The D-TUNA corpus: A Dutch dataset for the evaluation of referring expression generation algorithms

Ruud Koolen, Emiel Krahmer

Tilburg University, Department of Communication and Information Sciences
P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands
E-mail: R.M.F.Koolen@uvt.nl, E.J.Krahmer@uvt.nl

Abstract

In this paper, we present the D-TUNA corpus, which is the first semantically annotated corpus of referring expressions in Dutch. Its primary function is to evaluate and improve the performance of REG algorithms. Such algorithms are computational models that automatically generate referring expressions by computing how a specific target can be identified to an addressee by distinguishing it from a set of distractor objects. We performed a large-scale production experiment, in which participants were asked to describe furniture items and people, and provided all descriptions with semantic information regarding the target and the distractor objects. Besides being useful for evaluating REG algorithms, the corpus addresses several other research goals. Firstly, the corpus contains both written and spoken referring expressions uttered in the direction of an addressee, which enables systematic analyses of how modality (text or speech) influences the human production of referring expressions. Secondly, due to its comparability with the English TUNA corpus, our Dutch corpus can be used to explore the differences between Dutch and English speakers regarding the production of referring expressions.

1. Introduction

In everyday communication, speakers often produce referring expressions. Such expressions (for example: ‘the grey chair’) have therefore been studied extensively in research on Natural Language Generation (NLG). NLG is a subfield of Artificial Intelligence and aims to build systems that automatically convert non-linguistic information (e.g. from a database) into coherent natural language text (Reiter & Dale, 2000). Practical applications of NLG include, among others, the automatic generation of weather forecasts (Goldberg et al., 1994; Reiter et al., 2005), and summarization of medical information (Portet & Gatt, 2009).

Given the ubiquity of referring expressions in natural language, it is no surprise that NLG systems typically require algorithms that compute distinguishing descriptions to objects (Mellish et al., 2006). Various Referring Expression Generation (REG) algorithms have been proposed, including the Full Brevity Algorithm (Dale, 1989; 1992), the Incremental Algorithm (Dale & Reiter, 1995; van Deemter, 2002), and the Graph Algorithm (Krahmer et al., 2003). These REG algorithms, each in their own way, compute how a specific target can be identified to an addressee by distinguishing it from a set of distractor objects.

Many REG algorithms aim at generating referring expressions that match human referential behaviour (Dale & Reiter, 1995). Although some of the current REG algorithms generate distinguishing descriptions that are judged to be more helpful and better formulated than human-produced descriptions (Gatt et al., 2009), their applicability is still limited (Krahmer, 2010). Based on several psycholinguistic studies, Krahmer suggests that REG algorithms base the generation of their target descriptions on the wrong psycholinguistic assumptions. For example, while psycholinguistic research shows that human speakers adapt to their addressee when referring (e.g. Clark & Wilkes-Gibbs, 1986; Brennan & Clark, 1996), most current REG algorithms do not take the addressee into account. Furthermore, while human speakers often overspecify their referring expressions and include more information than is strictly needed for identification (e.g. Engelhardt et al., 2006; Pechmann, 1989), none of the current REG algorithms accounts for a systematic way to deal with such referential overspecification.

Given the above limitations, it is important to evaluate the performance of the current REG algorithms, and also to further improve the human-likeness of their generated output. Evaluating REG algorithms often occurs against human corpus data, and these data must be semantically transparent: All expressions need to be provided with information regarding the properties of both the target and the distractor objects. Semantic annotation usually occurs in XML format (Gatt, 2007). This format on the one hand permits the automatic generation of logical forms that correspond to human target descriptions, and on the other hand enables direct comparison of human target descriptions with the generated output of REG algorithms (for example in terms of the selected target attributes).

