G-TUNA: a corpus of referring expressions in German, including duration information

David M. Howcroft and **Jorrig Vogels** and **Vera Demberg**
Department of Language Science and Technology
Saarland Informatics Campus, Saarland University
66123 Saarbrücken, Germany
{howcroft,jorrig,vera}@coli.uni-saarland.de

Abstract

Corpora of referring expressions elicited from human participants in a controlled environment are an important resource for research on automatic referring expression generation. We here present G-TUNA, a new corpus of referring expressions for German. Using images of furniture as stimuli similarly to the TUNA and D-TUNA corpora, our corpus extends on these corpora by providing data collected in a simulated driving dual-task setting, and additionally provides exact duration annotations for the spoken referring expressions. This corpus will hence allow researchers to analyze the interaction between referring expression length and speech rate, under conditions where the listener is under high vs. low cognitive load.

1 Introduction

Referring expression generation (REG) is an important problem in natural language generation (Dale, 1989; Dale and Reiter, 1995; van Deemter, 2002; Krahmer et al., 2003). The challenge of generating referring expressions (REs) that can pick out a specific object among a set of similar objects represents a prominent subtask in REG. One important goal for REG is to produce human-like referring expressions which are not only logically correct but also sound natural to native speakers of the target language.

Corpora that include referring expressions and contain transparent semantic annotation are an important resource for being able to evaluate the naturalness of an REG algorithm. Because naturally occurring corpora vary wildly with respect to domain and genre, van Deemter et al. (2006) proposed the systematic construction of the TUNA corpus of REs by eliciting them from human subjects in a controlled setting. In their experiment, participants wrote descriptions of target objects in a scene of similar distractor objects and were told that they were interacting with a computer. While this provided the first systematic collection of REs written by humans that could be used to evaluate REG algorithms, Koolen & Krahmer (2010) argued that the written modality was not natural enough to be representative of typical language use. Indeed, most language use is spoken and involves an interlocutor, so when they collected the D-TUNA corpus of REs in Dutch, Koolen & Krahmer included two spoken-language conditions: one where the interlocutor was visible to the speaker and one where they were not.

Including a human addressee was a marked improvement over the text-only modality of the TUNA corpus, and showed that REs were longer on average and more overspecified (although the latter not significantly) in the spoken modality. However, the addressee was a confederate of the experimenters and explicitly instructed to give no feedback to the speaker. This is problematic as speakers usually expect a reaction from their listeners, and so a neutral interlocutor is unrealistic. It is also difficult to emulate the reactions of a naive subject throughout many experimental sessions. In addition, the D-TUNA corpus does not include any acoustic information about the recorded speech, such as RE duration, which is important to be able to assess reduction processes in human language production, and identify possible trade-offs between referring expression length and speaking rate.

Proceedings of The 10th International Natural Language Generation conference, pages 149–153,
Santiago de Compostela, Spain, September 4-7 2017. ©2017 Association for Computational Linguistics

2 Corpus Collection

We built a corpus of spoken referring expressions, aimed at investigating how speakers accommodate listeners who are under cognitive load. We created a more natural speech environment by having pairs of naive participants describe objects to each other in a simulated driving context, while retaining the same collection of furniture images used in both previous TUNA corpora. We did not use the people domain (Gatt et al., 2007; Koolen and Krahmer, 2010), because previous analyses on the TUNA corpora made clear that this domain results in a large amount of linguistic variation that is hard to capture by a semantic annotation. In addition, the detail in the images did not show up well in the driving simulator.

Furthermore, the fact that this corpus was collected in German means that it provides a third language for cross-linguistic comparison. There are two other corpora associated with the analysis of REs in German, namely the GIVE-2 corpus (Gargett et al., 2010) and the PENTOREF corpus (Zarrieß et al., 2016). However, the virtual environment in which the data for the GIVE-2 corpus were collected, coupled with the freedom subjects had to move around the environment, makes the corpus poorly suited to the sort of systematic evaluation of REs that is enabled by the TUNA, D-TUNA, and now the G-TUNA corpora. While the PENTOREF corpus provides data for both English and German accompanied by utterance-level timing information, the task-oriented dialogue with feedback from 'Instruction Followers' and the different target objects make comparisons to the TUNA corpora more challenging. Our corpus thus provides a testing ground for evaluating referring expression generation algorithms for German in a similarly controlled context to the TUNA and D-TUNA corpora in a more natural spoken language context without introducing the further confounds of collaborative dialogue.

