive Science (2012).

Van Deemter et al. (2012) K. van Deemter, A. Gatt, I. van der Sluis and R. Power. Generation of referring expressions: assessing the incremental algorithm. To appear in Cognitive Science (2012).

Vul et al. (2009) E. Vul, N.D. Goodman, T.L. Griffiths and J.T. Tenenbaum.

One and done? Optimal decisions from very few samples. In Proceedings of the 31st Annual meeting of the Cognitive Science Society (CogSci-2009).