Until now, only few semantically transparent corpora that can be used for the evaluation of REG algorithms were collected, and they all have limitations. The MAPTASK CORPUS (Anderson et al., 1991) and the COCONUT CORPUS (Di Eugenio et al., 1998) both consist of dialogues between two participants, but the referring expressions that occur in these corpora are rather specific to the kind of task used for collecting them (direction giving and furniture buying). This makes them less suitable for the evaluation of general REG algorithms (Gatt, 2007). This limitation was addressed by the TUNA corpus[1] (Gatt et al., 2007), which consists of English written referring expressions that are annotated in such a way that their underlying semantics is made explicit. However, also the TUNA corpus has some crucial limitations. Firstly, the corpus consists of written referring expressions, while speech is arguably the primary modality of communication. Secondly, the referring expressions were not uttered in the direction of an addressee, which contrasts with everyday communicative situations. Thirdly, the TUNA corpus contains only English referring expressions, which disables the possibility to investigate language differences in the production of referring expressions.

In order to address the limitations of other corpora, we decided to collect the Dutch D-TUNA corpus. In the current paper we describe the collection and annotation of this corpus, and its applications to psycholinguistic and computational linguistic research on the production of referring expressions.

2. Collection of the corpus

In order to collect the D-TUNA CORPUS, we performed a large elicitation experiment in which participants were asked to describe target objects and distinguish them from surrounding objects. This resulted in a corpus of 2400 Dutch referring expressions. Data collection was inspired by the English TUNA experiment (Gatt et al., 2007).

2.1 Participants

Sixty undergraduate students (14 males, 46 females) from Tilburg University participated in the experiment, either on a voluntary basis or for course credit. All participants (mean age 20.6 years old, range 18-27 years old) were native speakers of Dutch.

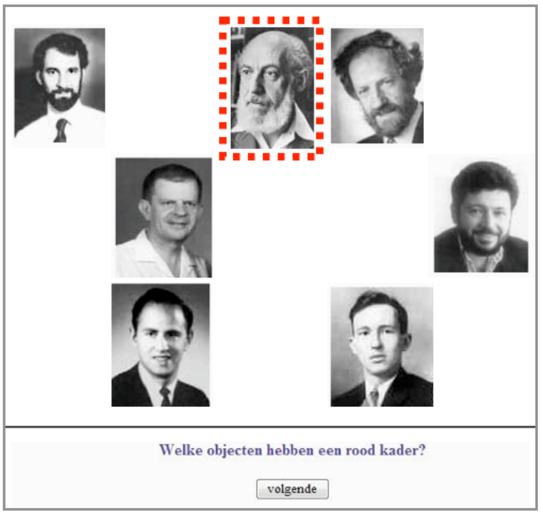
2.2 Materials

The materials consisted of forty trials, which all contained one or more target referents and six distractor objects. The target referents were clearly marked by red borders, so that they could easily be distinguished from the distractor objects.

For each participant and each trial, the target and distractor objects were positioned randomly on the screen in a 3 (row) by 5 (column) grid. In order to manipulate the properties of the target referents, the trials varied in terms of their types of domains and in terms of cardinality.

2.2.1. Two types of domains

A first manipulation of the target properties was that trials occurred in two different types of domains: The furniture domain and the people domain. For an example of a trial in the people domain, see figure 1.



Welke objecten hebben een rood kader?

Figure 1: A trial in the people domain.

The twenty trials in the *furniture domain* contained pictures of four types of furniture items[2]. These items differed along four dimensions (see table 1).

Attribute	Possible values
Type	Chair, sofa, desk, fan
Colour	Blue, red, green, grey
Orientation	Front, back, left, right
Size	Large, small

Table 1: Attributes and values of the pictures in the furniture domain.

The twenty trials in the *people domain* consisted of pictures of male mathematicians. A number of salient dimensions of variation were identified (see table 2).

Attribute	Possible values
Type	Person
Orientation	Front, left, right
Age	Young, old
Hair colour	Dark, light, other
Has hair	0 (false), 1 (true)
Has beard	0, 1
Has glasses	0, 1
Has shirt	0, 1
Has tie	0, 1
Has Suit	0, 1

Table 2: Attributes and values of the pictures in the people domain.

