Audience Design in the Generation of References to Famous People

Roman Kutlak (r04rk9@abdn.ac.uk)

Kees van Deemter (k.vdeemter@abdn.ac.uk)

Chris Mellish (c.mellish@abdn.ac.uk)

Computing Science Department
University of Aberdeen
Aberdeen AB24 3UE
Scotland, UK

Abstract

This paper seeks to fill a gap in existing computational models of the production of referring expressions, by addressing situations in which speakers have difficulty assessing what information is available to their audience. The paper describes a two-part experiment where speakers were given the name of a famous person and had to create a description that would enable a hearer to identify the person, and hearers used the created descriptions to guess the name of the described person. The experiment compares how confident hearers are that they have identified the referent and how well speakers can estimate this confidence. The results of the experiment suggest that speakers do not overestimate hearers' confidence as the psycholinguistic literature had led us to expect.

Keywords: Audience Design; Mutual Knowledge; Reference; Definite Descriptions; GRE

Introduction

Reference production has been investigated in two different research traditions: the psycholinguistic tradition and the computational-linguistics tradition.

Existing *psycholinguistic* research on reference has often focussed on mismatches of information between speakers and hearers. Researchers in this tradition have asked, for example, how well speakers and hearers are able to take such mismatches into account when they produce or interpret referring expressions (Horton & Keysar, 1996; Keysar, Barr, Balin, & Brauner, 2000; Lane, Groisman, & Ferreira, 2006). They have typically done this by putting speakers and hearers in small and cleverly constructed artificial situations, where there are things that speakers can observe, but hearers cannot (or the other way round). The artificiality of these situations has caused some researchers to question the validity of this research (Brown-Schmidt, 2009; Brennan & Hanna, 2009). Nonetheless, the results are extremely interesting and have led to an ongoing debate about the extent to which speakers "design" their utterances to maximise utility for their audience (Krauss & Fussell, 1991; Fussell & Krauss, 1992; Hanna, Tanenhaus, & Trueswell, 2003). The expression *audience design* (also, *perspective taking*) is associated with this issue.

Existing *computational* research on the generation of referring expressions (GRE) has thrived in recent years (Dale & Reiter, 1995; Nenkova, Siddharthan, & McKeown, 2005; Horacek, 2006; Mitchell, van Deemter, & Reiter, 2010). This computational research has focussed on small domains (typically containing less than 10 objects) in situations that were simple enough that speakers and hearers could be guaranteed to have the same information concerning the properties of the objects in the domain. Mismatches in information are therefore seldom addressed in GRE. This has arguably limited the interest and usefulness of these algorithms, because reference in daily life tends to be very different, involving large domains, about which different people have different information. A notable exception to this tradition is the work of Siddharthan & Copestake (2004) whose algorithm was designed to work in open domains.

The aim of the present study was to investigate reference in a situation where mismatches of information between speakers and hearers are normal and natural. Instead of focussing on small artificial situations (as is common in both above-mentioned research traditions), we focussed on large domains that are not directly observed, but remembered. More specifically, we chose to focus on situations where speakers had to describe *famous people* to hearers whom they did not personally know. A similar domain has been used by Nenkova, Siddharthan & McKeown (2005) to infer the cognitive status of a referent. Since these famous people are not directly observed during the experiment, their properties can only be remembered from past experience, and this introduces differences of information between speakers or hearers. We wanted to know how referential behaviour is affected by these differences.

The results of our experiment will inform algorithms that are able to describe people in a way that is likely to benefit hearers. (See our section on Algorithm Implications) Algorithms of this kind can *help readers to digest the news*, for example: the hearer clicks on a proper name (e.g., "Julian Assange"), whereupon the system responds with a description (e.g., "The founder of Wikileaks", or "A former journalist currently awaiting trial on charges of sexual misconduct"). This should help the hearer to know who the proper name refers to. The usefulness of such algorithms along with an example of such a system is also described in Radev & McKeown (1998).

Audience Design

Grice's maxim of quantity states that speakers should make their contribution as informative as is required but should not include more information than is required. Several researchers have pointed out that this kind of contribution requires the speaker to take the perspective of the hearer into account (Krauss & Fussell, 1991; Nickerson, 1999). As has been noted, one of the requirements of taking other's perspectives is to estimate other's knowledge relevant to the conversation (Clark & Marshall, 1981; Krauss & Fussell, 1991; Fussell & Krauss, 1992; Horton & Keysar, 1996).