2.1 Participants

Twenty pairs of Saarland University students participated in our experiment, with mean age 23.0 (*SD*=4.1). Twenty-one participants were women and the rest were men. We paid the students 10 euros each for their participation.

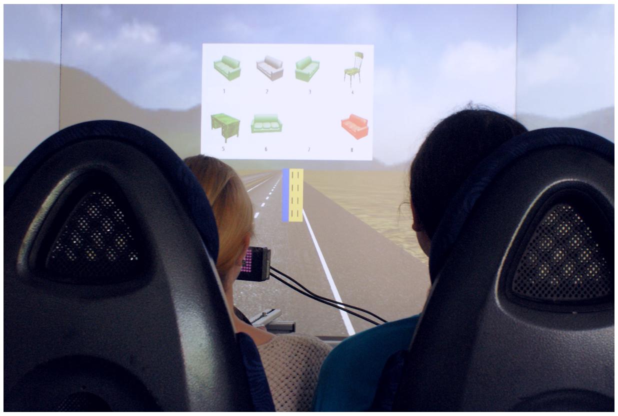

Figure 1: Driving simulator with 2 subjetcs and a stimulus. From left-to-right, top-to-bottom, the stimulus depicts a green sofa, a grey sofa, a green sofa, a green chair, a green desk, a green sofa, a blank space, and a red sofa. #1 was the target.

2.2 Materials

The stimuli were based on those used for the TUNA and D-TUNA corpora. They consist of scenes containing 7 images in a 2×4 grid, where each grid position is numbered 1-8.[1] The target image was identified for the speaker by a number appearing on a separate display not visible to the driver.

The images in a scene are furniture, taken from the Object Databank[2]. The image set in this domain is highly systematic, consisting of four different object types (chair, sofa, desk, fan) in four different colors (blue, red, green, grey), three different orientations (front-, left-, right-facing)[3], and two different sizes (large, small).

The scenes were constructed so that different numbers of modifiers are necessary for the listener to pick out the correct referent with a minimal description (MD). We systematically varied the length of the minimal description for each target, so that objects could be uniquely identified by mentioning 0 (i.e. mentioning only the object type), 1, 2, or 3 attributes. For example, in Figure 1, image #1 is identifiable by the attributes 'orientation=left' and 'color=green', allowing for descriptions like "Das

[1] We used a different grid size than in the previous TUNA experiments, which used a 3×5 grid, in order to accommodate the presentation of the images in the driving simulator.

[2] Available from: `http://wiki.cnbc.cmu.edu/Objects`

[3] We removed the backward-facing objects from the original image set as it was difficult to distinguish them from the forward-facing objects in the driving simulator.

grüne Sofa, das nach links zeigt"[4]. Each image appeared at most once as the target referent.

We created two lists of 60 items, each comprising two blocks of 30 items, such that each participant in a pair would describe different items in their role as speaker. Most items on a list required either one (18 trials) or two (26 trials) attributes to be mentioned. Each block began with 4 practice trials, one for each length of minimal description.

2.3 Procedure

Pairs of participants performed a referential communication task in a driving simulator A coin toss determined which participant in each pair was assigned to the role of speaker first. Speakers were instructed to describe the target referents in such a way that the listener-driver could identify the correct object from an array displayed on the driving simulator screen. They had 15 seconds to provide their description and were told that they were not allowed to use the image's location as a cue. All speech was recorded. To keep speakers aware of the listener's cognitive state, they were prompted to assess the driver's degree of cognitive load after every 10 trials.

