On task effects in NLG corpus elicitation: A replication study using mixed effects modeling

Emiel van Miltenburg **Merel van de Kerkhof**
Ruud Koolen **Martijn Goudbeek** **Emiel Krahmer**
Tilburg center for Cognition and Communication (TiCC), Tilburg University
C.W.J.vanMiltenburg@uvt.nl merelvdkerkhof@hotmail.com
R.M.F.Koolen@uvt.nl M.B.Goudbeek@uvt.nl E.J.Krahmer@uvt.nl

Abstract

Task effects in NLG corpus elicitation recently started to receive more attention, but are usually not modeled statistically. We present a controlled replication of the study by Van Miltenburg et al. (2018b), contrasting spoken with written descriptions. We collected additional written Dutch descriptions to supplement the spoken data from the DIDEC corpus, and analyzed the descriptions using mixed effects modeling to account for variation between participants and items. Our results show that the effects of modality largely disappear in a controlled setting.

1 Introduction

Natural Language Generation (NLG) systems are increasingly trained on the basis of datasets of human-produced examples, for example in the recent E2E-challenge (Dušek et al., 2018), or in automatic image description (Bernardi et al., 2016). The quality of the system output depends to a large extent on the quality of the data that is used to train the system, which in turn depends on the way that data is collected. A recent trend in NLG is to study task effects in the creation of corpora for natural language generation (Baltaretu and Castro Ferreira, 2016; van Miltenburg et al., 2017; Ilinykh et al., 2018). However, there does not seem to be an established methodology to investigate whether differences in task design lead to any significant differences in the output. This paper uses a tightly controlled approach to study task effects in NLG.

As a case study, we look at the effects of modality in an image description task. In their exploratory study, Van Miltenburg et al. (2018b) found that spoken and written descriptions differ in several ways, with the main result being that speakers have a greater tendency to show themselves through the use of 'egocentric language' (Akinnaso, 1982). The problem with this study is that it did not use matched corpora (containing exactly the same images) and their experiment did not control for the demographics of the participants. Therefore this paper presents a controlled replication of the study by Van Miltenburg et al. (2018b), to see if its findings are robust.

We carried out a between-subjects study where participants were assigned either to the SPOKEN or the WRITTEN condition. All participants were asked to describe the same images. For the former condition, we used the data from the Dutch Image Description and Eye-tracking Corpus (DIDEC; van Miltenburg et al. 2018a). For the latter condition, we collected additional data using a similar sample of participants. We analyzed the effects of modality on the elicited descriptions using mixed-effects models, controlling for variation in participants and images used to elicit the descriptions. We only found a significant effect for prepositions (used more in written descriptions); other effects disappear in a controlled setting.

This paper contributes to our understanding of the linguistic aspects of image descriptions (e.g., Ferraro et al. 2015; van Miltenburg et al. 2016; Alikhani and Stone 2019). Still, the main takeaway from our study is methodological: for studying task effects in elicitation tasks, we should control for individual variation and the effects of the stimuli used in the experiment. We hope that this study can serve as an example for the use of mixed effects modeling in natural language generation.[1]

2 The original study

Van Miltenburg et al. (2018b) aimed to identify

[1] All our code and data is publicly available online.
The interface for the written descriptions is available through:
https://github.com/evanmiltenburg/DIDEC-written.
The data analysis is available through:
https://github.com/evanmiltenburg/SpokenWritten-INLG
More information on the DIDEC website.

Feature	Terms
Consciousness-of-projection	*Lijkt, waarschijnlijk, misschien, duidelijk, mogelijk, zeker, vermoedelijk, eigenlijk*
Negations	*Geen, niet, niemand, nergens, noch, nooit, niets*
Positive allness	*Alle, elke, iedere, iedereen*
Pseudo-quantifiers	*Veel, vele, weinig, enkele, een paar, een hoop, grote hoeveelheid, kleine hoeveelheid*
Self-reference	*Ik, me, mij*

Table 1: Terms that were used for each feature. We added *vermoedelijk* ('presumably'), and *eigenlijk* ('actually').

differences between spoken and written image descriptions in both English and Dutch. Since our replication is carried out in Dutch, we will focus on the Dutch part of the original experiment.