Fussell & Krauss (1992) report experiments that focus on referring expressions and audience design. In their first experiment, participants were presented with pictures of men prominent in business, politics, or entertainment. The participants were asked to rate how identifiable the stimulus person was on a 7-point scale (from *not identifiable* to *very identifiable*) for themselves and for other students. The participants were also asked to provide the name of the stimulus person if they knew it. The identifiability of stimuli was defined as the likelihood of knowing the stimulus person's name. The experiment showed that the participants could judge reasonably well the knowledge of others. The data also showed a bias in the direction of the participants' knowledge. In other words, when a participant knew the name of the target person, he or she believed that a higher proportion of people than in reality would know the name. Similarly, when the participant did not know the name of the target person, he or she believed that a smaller proportion of people would know the person's name. A similar pattern emerged from another experiment, where participants were estimating the proportion of students knowing a name of an everyday object. "Even items that were identified by 10% or less of the subjects were estimated by those who knew its name to be identifiable to 40-80% of the population" (Fussell & Krauss, 1992).

Another line of research has shown that speakers tend to overestimate the effectiveness of their communication. For example, Keysar & Henly (2002) presented speakers with ambiguous sentences, explained the ambiguity to the speakers and asked them to read such sentences to hearers. The experimenters were hoping that the speakers would try to disambiguate the sentence meaning using prosody whenever that was possible. The speaker was then asked to assess the hearers understanding. Similarly, the hearer was asked to select which meaning he or she believed the speaker intended, and indicate his or her certainty on a 5-point scale (1 = *very uncertain*, 5 = *very certain*). The results showed that speakers overestimate their ability to disambiguate (i.e., their estimate of a hearer's certainty was higher than the speaker's actual certainty). Both speakers' overestimate of hearers' knowledge and speakers' overestimate of their effectiveness motivate our hypothesis:

H_1 Speakers are more confident that the hearers will identify the referent given their description than hearers.

Experiment Design

The experiment was conducted online. Participants were presented with a website that described the experiment setup and the two available tasks. Although we did not anticipate an influence of one task on the other one, participants were asked to avoid doing both tasks or indicate in the comments which task they had done first. Only the data from the first task attempted by a participant were then used in the analysis.

The first task was describing famous people and the second task was guessing the name of a famous person given a description of such a person. In each of the tasks we collected the description or the guessed name and asked the participants to rate to what extent they agree with 3 additional statements. For each of the presented statements participants could select one of the following agreement options:

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly disagree

Each of the tasks allowed for any additional comments. The website also provided an introduction that informed the participant about the nature of the tasks and what kind of descriptions we were interested in. In particular, we required participants to provide a description of a famous person that would enable "a general reader" to identify the person given the description.

Describing

Each participant (speaker) in the description task was presented with the name of a famous person. The participant could skip the person if he or she was not comfortable with creating a description for the particular person (e.g., did not know the person). When the participant decided to write a description for the presented person, he or she also addressed the following three statements:

S_a I think a general reader will know who I mean

S_b I know several people of that description

S_c I am sure about the facts in my description

Naming

Each participant (hearer) in the naming task was presented with a description of a famous person. The participant could skip the description if he or she did not want to guess the name of the person. When the participant decided to guess the name of the described person, he or she also addressed the following three statements:

H_a I am sure I know who this description refers to

H_b I am sure the name I provided is correct

H_c I am sure about the facts in the description

The reason for having two very similar questions, H_a and H_b, was to address the tip-of-the-tongue experience (Brown, 1991) where a person can not recall a particular name or a word despite knowing it. H_b was not used in the analysis and is not further discussed in this paper.

Results

The experiment produced two related datasets; one with descriptions and one with names. Each dataset was cleaned by removing non-native participants and descriptions that were not seen by any hearer. Whenever participants performed both tasks, only the data from the first task were used.

From the 34 native speakers (21 females, 12 males and one not stated) only 29 native speakers (17 females, 11 males and 1 not stated) produced descriptions that were viewed by native hearers (11 males, 7 females). The total number of descriptions and corresponding name guesses were 215 and 261 respectively. The speakers produced on average 7.4 (sd 5.3) descriptions and the hearers named on average 14.5 (sd 10.7) descriptions.