For several reasons, the people domain was the more complex of the two. Firstly, targets in the people domain

[1] The TUNA corpus is distributed by the Evaluations and Language resources Distribution Agency (ELDA). URL: http://catalog.elra.info/product_info.php?products_id=1074

[2] All pictures were taken from the Object Databank, developed by Michael Tarr at Brown University and freely distributed. URL: http://titan.cog.brown.edu:8080/TarrLab/stimuli/objects/

cannot be distinguished in terms of their type (since they all have ‘type = person’). Secondly, the pictures of the persons are arguably more similar to each other than the furniture items, which makes them more difficult to distinguish from the distractor objects. Furthermore, the pictures of people were not as controlled as the artificial pictures in the furniture domain and hence there may be more information in them that participants may use in their references. Last, the possible descriptions of people are somewhat open-ended, in that there are many unpredictable attributes that can be mentioned.

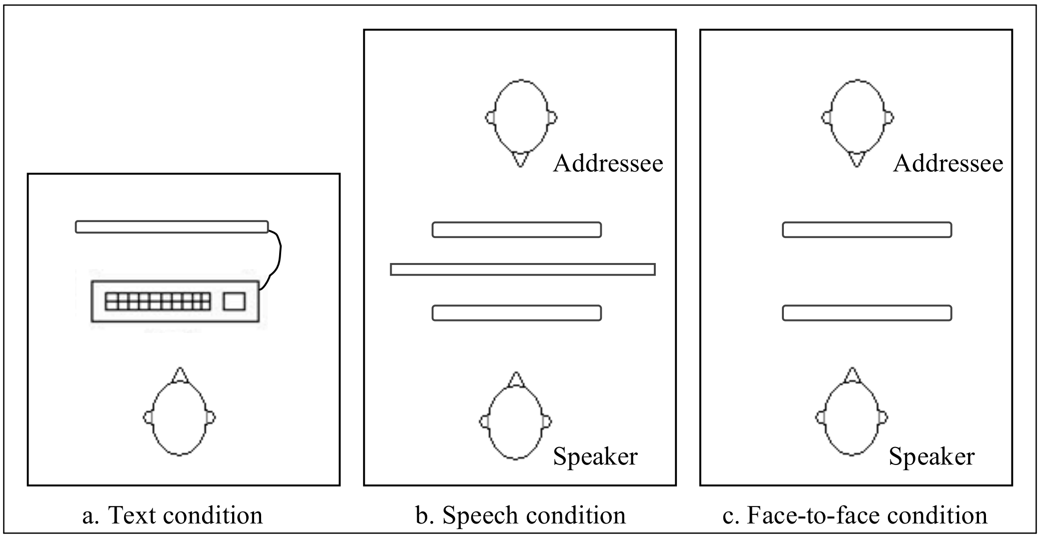
Since speakers need a head noun in their references and therefore always use ‘type’ in their formulation (Levelt, 1989), trials were built in such a way that the attribute ‘type’ could never be a distinguishing attribute.

2.2.2. Two levels of cardinality

A second manipulation of target properties was that trials differed in terms of cardinality, i.e. the number of target referents that they contained. Twenty trials were singular (SG, ten per domain) and contained one target referent. Furthermore, twenty trials (again ten per domain) were plural (PL) trials containing two target referents. An extra manipulation of the target properties occurred by including two levels of similarity. Plural/similar trials (PS, five per domain) trials contained two target objects with both identical distinguishing attributes, for example *‘the table and the sofa that are both red’*, where the two target objects are distinguished from the distractors by means of their (shared) red colour. The plural/dissimilar trials (again five per domain) contained two target objects with different distinguishing attributes, for example *‘the large fan and the red sofa’*, where the two target objects are distinguished by means of different attributes: size and colour.

2.2 Procedure

Each participant was presented the forty trials in a different random order. The experiments were individually performed in an experimental room, with an average running time of twenty minutes. All participants were filmed during the experiment. The participants were asked to describe the target referents in such a way that an addressee could uniquely identify them. In order to manipulate properties of the communicative setting, the participants were randomly assigned to three conditions (text, speech and face-to-face). The *text* condition was a replication (in Dutch) of the TUNA experiment: participants produced written identifying experimental room. In the *speech* condition and the *face-to-face* condition, participants were asked to utter their descriptions to an addressee inside the experimental room. The addressee was a confederate of the experimenter, instructed to act as though he understood the references, but never to ask clarification questions. In the instructions, the participants were told that the location of the objects on the addressee’s screen had been scrambled; hence, they could not use location. In the face-to-face condition, the addressee was visible to the participants; in the speech condition this was not the case, because a screen was placed in between speaker and addressee. A schematic overview of the three conditions is displayed in figure 2a-c.