Data. For the written sample, Van Miltenburg et al. used crowdsourced Dutch descriptions for the Flickr30K validation split (1000 images, 5 descriptions per image, collected by van Miltenburg et al. 2017). For the spoken descriptions, they used the Dutch Image Description and Eye-tracking Corpus (DIDEC; van Miltenburg et al. 2018a). This dataset contains 307 different images from the MS COCO dataset, with 14–16 spoken descriptions per image. The authors measured the following kinds of dependent variables:

LENGTH: Token length (in syllables or in characters), description length (in tokens). Both are measured after tokenizing the text.

PART-OF-SPEECH: (Attributive) adjectives, adverbs, prepositions. These are detected using a part-of-speech tagger (SpaCy 2.0.4).

SEMANTIC CATEGORIES: negations (*no, not*), pseudo-quantifiers (*few, lots*), consciousness-of-projection terms (*seem, appear, maybe*, positive allness terms (*all, every*), and self-reference terms (*I, me, my*) are detected by matching word tokens with a word list. Table 1 provides an overview.

OTHER: Propositional Information Density (PID; Turner and Greene 1977), which corresponds to the average number of propositional ideas per word in a text, and is computed through an external tool (Marckx, 2017). Mean-segmental type-token ratio (MSTTR; Johnson 1944), which is a measure of diversity (the average number of types per segment).

Findings. Van Miltenburg et al. (2018b) found no consistent differences between spoken and written descriptions for token length, MSTTR, PID, or the use of adjectives or prepositions. The authors did find that spoken descriptions are longer, and contain more adverbs, negations, positive allness terms, self-reference terms, pseudo-quantifiers, and consciousness-of-projection terms. This led them to conclude that speakers have a greater tendency to show themselves through the use of 'egocentric language' (Akinnaso, 1982). What the authors mean by this is that spoken descriptions are not just neutral and detached, but that they also tend to communicate something about the observer who generated the description. For example, if a participant says that some entity X *looks like* or *might be* a sheep (i.e., describing the entity using consciousness-of-projection terms), then their description also signals their uncertainty about whether X is a sheep or not. Written descriptions typically avoid this kind of language (Akinnaso, 1982).

Limitations. The original study did not control for the content of the images, or for the demographics of the participants. Furthermore, it did not control for the setting: the DIDEC dataset was collected in a laboratory setting, whereas the written sample was collected through a crowdsourcing task. This makes it hard to determine whether the results were actually due to the difference in modularity, and not due to any other difference. Hence we set out to provide a controlled replication.

3 The current study

The current study was set up to provide a more controlled comparison between spoken and written image descriptions. We collected written descriptions for the images from the Dutch Image Description and Eye-tracking Corpus, so that we could compare these written descriptions to the existing spoken data. We used a different sample of participants from the exact same population (the Tilburg University participant pool) to generate the descriptions, so that we could isolate the effect of modality on the generated descriptions.

Participants. Our participants were 48 Dutch students (33 women, 15 men, with a mean age of 21.6) who earned course credits for their participation. Our study followed standard ethical procedures. We obtained IRB approval for this study, and all participants were asked for their informed consent. Participants were allowed to quit the experiment at any stage and still earn credits.

Materials. We used the same 307 images (originally from MS COCO) that were used for the creation of the DIDEC dataset. In the original task, participants provided spoken descriptions for 102 or 103 images in one session. However, written language is typically slower to produce than spoken language; data from Van Miltenburg et al. (2017) shows that the median time for crowdworkers to write 5 descriptions is 294 seconds. Extrapolating from this, we expect that it would take 49 minutes to write descriptions for 50 images. To ensure that participants are able to finish the experiment within one hour (and to avoid fatigue), we shortened the lists to 51 or 52 images.

Design. We used a single-factor (modality) between-subjects design, where the participants who took part in the DIDEC study serve as the SPOKEN group, and we collect additional responses for the WRITTEN condition. Because both sets of participants are sampled from the same population, we can compare their descriptions for the same images to examine the effect of modality. However, we do note that a within-subjects design would have more statistical power, since we would also have information about the effects of modality for each participant.[2] Our choice for a between-subjects design was motivated by economic reasons: it would have been very time-consuming to build a new corpus of spoken image descriptions.