One problem that emerged during the analysis was the treatment of descriptions that were viewed but the hearer did not guess the name of the famous person. We could either discard the unsuccessful name guesses and the corresponding descriptions or treat the unsuccessful name guesses as valid guesses with *the lowest rating* (i.e., rating 1, see below) for each of the statements. Both of the approaches seemed valid so we analysed both sets.

The set that included the answers where the participant did not guess the name of the described person was labeled A. The set B contains only the answers where the hearer guessed the name of the described person. The set A has 215 descriptions and 261 name guesses (hearers did not guess the name of the described person in 47 cases) and set B has 180 descriptions and 214 names. The reduced number of descriptions in set B is the result of removing descriptions where the hearers did not guess the name of the described person. The hearers incorrectly identified the referent 56 times (21%) in the set A (this includes the cases where the hearer viewed a description but did not guess the name of the described person) and 9 times (4%) in the set B.

The agreement levels were converted into numerical values to allow analysis. The levels *strongly agree* to *strongly disagree* were assigned values 5 to 1 where 5 denoted *strongly agree* and 1 denoted *strongly disagree*. All calculations were performed using the R statistical package (R Development Core Team, 2010). We have used the Mann-Whitney U test to accommodate for ordinal values and non-normal distributions.

Pre-hypothesis Tests

A number of pre-tests were performed to make sure that our experiment was measuring what we intended it to measure. Given the experimental setup, for example, one would expect that a speaker is more confident about the truth of the facts in the descriptions that he or she produces than hearers are about the descriptions produced (after all, if a speaker includes an incorrect fact but believes it to be correct, his or her confidence in such fact can still be high.). This expectation was borne out by our findings, as the following analysis shows:

Set A The median [quartiles] rating for speakers and hearers respectively are 5 [4, 5] and 4 [3, 5]. The ratings of certainty about the facts are significantly higher for the speakers (Mann-Whitney U = 37353, n1 = 215, n2 = 261, $p < 0.01$ one-tailed). The graph in figure 1 below shows the percentages of answers corresponding to individual confidence levels.

Set B The median [quartiles] rating for speakers and hearers respectively are 5 [4, 5] and 5 [4, 5]. The ratings of certainty about the facts are significantly higher for the speakers (Mann-Whitney U = 22495, n1 = 180, n2 = 214, $p < 0.01$ one-tailed). The graph in figure 2 shows the percentages of answers corresponding to individual confidence levels.

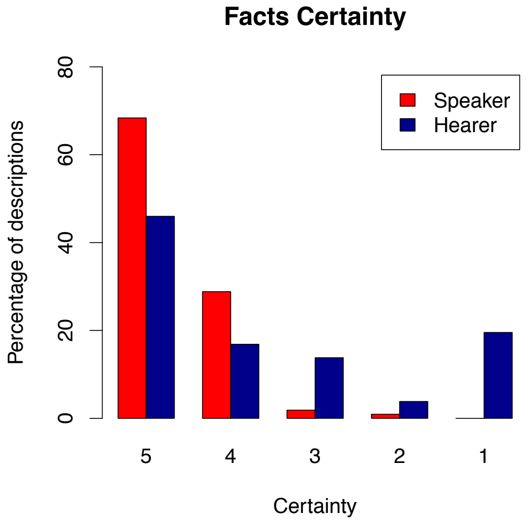

Figure 1: The figure shows the confidence in facts. The data represents set A.

Having performed a number of sanity tests, we proceeded to test our main hypothesis, H1.

H_1: Speakers are more confident that the hearers will identify the referent given their description than hearers

Set A The median [quartiles] rating for speakers and hearers respectively are 4 [4, 5] and 5 [3, 5]. The confidence of correct identification of the referent estimated by the speakers is not significantly bigger than the confidence of hearers (Mann-Whitney U = 27426, n1 = 215, n2 = 261, $p > 0.5$ one-tailed). The graph in figure 3 shows the percentage of answers.