The listener-drivers were instructed to verbally respond with the number of the object that they believed was described. They were allowed to ask for clarification if the description was not clear. During the identification task, drivers were either holding the steering wheel stationary (EASY driving condition) or steering to follow a moving object on the road (HARD driving condition). We refrained from the use of a confederate for the listener role, because experience with the dual task may decrease cognitive load, and visibly performing the role of a driver under increased cognitive load was expected to be too difficult for reliable results. Since both participants played both roles, an additional factor was whether the subject played the role of the driver first or the speaker first. A complete experimental session took about 1.5 hours.

3 Corpus Format and Statistics

3.1 Format

The corpus uses the same XML format as the earlier TUNA corpora to facilitate comparisons (Gatt et al., 2008). This annotation scheme includes information about the target image and distractors along with the transcribed referring expression and a flat semantic representation of the mentioned attributes.

We supplement the annotation scheme with duration information for each trial to facilitate more detailed analyses. So far results are mixed as to whether speakers vary word durations based on communicative setting (Bard et al., 2000; Galati and Brennan, 2010), so it is important to provide this information for studies on accommodation, in addition to variation in the number of words used and the degree of over- or under-specification.

3.2 Statistics and Comparison to other Corpora

The current version of the G-TUNA corpus contains data from 40 native speakers of German, each of which completed 60 trials as a driver and 60 trials as a speaker. This resulted in 2331 descriptions after removing problematic items[5], which is comparable to the other two corpora. Table 1 compares the three TUNA corpora on their main properties.

Out of all referring expressions, 45.5% were overspecified, 51.4% were minimally specified, 1.9% were underspecified, and 1.2% were wrongly specified (e.g. mentioning color and orientation where color and size were required). These figures are similar to those found for the TUNA and D-TUNA corpora (Koolen and Krahmer, 2010; Koolen et al., 2011), confirming that referential overspecification is ubiquitous for these items in German as well as in English and Dutch. The rate of overspecification was very similar between the EASY (*M*=0.53; *SD*=0.70) and HARD (*M*=0.56; *SD*=0.72) driving conditions, which is also in line with the findings for D-TUNA that the communicative situation does not affect the degree of overspecification.

An important addition in G-TUNA is the duration annotation for the descriptions. Both the average duration of referring expressions and the number of words showed an influence of the task manipulation. Referring expressions were significantly shorter on

[4] English: The green sofa facing left

[5] We removed items where subjects described the wrong item, identified the target by number, took too long to respond, or made reference to earlier trials as well as items which involved experimental errors, interruptions, etc. Any additions triggered by listener feedback were also removed.

	TUNA	D-TUNA	G-TUNA
# subjects	45	60	40
language	English	Dutch	German
# trials	20	40	60
grid size	3 × 5	3 × 5	2 × 4
# targets/grid	1–2	1–2	1
# distractors/grid	6	6	6
communicative situation	human-computer	no v. invisible v. visible addressee	driver & passenger in driving simulation
modality	written	written + spoken	spoken
domains	furniture, people	furniture, people	furniture
# comparable REs / total	420 / 2280	400 / 2400	2331 / 2331

Table 1: Comparison table for the three versions of TUNA released so far. The 'comparable' RE counts are based on domain & cardinality matches. The TUNA corpus' REs are all in the textual modality, while the 400 D-TUNA REs listed here are in the spoken modality as in our experiments. There are an additional 200 textual furniture REs in the D-TUNA corpus as well.

average in the HARD condition than in the EASY condition, but only for those speakers that had already experienced the driving task themselves (5.9 words / 2464 ms vs. 6.3 words / 2687 ms). This suggests that speakers do adapt some aspects of their descriptions to the communicative situation.

As shown in Figure 2, our corpus provides a balanced middle ground between the TUNA and D-TUNA corpora with respect to description length. Here we observe that the D-TUNA descriptions are often longer, involving longer full sentences, where the time pressure of our experimental setting encouraged subjects to use shorter utterances on average. At the same time, our utterances are longer than the TUNA expressions on average, perhaps due to the difference in modalities as well as the difference in language. That differences between the corpora are already visible with such a coarse analysis suggests that there are many more interesting nuances available to study while accounting for differences in modality and presentation as appropriate.