a. Text condition b. Speech condition c. Face-to-face condition

Figure 2a-c: A schematic overview of the three conditions

```
<TRIAL ID="A03t21" CARDINALITY="1" CONDITION="text" DOMAIN="people" MODALITY="written">
    <DOMAIN>
        <ENTITY ID="54" IMAGE="Eilenberg.jpg" TYPE="target">
            <ATTRIBUTE NAME="hasBeard" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="hasTie" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="type" TYPE="literal" VALUE="person"/>
            <ATTRIBUTE NAME="hasHair" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="hasGlasses" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="hasSuit" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="age" TYPE="literal" VALUE="old"/>
            <ATTRIBUTE NAME="hairColour" TYPE="literal" VALUE="light"/>
            <ATTRIBUTE NAME="orientation" TYPE="literal" VALUE="left"/>
            <ATTRIBUTE NAME="hasShirt" TYPE="boolean" VALUE="1"/>
        </ENTITY>
        <ENTITY ID="4" IMAGE="Fefferman.jpg" TYPE="distractor">
            <ATTRIBUTE NAME="hasBeard" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="hasTie" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="type" TYPE="literal" VALUE="person"/>
            <ATTRIBUTE NAME="hasHair" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="hasSuit" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="hasGlasses" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="age" TYPE="literal" VALUE="young"/>
            <ATTRIBUTE NAME="hairColour" TYPE="literal" VALUE="dark"/>
            <ATTRIBUTE NAME="orientation" TYPE="literal" VALUE="front"/>
            <ATTRIBUTE NAME="hasShirt" TYPE="boolean" VALUE="1"/>
        </ENTITY>
        <ENTITY ID="48" IMAGE="Wall.jpg" TYPE="distractor">
            <ATTRIBUTE NAME="hasBeard" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="hasTie" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="type" TYPE="literal" VALUE="person"/>
            <ATTRIBUTE NAME="hasHair" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="hasSuit" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="hasGlasses" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="age" TYPE="literal" VALUE="young"/>
            <ATTRIBUTE NAME="hairColour" TYPE="literal" VALUE="dark"/>
            <ATTRIBUTE NAME="orientation" TYPE="literal" VALUE="front"/>
            <ATTRIBUTE NAME="hasShirt" TYPE="boolean" VALUE="0"/>
        </ENTITY>
        . . . . .
    </DOMAIN>
    <STRING-DESCRIPTION>
        De man met een witte baard en zonder bril.
    </STRING-DESCRIPTION>
    <DESCRIPTION NUM="singular">
        <DET VALUE="definite">De</DET>
        <ATTRIBUTE ID="a1" NAME="type" VALUE="person">man</ATTRIBUTE>
            met
        <ATTRIBUTE ID="a3" NAME="hasBeard" VALUE="1">een<ATTRIBUTE ID="a2"
        NAME=”hairColour” VALUE=”light”>witte</ATTRIBUTE>baard</ATTRIBUTE>
            en
        <ATTRIBUTE ID="a4" NAME="hasGlasses" VALUE="0">bril</ATTRIBUTE>
    </DESCRIPTION>
    <ATTRIBUTE-SET>
        <ATTRIBUTE ID="a1" NAME="type" VALUE="person"></ATTRIBUTE>
        <ATTRIBUTE ID="a2" NAME="hairColour" VALUE="light"></ATTRIBUTE>
        <ATTRIBUTE ID="a3" NAME="hasBeard" VALUE="1"></ATTRIBUTE>
        <ATTRIBUTE ID="a4" NAME="hasGlasses" VALUE="0"></ATTRIBUTE>
    </ATTRIBUTE-SET>
</TRIAL>
```

Figure 3: Example of an XML file of a reference in the people domain.

2.4 Experimental design

The experiment had a 2x2x3 design (see table 3), with two within-subjects factors: *domain* (levels: furniture, people) and *cardinality* (levels: singular, plural), and one between-subjects factor representing communicative setting: *condition* (levels: text, speech, face-to-face).