Set B The median [quartiles] rating for speakers and hearers respectively are 4 [4, 5] and 5 [5, 5]. The confidence of

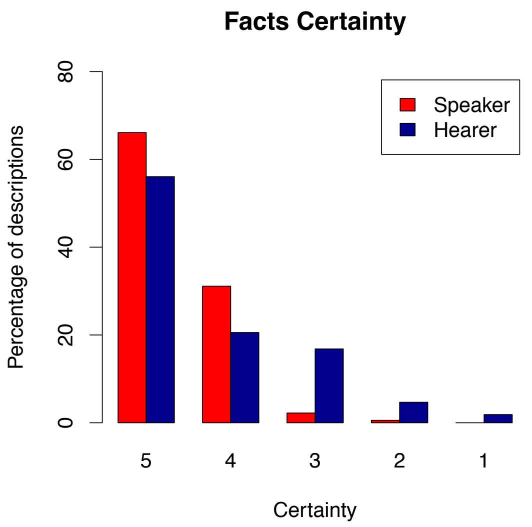

Figure 2: The figure shows the confidence in facts. The data represents set B.

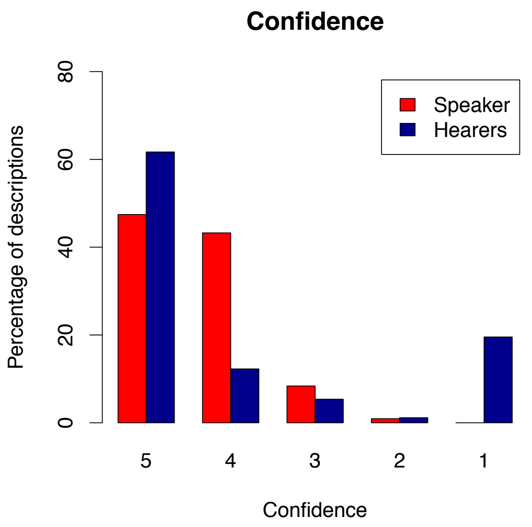

Figure 3: The figure shows the confidence in the identification of the referent as estimated by speakers and as given by hearers. The data represents set A

correct identification of the referent estimated by the speakers is not significantly bigger than the confidence of hearers (Mann-Whitney U = 14748, n1 = 180, n2 = 214, p = 1 one-tailed). The graph in figure 4 shows the percentage of answers.

The result for set B is interesting and it suggests that the converse of the tested hypothesis might be true and that hearers are in fact more confident than speakers. One possible reason for this result is the difference between the actual hearers and the fictitious hearers assumed by the speakers. The speakers were instructed to assess how confident they are that a *general reader* will identify the person whereas the participants for the experiment were recruited through the *linguist list*, an email list for academics, who are likely to be more knowledgeable than a general reader.

A second explanation for this seemingly contradictory result might be the difference in the task at hand and the tasks used by other researchers. Participants in our experiment were estimating hearers' confidence in identifying the famous person. Participants in, for example, Fussell & Krauss (1992) were estimating the percentage of population that would recognise a famous person upon seeing that person regardless of the confidence with which such identification occurs. It thus might be the case that speakers overestimate in one direction (e.g., estimating the commonality of a particular knowledge) and underestimate in other direction (e.g., confidence of identification).

Algorithm Implications

Traditionally, GRE algorithms take as an input the referent and a description of the domain. The algorithm then determines which properties are true of the referent and composes them into an expression that is true of the referent but not of any other object in the domain. The algorithms also employ a mechanism that determines when to stop adding properties. This stop condition is usually triggered when only the referent matches the description. Furthermore, many algorithms assume that the knowledge base contains only information known to the hearer or that the algorithm can determine whether hearer knows a particular fact and avoid the inclusion of facts that are not known to the hearer.

A good example of this standard approach to GRE is the Incremental Algorithm (IA) of Dale and Reiter (Dale & Reiter, 1995). Simplifying considerably, this algorithm operates by addressing the various properties available to the generator one by one, always including the property in the description if it is true of the target referent and false of at least one other object in the domain. The IA uses no backtracking and stops once the target referent is the only domain object of which all the properties included in the description hold true (or if there are no properties left, in which case no referring expression is generated). Additionally, the algorithm makes sure that the description contains a property expressible as a noun; if no such property is included by the mechanism outlined above, one is added at the end of the algorithm. Although the IA will often produce descriptions that are slightly longer than logically necessary (i.e., they are not the shortest identifying descriptions possible), the descriptions generated are always very short. A good way to understand the IA is as a computationally tractable approximation of the idea of generating the shortest identifying description possible (cf. the Gricean maxim of Brevity, (Grice, 1975)). Other algorithms, such as the Greedy Algorithm (Dale, 1992), can be seen in the same light.