Procedure. The elicitation task is similar to the one carried out by Van Miltenburg et al. (2018a) for the DIDEC dataset. We implemented the task using Qualtrics,[3] so as to have a simple web interface. The participants sat in a computer room with 20 computers. They were not allowed to communicate with each other. After reading the instructions and signing the consent form, participants first carried out a practice trial, after which they could ask clarification questions. For the main task, participants were presented with a list of images, and asked to describe each of the images in one short but complete sentence.

Dependent variables. Our dependent variables are almost the same as in the original study; we ignore MSTTR for reasons of space.[4] We modified the original (public) scripts to prepare the results for our analysis. Whereas the original study reported average results over the aggregated data (per 1000 tokens or per description), we measure the variables for each individual description.

	Spoken	Written
Number of participants	45	48
Number of descriptions	4604	2547
Descriptions per image	14–16	7–9

Table 2: General statistics for the two datasets. Spoken data comes from the DIDEC dataset (van Miltenburg et al., 2018a), written data was collected for this paper.

4 Statistical analysis

In addition to the effects of modality (SPOKEN and WRITTEN), our observations (the individual descriptions) are influenced by two other factors; namely PARTICIPANT and IMAGE. To capture the random effects of both participants and images, we use a linear mixed effect model (Baayen et al. 2008; see Winter 2013 for a tutorial). We used the lme4 package (Bates et al., 2015) to build our models in R (R Core Team, 2017) and the lmertest package (Kuznetsova et al., 2017) to provide p-values for linear mixed effect models. We created a separate model for each dependent variable and assessed the effect of modality for significance.[5] When significant, the null hypothesis of no difference between the means of the written and spoken condition is rejected (implying there is a task effect). For each model, we specify the relevant type of distribution. We model sentence length, token length, and propositional idea density as continuous data, and assume a standard Gaussian distribution. The other variables correspond to count data, modeled using the Poisson distribution (through the glmer function).

5 Results

We collected 2457 descriptions from 48 participants. Table 2 provides general statistics about the spoken and written descriptions. Descriptive statistics are provided in Table 3. Compared to the spoken descriptions, written descriptions are longer, have longer tokens, and (with the exception

[2] This is the set-up of Drieman (1962), who asked participants to describe different paintings using either spoken or written language. However, they did not use mixed effects models, and could not investigate stimulus effects.

[3] An online survey platform: https://www.qualtrics.com

[4] MSTTR should be analyzed using a t-test, since we cannot analyze diversity at the item level. We can only aggregate the descriptions for each participant, compute the MSTTR scores (one per participant) and see whether there is a significant difference in the scores between the two conditions.

[5] We use the traditional significance level of $\alpha = 0.05$, and correct for multiple comparisons using the Bonferroni method: $\alpha' = 1 - (1 - \alpha)^{1/k}$. With k=12 models, α=0.00427.

#	Variable	Expectation	μ_{spoken}	μ_{diff}	Data type	β_{written}	SE	Statistic	p
1.	Sentence length	s>w	12.621	+2.632	Continuous	2.630	0.993	t: 2.648	0.010
2.	Token length (characters)	s=w	4.679	+0.003	Continuous	0.005	0.046	t: 0.111	0.912
3.	Token length (syllables)	s=w	1.519	+0.001	Continuous	0.001	0.014	t: 0.087	0.931
4.	Propositional idea density	s=w	0.443	+0.003	Continuous	0.002	0.006	t: 0.367	0.714
5.	Attributive adjectives	s=w	0.495	+0.071	Count	0.151	0.105	z: 1.434	0.152
6.	Adverbs	s>w	0.648	+0.070	Count	0.092	0.127	z: 0.726	0.468
7.	Prepositions	s=w	1.810	+0.821	Count	0.260	0.069	z: 3.768	<0.001
8.	Negations	s>w	0.010	+0.005	Count	0.438	0.288	z: 1.520	0.128
9.	Pseudo-quantifiers	s>w	0.050	+0.024	Count	0.459	0.201	z: 2.288	0.022
10.	Consciousness-of-projection	s>w	0.031	–0.018	Count	–0.852	0.364	z: –2.339	0.019
11.	Self-reference	s>w	0.145	+0.072	Binary	–2.291	1.010	z: –2.268	0.023
12.	Positive allness	s>w	0.004	+0.001	Binary			Failed to converge.	

