Varying image description tasks: spoken versus written descriptions

Emiel van Miltenburg
Vrije Universiteit Amsterdam
Emiel.van.Miltenburg@vu.nl

Ruud Koolen
Tilburg University
R.M.F.Koolen@tilburguniversity.edu

Emiel Krahmer
Tilburg University
E.J.Krahmer@tilburguniversity.edu

Abstract

Automatic image description systems are commonly trained and evaluated on *written* image descriptions. At the same time, these systems are often used to provide *spoken* descriptions (e.g. for visually impaired users) through apps like *TapTapSee* or *Seeing AI*. This is not a problem, as long as spoken and written descriptions are very similar. However, linguistic research suggests that spoken language often differs from written language. These differences are not regular, and vary from context to context. Therefore, this paper investigates whether there are differences between written and spoken image descriptions, even if they are elicited through similar tasks. We compare descriptions produced in two languages (English and Dutch), and in both languages observe substantial differences between spoken and written descriptions. Future research should see if users prefer the spoken over the written style and, if so, aim to emulate spoken descriptions.

1 Introduction

Automatic image description systems (Bernardi et al., 2016) are commonly trained and evaluated on datasets of described images, such as Flickr30K and MS COCO (Young et al., 2014; Lin et al., 2014). These datasets have been collected by asking workers on Mechanical Turk to write English descriptions that capture the contents of the images that are presented to them. But how much are these descriptions influenced by the modality of the task? This paper explores the differences between spoken and written image descriptions. While many papers at VarDial aim to distinguish similar *dialects* using machine learning (e.g. in shared tasks such as those in Malmasi et al., 2016; Zampieri et al., 2017), we aim to identify the features distinguishing two similar *varieties* (spoken and written language) of the same language (either English or Dutch) in a particular domain (image descriptions).

One of the motivations behind automatic image description research is to support blind or visually impaired people (e.g. Gella and Mitchell, 2016), and indeed *apps* are starting to appear which describe visual content for blind users (e.g. *TapTapSee* or Microsoft's *Seeing AI*[1]). These apps are commonly used together with screen readers, which convert on-screen text to speech. Given this presentation through speech, it is worth asking: should we not also collect *spoken* rather than *written* training data? That might give us more natural-sounding descriptions. But a big downside of collecting spoken training data is that it also requires a costly transcription procedure (unless we go for an *end-to-end* approach, see Chrupała et al., 2017). An alternative is to try to understand what the differences are between written and spoken image descriptions. Once we know those differences, and we know what kind of descriptions users prefer, we may be able to direct image description systems to produce more human-like descriptions, similar to the way we can modify the style of the descriptions, for example with positive/negative sentiment (Mathews et al., 2016), or humorous descriptions (Gan et al., 2017).

This paper presents an exploratory study of the differences between spoken and written image descriptions, for two languages: English and Dutch. We provide an overview of the variables that have been found to differ between spoken and written language, and see whether these differences also hold

[1] TapTapSee: https://taptapseeapp.com/; Seeing AI: https://www.microsoft.com/en-us/seeing-ai/

Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 88–100
Santa Fe, New Mexico, USA, August 20, 2018.

between English spoken and written image descriptions. Following this, we repeat the same experiment for Dutch. Our main findings are that spoken descriptions (1) tend to be longer than written descriptions, (2) contain more adverbs than written descriptions, (3) contain more pseudo-quantifiers and allness terms (DeVito, 1966), and (4) tend to reflect the certainty of the speaker's beliefs more-so than written descriptions. Our work paves the way for a future controlled replication study, and follow-up studies to assess what kind of descriptions users prefer. All of our code and data is available online.[2]

2 Technical background: Manipulating the image description task

Recently, researchers have started to manipulate the image description task to obtain a better understanding of how this influences the resulting descriptions. This section presents a brief list of variables that have been considered in the literature.