Given that our domain consisted of famous (i.e., widely known) people, it might be thought that speakers and hearers would mostly agree on the facts in the domain, but the dif-

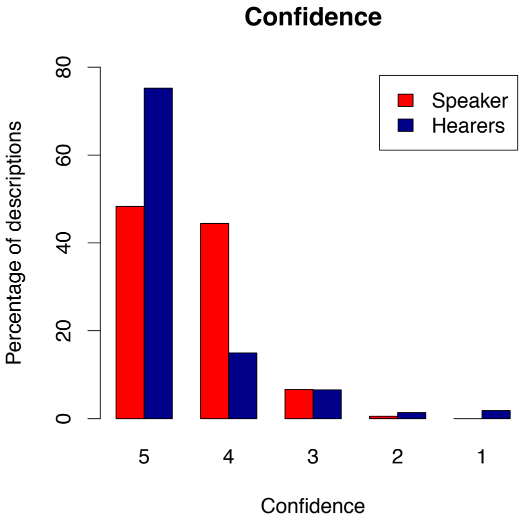

Figure 4: The figure shows the confidence in the identification of the referent as estimated by speakers and as given by hearers. The data represents set B

ference in certainty between speakers and hearers about the facts was statistically significant (statements S_c and H_c). A GRE algorithm for this domain will thus have to be robust against differences in knowledge between speakers and hearers. What might seem to be an appropriate description for one person might be unintelligible for another.

Closer examination of the comments offered by hearers showed that they are very tolerant of the information speakers provided. It seems that hearers can not only accept descriptions that contain additional information previously unknown to hearers but also correctly interpret facts on which the hearers do not agree with the speakers. The text below shows three examples where hearers were not certain about the facts in the description but managed to correctly identify the referent. The lines S1, S2 and S3 show speaker produced descriptions and lines H1, H2 and H3 show comments left by hearers who viewed the corresponding description.

S1 *This person is/was the inventor of the telephone.*

H1 There is some dispute.

S2 *This person is/was the Cambridge Professor of Theoretical physics with Lou Gehrig's disease.*

H2 Didn't know he had Lou Gehrig's disease, just knew he had some degenerative illness.

S3 *This person was credited with the invention of the telephone. He was also interested in flight and assisting the deaf.*

H3 I'm just responding to the "telephone" prompt, basically. I have no idea about the other info.

There were 14 cases where the hearer disagreed or strongly disagreed with the facts provided in the description (statement H_c) but only in 3 out of 14 cases did hearers incorrectly identify the referent.

Given that most contemporary GRE algorithms favour short descriptions and avoid information unknown to the hearer, they are unable to produce the above mentioned successful descriptions. Engelhardt, Bailey & Ferreira (2006) showed that hearers do not judge over-specified expressions worse than concise ones and our hearers' comments suggest that such expressions can be beneficial. We propose that algorithms that generate referring expressions where the knowledge can not be assumed to be shared by the speaker and the hearer should include additional information to allow hearers to correctly identify the referent even if they differ over some of the facts. If an incremental approach to GRE is chosen (as in the IA of Dale and Reiter) It would also make sense to let the algorithm inspect properties in order of their familiarity, instead of their discriminatory value (as in the Greedy Algorithm). One computationally feasible way in which the familiarity of a property might be assessed might be to link this to the frequency of the property in a large corpus of text (Sluis, Gatt, & van Deemter, 2007).

Our results have implications for algorithm testing as well. Normally, GRE algorithms are tested by comparing their output to a corpus of human-produced descriptions (Passonneau, 2006; Jordan & Walker, 2005; Belz & Gatt, 2007). Normally such a corpus includes descriptions by all speakers, regardless of their confidence. In situations like the ones we studied, where human-generated descriptions are at risk of being misinterpreted, one possible approach is to compare the output of the algorithms to the descriptions produced by speakers with high confidence only. It seems reasonable to expect that this will help hearers to identify the described referent, but this is an assumption that would need to be tested.