Table 3: All models with their dependent variables, whether we expect a difference (less/greater than: difference, equals: no difference), the mean results for the spoken descriptions, difference between written and spoken descriptions (w–s), the data type used in our analysis, the fixed effect (β) of the written modality on the outcome, the standard error, statistic, and the p-value for the model.

of consciousness-of-projection terms) contain more terms from each category. The direction of these differences is surprising, because they are opposite to our expectations (again with the exception of consciousness-of-projection terms). For example, we expected spoken descriptions to be longer than written ones, indicated as ‘s>w’ in Table 3.

To assess whether these observed differences generalize outside of this particular dataset, we assessed their statistical significance using the linear mixed effect models described earlier.

Model convergence. Initially, the models for token length (syllables), self-reference, and allness terms failed to converge (i.e. find stable estimates of the effects). We addressed this issue in two ways: 1. For the token length model, we used a different optimizer (`bobyqa`); 2. For self-reference and allness terms, we modeled the presence or absence of the relevant terms with a binomial distribution. After this, only the model for positive allness failed to converge; likely because only 30 out of 7,056 descriptions contained positive allness terms —not enough positive examples.

Main results. The last four columns of Table 3 show the effect of modality on the dependent variables (full models are in the supplementary materials). We only found a significant effect of modality on the use of prepositions: written descriptions use more prepositions than spoken ones.

We found no significant effect of modality on any of the other dependent variables. (Note that this is partly due to the Bonferroni correction we applied earlier. If we had not corrected for multiple comparisons, we would have judged the models for sentence length, pseudo-quantifiers, consciousness-of-projection terms, and self-reference terms to be significant at $\alpha = 0.05$.) This means that while those models may capture general tendencies in the data, there are no consistent differences between spoken and written language for these variables.

Model interpretation. Although most of our analyses do not show significant differences, we can still interpret the way they capture the overall distribution of the data. The strongest non-significant effect is observed for sentence length; on average, written descriptions are β=2.6 words longer than spoken ones.

6 Discussion

We will now briefly summarize and explain our results, before discussing their implications.

6.1 Summary of the results

We aimed to replicate the findings by van Miltenburg et al. (2018b), who looked at modality effects in the elicitation of NLG corpus data. Like the original authors, we found no significant difference for token length, PID, or the use of adjectives. While Van Miltenburg et al. did not find a consistent difference in the use of prepositions for both English and Dutch prepositions, we replicate their finding that written Dutch descriptions contain more prepositions than spoken ones. This is in line with earlier findings by Drieman (1962) and Chafe and Danielewicz (1987).

All other effects disappear in a controlled setting. This is not to say that there is no effect of modality, but that the effect is smaller than can be detected with these variables and this elicitation method. We are unsure how the original effects emerged, but a likely explanation lies in the differences between the datasets used in the original study (which contain different images, and were collected in a different setting, with less comparable participants). This shows the importance of setting up a controlled study, where such differences are minimized, and we can isolate the factor that we are interested in (here: modality).

6.2 Rarity and the need for guidelines

One other factor contributing to the difficulty of finding statistically significant effects for modality is that many of the phenomena under investigation are low-frequent. Positive allness terms are the most extreme case, occurring in 0.4% of all spoken descriptions. But attributive adjectives, negations, pseudo-quantifiers, consciousness-of-projection terms, and self-reference terms also occur in less than half of the spoken descriptions.

It appears that only changing the modality is not enough to observe a (strong) task effect. If we want participants to produce different kinds of descriptions, they will probably need guidelines, with explicit instructions to change their behavior. But this raises the next issue: what should those guidelines look like?