Language. The most common modification is the *language* in which the task is carried out (e.g. Elliott et al., 2016; Li et al., 2016; Miyazaki and Shimizu, 2016). This is typically done to be able to train an image description system in a different language, but van Miltenburg et al. (2017) use this manipulation to show that speakers of different languages may also provide different descriptions. For example, speakers of American English described sports fans barbecuing on a parking lot as *tailgating*, a concept unknown to Dutch and German speakers.

Style. Another possible manipulation is the requested *style* of the descriptions. Gan et al. (2017) asked crowd workers to provide 'humorous' and 'romantic' descriptions, but found that it is impossible to control the quality of the resulting descriptions. So they, like Mathews et al. (2016), further changed the description task to a *description editing* task.

Content. Gella and Mitchell (2016) emphasize the importance of *emotional or descriptive content* and *humor* in the image, and explicitly ask for these to be annotated. This makes the elicited descriptions useful for training an assistive image description system which can provide descriptions for blind people.

Other. Baltaretu and Castro Ferreira (2016) present variations on an *object description* task (the *ReferIt* task, by Kazemzadeh et al. (2014)). The authors show that asking participants to work very fast, or produce thorough or creative descriptions, results in very different kinds of descriptions.

While the studies listed above cover a wide range of variables, there are many more possibilities that are still unexplored. Van Miltenburg et al. (2017) provide a (non-exhaustive) list of other factors that may influence the image description process. This paper aims to identify the role of modality.

3 Theoretical background: Spoken versus written language

The differences between spoken and written language have been thoroughly studied in the linguistics literature since the 1960s. Extensive overviews are provided by Akinnaso (1982), Chafe and Danielewicz (1987), Chafe and Tannen (1987), Biber (1988), and Miller and Fernandez-Vest (2006). Why should we study differences between spoken and written image descriptions, when so many linguists before us have studied differences between spoken and written language? Because spoken and written language are not monoliths. Biber (1988) notes that there is often as much variation within each modality, as there is between the two modalities. Biber attributes this variation to situational, functional, and processing considerations (p. 24-25). So while there may be general tendencies for particular linguistic phenomena to occur more in written than in spoken language (or vice versa), the only way to know for sure how speech differs from writing in a particular domain is to investigate that particular domain. For image description, the seminal study by Drieman (1962a; 1962b) is of particular interest to us. Drieman asked eight participants to describe two realistic paintings (one by Renoir and one by Weissenbruch), providing either spoken or written descriptions. He found that written texts (1) are shorter; but (2) have longer words (fewer words of one syllable, more words of more than one syllable); (3) have more attributive adjectives[3]; and (4) a more varied vocabulary. The drawback of this study is its limited size. Moreover, it is unclear if Drieman's conclusions about descriptions of paintings extend to one-sentence image

[2] Our code and data is available at https://github.com/cltl/Spoken-versus-Written

[3] In English, this means that the adjective is used in the prenominal position (*the **good** book*) rather than postnominal (*the book is **good***). The same holds for Dutch.

MS COCO instructions	Flickr30K instructions
1. Describe all the important parts of the scene. 2. Do not start the sentences with "There is." 3. Do not describe unimportant details. 4. Do not describe things that might have happened in the future or past. 5. Do not describe what a person might say. 6. Do not give people proper names. 7. The sentences should contain at least 8 words.	1. Describe the image in one complete but simple sentence. 2. Provide an explicit description of prominent entities. 3. Do not make unfounded assumptions about what is occurring. 4. Only talk about entities that appear in the image. 5. Provide an accurate description of the activities, people, animals and objects you see depicted in the image. 6. Each description must be a single sentence under 100 characters.

Figure 1: Instructions for the written English data. MS COCO instructions are from Chen et al. (2015). Flickr30K instructions are from the appendix of Hodosh et al. (2013), edited for brevity.

descriptions like those in MS COCO and Flickr30K. This is what we intend to study.