	Furniture		People	
	Sing.	Plur.	Sing.	Plur.
Text	200	200	200	200
Speech	200	200	200	200
Face-to-face	200	200	200	200

Table 3: Overview of the experimental design and number of descriptions within each cell.

3. Data annotation

The 2400 (3x20x40) identifying descriptions of the D-TUNA corpus were all semantically annotated using an XML annotation format: they were provided with information regarding attributes of both the target and distractor objects. For this annotation, we used the XML annotation scheme of the TUNA corpus (Gatt, van der Sluis & van Deemter, 2008b).

The annotation tool Callisto[3] was used for the annotation of the expressions. An example of an XML file of a reference to the target shown in figure 1 is depicted in figure 3. In this expression, the target is (in Dutch) referred to as 'De man met een witte baard en zonder bril' (meaning 'The man with the white beard and without glasses').

All XML files consist of a trial node, containing a trial ID and specific conditions under which the expression was produced (such as domain, modality and cardinality). Furthermore, each trial node subsumes four nodes: a domain node, a string-description node, a description node and an attribute-set node.

- The DOMAIN node contains a representation of the domain of the particular trial and consists of seven entity nodes: one or two target entities (depends on cardinality) and five or six distractor entities. Each entity node depicts a list of properties of the particular entity.

- The STRING-DESCRIPTION node contains the full target description, as produced by the participant.

- The DESCRIPTION node contains the annotated version of the target description. All determiners and content words that are part of the string description were provided with the attributes that they represent. For example, the adjective *'witte'* (meaning 'white') corresponds to the attribute <hair colour: *light*>. In case a participant mentioned an attribute that was not present in the domain at all (e.g. 'the laughing man), the attribute 'laughing' was annotated as <other: *other*>.

- The ATTRIBUTE-SET node contains an overview of all properties that are mentioned in the string description and thus represents the flat semantic structure of the referring expression.

4. Applications

The D-TUNA corpus can be used in computational linguistic and psycholinguistic studies on the production of referring expressions.

The D-TUNA corpus is a useful tool in computational linguistic research on the generation of referring expressions, since its semantic annotation in XML format permits using the referring expressions as input for REG algorithms. In line with Gatt et al. (2009), who used the English TUNA corpus to evaluate and compare the performance of several REG algorithms, Theune et al. (2010) used the Dutch references of the D-TUNA corpus as input for the Graph Algorithm (Krahmer et al. 2003).

Since the data collection of the Dutch D-TUNA corpus was inspired by the data collection of the English TUNA corpus, it is possible to explore the differences between Dutch and English speakers regarding the production of referring expressions. For example, Koolen et al. (2010) used the two corpora to compare Dutch and English referring expressions in terms of overspecification. They found roughly similar patterns for references in the two languages regarding which and how many redundant target attributes they contain. In line with Theune et al. (2010), this suggests that our Dutch corpus can be used to train and improve non-Dutch REG algorithms.

Furthermore, the D-TUNA corpus is a useful tool in psycholinguistic research on human referring behaviour. Since it contains both written and spoken references that are produced for an addressee, the D-TUNA corpus enables systematic analyses of how modality (text or speech) influences the human production of referring expressions. For example, Koolen at al. (2009) used the corpus to explore which factors cause speakers to overspecify their referring expressions. They found that references to plural targets uttered in the complex people domain contain more redundant target attributes than references to singular targets uttered in the simple furniture domain. Koolen et al. also found that written and spoken referring expressions do not differ in terms of redundancy, but do differ in terms of the number of words they contain: Speakers need more words to provide the same information as people who type their expressions.

5. Conclusion

We have presented the D-TUNA corpus, which is the first semantically annotated corpus of referring expressions in Dutch. Due to the XML annotation format, the corpus can be used for evaluating and improving the performance of REG algorithms. Furthermore, due to its

[3] URL: http://callisto.mitre.org/

comparability with the English TUNA corpus, our Dutch corpus can be used to explore the differences between Dutch and English speakers regarding the production of referring expressions. Last, the D-TUNA corpus is a useful tool in psycholinguistic studies on human referential behaviour.