Conclusion

This paper described a two-part experiment in which speakers described famous people, and hearers attempted to guess the name of the described persons from these descriptions. We were interested in finding out how well speakers can judge the success of their own descriptions, and how confident hearers are that they have identified the referent. We were also interested in how confident speakers and hearers are about the facts in the descriptions. The results did not confirm that speakers overestimate the confidence with which hearers identify the referent (relevant statements S_a and H_a in the Experiment Design section). The results suggests that hearers are more confident about the correct identification of the referent than speakers estimate, but this would have to be tested in a separate experiment. We also found that hearers are less certain of the facts provided by the speakers (relevant statements S_c and H_c in section Experiment Design). We interpreted this as a disagreement between speakers' and the hearers' knowledge and we suggest that GRE algorithms should take this disagreement into consideration.

The comments provided by hearers strongly indicate that their identification can be successful even when hearers do not agree with all of the facts in a description. The hearers successfully identified the described person in 11 out of 14 cases where they did not agree with the facts used in the description.

The main implications for the design of computational models are the following. In domains where there is a possibility of mismatch between the knowledge of a speaker and the knowledge of a hearer, the algorithms should generate descriptions that are robust enough to allow hearers to identify the referent even in cases where such mismatch occurs. In particular, we suggest that the GRE algorithms should intentionally over-specify the referring expressions in cases where there is a risk of disagreement between the speaker's and the hearer's knowledge.

Acknowledgments

We would like to thank the reviewers and the University of Aberdeen NLG group for their valuable comments. This research is sponsored by SICSA.

References

Belz, A., & Gatt, A. (2007). The attribute selection for gre challenge: Overview and evaluation results. In *Proc. 2nd ucnlg workshop: Language generation and machine translation (ucnlg+mt)* (p. 75-83). Citeseer.

Brennan, S., & Hanna, J. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2), 274–291.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, *109*(2), 204 - 223.

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of memory and language*, *61*(2), 171–190.

Clark, H. H., & Marshall, C. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge University Press.

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

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. In *Cognitive science* (Vol. 19, pp. 233–263).

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, *54*(4), 554 - 573.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, *62*(3), 378 - 391.

Grice, P. (1975). Logic and conversation. *Syntax and Semantics*, *3*, 43–58.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43 - 61.

Horacek, H. (2006). Generating references to parts of recursively structured objects. In *Inlg '06: Proceedings of the fourth international natural language generation conference* (pp. 47–54). Morristown, NJ, USA: Association for Computational Linguistics.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91 - 117.

Jordan, P. W., & Walker, M. A. (2005). Learning content selection rules for generating object descriptions in dialogue. *J. Artif. Int. Res.*, *24*(1), 157–194.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32-38.

Keysar, B., & Henly, A. S. (2002). Speakers' overestimation of their effectiveness. *Psychological Science*, *13*(3), 207-212.

Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, *9*(1), 2–24.

Lane, L. W., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants! *Psychological Science*, *17*(4), 273-277.

Mitchell, M., van Deemter, K., & Reiter, E. (2010). Natural reference to objects in a visual domain. In *Proceedings of the 6th international natural language generation conference* (pp. 95–104). Stroudsburg, PA, USA: Association for Computational Linguistics.

Nenkova, A., Siddharthan, A., & McKeown, K. (2005). Automatically learning cognitive status for multi-document summarization of newswire. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 241–248). Morristown, NJ, USA: Association for Computational Linguistics.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737 - 759.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proc. 5th international conference on language resources and evaluation (lrec-06)* (pp. 831–836).

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (ISBN 3-900051-07-0)

Radev, D. R., & McKeown, K. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, *24*(3), 469-500.

Siddharthan, A., & Copestake, A. (2004). Generating referring expressions in open domains. In *Acl '04: Proceedings of the 42nd annual meeting on association for computational linguistics* (p. 407). Morristown, NJ, USA: Association for Computational Linguistics.

Sluis, I. van der, Gatt, A., & van Deemter, K. (2007). *Evaluating algorithms for the generation of referring expressions using a balanced corpus.*