Following Drieman's study, researchers have proposed many other variables that seem to correlate with the speech/writing distinction. After surveying the literature on spoken versus written language, Biber (1988) presents an extensive list of linguistic features. The features used in this paper are based on Biber's list, see Section 4.3 for an overview. Noted in almost all surveys is the *ephemeral nature* of speech; whereas writing samples can be edited and reworded, speech cannot be edited the same way. Hence, spoken language also contains false starts, speech errors, and subsequent repairs. But despite those flaws, we must not think of spoken language as somehow *inferior* to written language. Halliday (1989) notes that the two are simply different media that serve different functions, which may require different forms of language. It is our task, as language users, to pick the right form (and medium) for the right job. If we find significant differences between spoken and written language, we should ask ourselves: now that we know about these differences in the way people describe images, which form is the most suitable for an image description system?

4 Data and methods for analyzing image descriptions

We present an analysis for both Dutch and English image descriptions. For each language, we took existing sets of spoken and written image descriptions, and automatically computed their differences in terms of the literature discussed above. The rationale here is that, even if these corpora are not perfectly comparable, they do provide an indication of the extent to which spoken and written image descriptions may differ. If we find structural differences between spoken and written image descriptions, it may be worth it to explore these differences further in a more controlled environment. If we fail to find any differences, we should conclude that there is no evidence for the effect of modality on the image description task. But, as we will see later, there do seem to be structural differences between spoken and written descriptions in both Dutch and English.

4.1 English data

For the written sample, we use the Flickr30K and the MS COCO datasets. Both were collected through Mechanical Turk, and have 5 written descriptions per image. We only use the training splits from both datasets, so that we remain ignorant of the properties of the validation and test splits. Figure 1 provides the instructions for both datasets. One of the main differences between the two is that the MS COCO instructions explicitly forbid the use of *there is* at the start of a sentence, which leads to the use of different syntactic constructions. Otherwise the instructions are very similar.

For the spoken sample, we use the Places Audio Caption Corpus, Part 1 (Harwath et al., 2016; Harwath and Glass, 2017), which contains about 230,000 spoken descriptions for a selection of images that were equally sampled from the 205 scene categories in the Places205 dataset (Zhou et al., 2014). The spoken descriptions were collected through Mechanical Turk using the Spoke framework (Saylor, 2015). These were then automatically transcribed by Harwath et al. (2016) using the Google Speech API. Because the transcriptions were not manually corrected, they have a word error rate of about 20%. It is unclear how participants were instructed to describe the image. The authors only mention that the descriptions are

free-form, and that they should describe the salient objects in the scene.[4]

Image selection. The images from Flickr30K, MS COCO, and Places205 were all collected from online sources. Flickr30K and MS COCO exclusively use images from Flickr,[5] while Places also contains images found through general image search engines (Google and Bing). The main difference between the datasets is in the kind of images that are included. For the Flickr30K dataset, the authors downloaded images from six different user groups on the Flickr website.[6] The MS COCO authors compiled a list of 91 object categories, and searched for different object+object combinations of different categories on Flickr. They also selected 60 scene categories from the SUN database (Xiao et al., 2010), and searched for different object+scene combinations to diversify their data. Finally, the Places205 dataset is built by querying different search engines for adjective+scene combinations. The 205 scenes come from the SUN database, and the adjectives come from a manually curated list. Examples are: *messy, spare, sunny*.

Comparability. To what extent can we compare the descriptions in these datasets? Ideally, we would have one set of images that is provided with both spoken and written descriptions. But if the tasks are similar enough, and we compare the image descriptions on a large scale, we may still be able to confirm general tendencies of spoken versus written data, e.g. that spoken descriptions tend to be longer than written ones (Drieman, 1962a), or use more self-reference terms (DeVito, 1966). What we *cannot* do, is compare how often particular properties or kinds of entities are mentioned, because the distribution of those properties or entities might be dramatically different. Generally speaking, using different sets of images also means that we can never exclude the possibility that the underlying cause of the differences between the descriptions lies with the images rather than the modality. However, as the sets of images become more similar, chances of the images being a major source of the differences between the written and spoken descriptions become smaller. So how big *are* the differences between existing datasets?