4 Conclusion

We presented a German corpus of referring expressions, designed to examine listener accommodation in an image identification task where listeners are under cognitive load. The stimuli and design were crafted to make the corpus comparable to the existing TUNA and D-TUNA corpora of referring expressions in English and Dutch. Moreover, we extended our annotations to include word durations, enabling us to evaluate more nuances of speaker adaptation, and to investigate the relationship between referring expression length and speech rate.

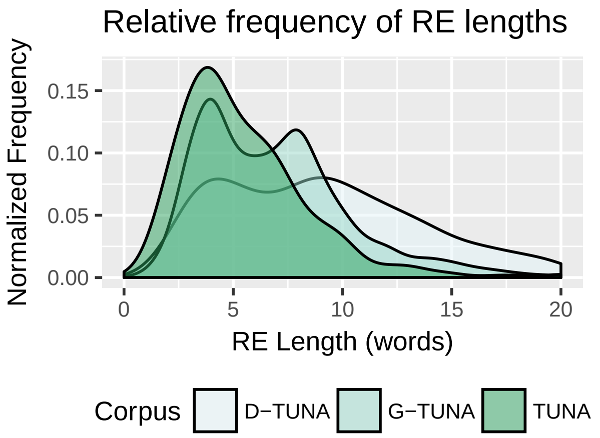

Figure 2: Density plot of RE lengths in the 3 TUNA corpora for comparable REs. The density plot is used so the distribution over different lengths is more easily compared across corpora despite the different numbers of REs in each corpus.

In addition to enabling cross-linguistic comparison and the evaluation of algorithms for referring expression generation in German, this corpus will provide insight into human behavior when describing objects for identification by listeners with varying levels of linguistic attention.

Acknowledgments

Thank you to student research assistants Fabio Lu & Matthias Lindemann.Thanks to Dietrich Klakow for feedback on drafts and Marc Schulder, Elisabeth Rabs, & Sebastien Le Maguer for photography. This work was supported by the DFG through SFB 1102 'Information Density and Linguistic Encoding'.

References

Ellen Gurman Bard, Anne H Anderson, Catherine Sotillo, Matthew Aylett, Gwyneth Doherty-sneddon, and Alison Newlands. 2000. Controlling the Intelligibility of Referring Expressions in Dialogue. *Journal of Memory and Language*, 42(1):1–22, jan.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Robert Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*, pages 68–75. Association for Computational Linguistics.

Alexia Galati and Susan E. Brennan. 2010. Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62(1):35–51.

Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Kristina Striegnitz. 2010. The GIVE-2 Corpus of Giving Instructions in Virtual Environments. In *LREC*.

Albert Gatt, Ielka van der Sluis, and Kees van Deemter. 2007. Evaluating algorithms for the Generation of Referring Expressions using a balanced corpus. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG)*, pages 49–56, Schloss Dagstuhl, Germany. Association for Computational Linguistics.

Albert Gatt, Ielka van der Sluis, and Kees van Deemter. 2008. XML format guidelines for the TUNA corpus. Technical report, Technical report, Computing Science, Univ. of Aberdeen, http://www. csd. abdn. ac. uk/ agatt/home/pubs/tunaFormat. pdf.

Ruud Koolen and Emiel Krahmer. 2010. The D-TUNA Corpus: A Dutch Dataset for the Evaluation of Referring Expression Generation Algorithms. In *LREC*.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250.

Emiel Krahmer, Sebastiaan Van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 130–132. Association for Computational Linguistics.

Kees van Deemter. 2002. Generating Referring Expressions: Boolean Extensions of the Incremental Algorithm. *Computational Linguistics*.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernndez, and David Schlangen. 2016. Pentoref: A corpus of spoken references in task-oriented dialogues. In *Proc. of the Language Resources and Evaluation Conference (LREC)*.