6. Acknowledgements

We thank Albert Gatt and Martijn Goudbeek for their technical support during the annotation of the corpus, and three anonymous reviewers for their comments concerning our abstract.

The research reported in this paper forms part of the VICI project 'Bridging the gap between psycholinguistics and computational linguistics: the case of referring expressions', funded by the Netherlands Organization for Scientific Research (NWO Grant 277-70-007).

The D-TUNA CORPUS is publicly available via the TST-Centrale (Dutch HLT Agency: http://www.tst.inl.nl/) of the Nederlandse Taalunie (Dutch Language Union: http://taalunieversum.org/).

7. References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34, pp. 351--366.

Brennan, S. & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 22(6), pp. 1482--1493.

Clark, H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, pp. 1--39.

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics.* University of British Columbia, Vancouver, BC, Canada, pp. 68--75.

Dale, R. (1992). *Generating referring expressions: Building descriptions in a domain of objects and presses.* Cambridge: MIT Press.

Dale, R. & Reiter, E. (1995). Computational interpretation of the Gricean maxims of Quantity in the generation of referring expressions. *Cognitive Science,* 19(8): 233--263.

Di Eugenio, B., Jordan, P., Moore, J. & Thomason, R. (1998). An empirical investigation of proposals in collaborative dialogues. In *Proceedings of COLING-ACL*. University of Montreal, Montreal, Quebec, Canada, pp. 325--329.

Engelhardt, P., Baily, K., Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54, pp. 554--573.

Gatt, A. (2007). *Generating coherent references to multiple entities*. Unpublished PhD thesis, University of Aberdeen.

Gatt, A., Van der Sluis, I. & van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th workshop on Natural Language Generation, ENLG-07.* Dagstuhl, Germany.

Gatt, A., Van der Sluis, I. & van Deemter, K. (2008b). XML formatting guidelines for the TUNA corpus. Technical report, Department of Computing Science, University of Aberdeen, Scotland.

Gatt, A., Belz, A. & Kow, E. (2009). The TUNA-REG challenge 2009: Overview and evaluation results. *Proceedings of the 12th European workshop on Natural Language Generation, ENLG 2009*. Athens, Greece.

Goldberg, E., Driedger, N., Kittredge, R. (1994). Using Natural-Language Processing to produce weather forecasts. *IEEE Expert: Intelligent Systems and their applications,* 9 (2), pp. 45--53.

Koolen, R., Gatt, A., Goudbeek, M. & Krahmer, E. (2009). Need I say more? On factors causing referential overspecification. *Proceedings of the workshop on the Production of Referring Expressions: Bridging the gap between Computational and Empirical approaches to reference (PRE-CogSci 2009).* Amsterdam, The Netherlands.

Koolen, R., Gatt, A., Goudbeek, M. & Krahmer, E. (2010). Overspecification in referring expressions: Causal factors and language differences. Submitted.

Krahmer, E., van Erk, S. & Verleg, A. (2003). Graph-based Generation of Referring Expressions. *Computational Linguistics*, 29(1), pp. 53--72.

Krahmer, E. (2010). What computational linguists can learn from psychologists (and vice versa). *Computational linguistics,* 36(2).

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Mellish, C., Scott, D., Cahill, L., Evans, R. & Reape, M. (2006). A reference architecture for Natural Language Generation systems. *Natural Language Engineering*, 12(1), pp. 1--34.

Portet, F. & Gatt, A. (2009). Towards a possibility -theoretic approach to uncertainty in medical data interpretation for text generation. *Proceedings of the workshop on Knowledge Representation for HealthCare (KR4HC-09).* Verona, Italy.

Reiter, E. & Dale, R. (2000). *Building Natural Language Generation systems*. Cambridge University Press, Cambridge, UK.

Reiter, E., Sripada, S., Hunter, J., Yu, J. & Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence,* 167, pp. 137--169.

Theune, M., Koolen, R. & Krahmer, E. (2010). Cross-linguistic attribute selection for REG: Comparing Dutch and English. Submitted.

Van Deemter, K. (2002). Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1), pp. 37--52.