6.3 Usefulness of different modalities

One of the reasons cited by van Miltenburg et al. (2018b) to look at spoken image descriptions is that they might provide more natural examples of how people generally talk about images. After all: speech is a more primary form of language (cf. Biber 1988, Chapter 1). Their naturalness would make spoken descriptions more suitable for training voice-operated image description systems.

Our results show that changing the modality of the elicitation task does not necessarily yield qualitatively different descriptions, let alone more natural descriptions. Importantly, our study does not say anything about the *usefulness* of typical features of spoken language. User studies may still find that emulating the spoken style (as presented in the literature) positively/negatively affects users' appreciation of the output. After establishing desirable properties that image descriptions should have, we can define guidelines for what image descriptions should look like. We may then be able to alter the elicitation task in such a way that participants provide suitable descriptions. Here, the question arises: how do you know whether the elicitation task is successful? This brings us to our next point.

6.4 Statistics as a manipulation check

Our failure to replicate the effects of modality for all variables (except the use of prepositions) suggests that (at least for those other variables) it does not matter in which modality you collect the descriptions; they will look more or less the same. In other words, our study served as a *manipulation check*, to see if the manipulation (changing the modality of the elicitation task) had the desired effect (changing the style of the descriptions). In this case, the manipulation turned out to be unsuccessful. We hope that our study provides a good example for showing (or refuting) the robustness of different task effects in NLG. Note that, for a check like this to be possible, one needs to establish a metric or set of metrics that can be used to quantify the phenomenon that you're interested in.

7 Conclusion

We presented a controlled study to evaluate task effects in an NLG elicitation task, namely image description. We used mixed effects models to filter out the effects of participants and individual stimuli. Using these models, we learned that modality alone has a minimal effect on the content of the descriptions. Thus, a stronger manipulation is needed to obtain different kinds of descriptions. The methodology used in this paper is suitable for running pilot studies to check whether task manipulations are successful. We hope that future studies will adopt this methodology, so as to ensure fruitful data collection.

8 Acknowledgments

This study is based on the MA thesis of the second author. The design was approved by the Research Ethics and Data Management Committee at Tilburg University, reference: REC #2018/56.

References

F Niyi Akinnaso. 1982. On the differences between spoken and written language. *Language and speech*, 25(2):97–125.

Malihe Alikhani and Matthew Stone. 2019. "caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67, Minneapolis, Minnesota. Association for Computational Linguistics.

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Adriana Baltaretu and Thiago Castro Ferreira. 2016. Task demands and individual variation in referring expressions. In *Proceedings of the 9th International Natural Language Generation conference*, pages 89–93, Edinburgh, UK. Association for Computational Linguistics.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

Wallace Chafe and Jane Danielewicz. 1987. Properties of spoken and written language. In R. Horowitz and F.J. Samuels, editors, *Comprehending oral and written language*. New York: Academic Press.

Gerard HJ Drieman. 1962. Differences between written and spoken language: An exploratory study, I. quantitative approach. *Acta Psychologica*, 20:36–57.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328, Tilburg University, The Netherlands. Association for Computational Linguistics.

Francis Ferraro, Nasrin Mostafazadeh, Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell, et al. 2015. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 201–213.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2018. The task matters: Comparing image captioning and task-based dialogical image description. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 397–402, Tilburg University, The Netherlands. Association for Computational Linguistics.

Wendell Johnson. 1944. I. a program of research. *Psychological Monographs*, 56(2):1.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Silke Marckx. 2017. Propositional idea density in patients with alzheimer's disease: An exploratory study. Master's thesis, Universiteit Antwerpen.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.

Emiel van Miltenburg, Ákos Kádar, Ruud Koolen, and Emiel Krahmer. 2018a. DIDEC: The Dutch Image Description and Eye-tracking Corpus. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*. Resource available at https://didec.uvt.nl.

Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018b. Varying image description tasks: spoken versus written descriptions. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*, pages 54–59, Berlin, Germany. Association for Computational Linguistics.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Althea Turner and Edith Greene. 1977. *The construction and use of a propositional text base*. Institute for the Study of Intellectual Behavior, University of Colorado Boulder.

Bodo Winter. 2013. Linear models and linear mixed effects models in r with linguistic applications. *arXiv preprint arXiv:1308.5499*.