To answer this question, we tagged the descriptions in all three datasets using the SpaCy part-of-speech tagger.[7] Table 1 shows the top-10 most frequent nouns in all three datasets. These correspond to the most frequent entities. We observe that while the Flickr30K and MS COCO datasets are fairly similar (sharing 5 words in their top-10), the Places dataset stands out from the other two (sharing only 2 words). Thus, the only spoken English descriptions that are available, describe images that are fairly different from the other datasets. Luckily we have more comparable data for Dutch.

	Flickr30K		MS COCO		Places	
#	Word	Count	Word	Count	Word	Count
1	man	42595	man	48847	picture	36020
2	woman	22197	people	25723	people	26094
3	people	17338	woman	22992	building	25735
4	shirt	14341	table	21104	trees	22449
5	girl	9656	street	20527	water	20324
6	men	9499	person	16857	man	18609
7	boy	9399	top	14755	front	16584
8	dog	9093	field	14597	background	15484
9	street	8012	group	14450	side	15254
10	group	7852	tennis	13411	room	12985

Table 1: Top-10 most frequent nouns for all three datasets. Flickr30K and MS COCO are fairly similar (they have a larger overlap), but Places differs from the other two.

4.2 Dutch Data

For the written sample, we use the data collected by van Miltenburg et al. (2017). The authors crowdsourced Dutch descriptions for the Flickr30K validation and test sets (1014 + 1000 images, with 5 descriptions per image). The annotation task was translated from the Flickr30K and Multi30K templates (Elliott et al., 2016), to stay as close to these datasets as possible. We only use the validation split for our comparison, so that we remain ignorant of the properties of the test set.

For our spoken sample, we use data from a task that we carried out for another purpose, and that will be published separately (van Miltenburg et al., 2018). 45 Dutch students participated in a lab experiment

[4] We contacted the authors for more information about the crowd-sourcing task, but have not received any response.

[5] A social image sharing platform, see: www.flickr.com.

[6] These user groups are: *strangers!; Wild-Child (Kids in Action); Dogs in Action (Read the Rules); Outdoor Activities; Action Photography; Flickr-Social (two or more people in the photo)*. See (Hodosh et al., 2013) for the full methodology.

[7] We use version 2.0.4. See: http://spacy.io/

where they were asked to describe a series of images, while we also measured their eye movements. We used the 307 images from MS COCO that both appear in SALICON and the Visual Genome dataset (Jiang et al., 2015; Krishna et al., 2016). We transcribed and annotated the recorded descriptions, so that we ended up with three layers: (1) a raw layer; (2) an annotated layer indicating (filled) pauses, corrections and repetitions; and (3) a normalized layer, with the 'intended' description. In total, we collected 14-16 descriptions per image, resulting in a grand total of 4604 descriptions for the entire dataset. This study uses the normalized descriptions, so that our metrics are unaffected by corrections and repetitions, and we can focus more on the content of the descriptions.

4.3 Preprocessing, metrics, and hypotheses

We tokenize, tag, and parse the descriptions using SpaCy. Then, we compute the following metrics:

1. **Average token length** Drieman (1962a) and others have found that the tokens in spoken language are shorter than those in written language. We measure token length in terms of syllables (following e.g. Drieman 1962a) and characters (following e.g. Biber 1988), using Hunspell to obtain the syllables.[8]
2. **Average description length** Drieman (1962a) and others have shown that spoken language has a higher sentence length than written language. We measure description length in tokens and syllables.
3. **Mean-segmental type-token ratio** (MSTTR) corresponds to the average number of types per 1000 tokens (Johnson, 1944). It is used as a measure of lexical variation. Because it is computed for a fixed number of tokens, it is unaffected by corpus size or sentence length. Drieman (1962a) shows that written language is more diverse than spoken language. One issue is that the Places Audio Caption Corpus has only one description per image, versus five descriptions per image for MS COCO and Flickr30K. This means that for every description in Flickr30K or MS COCO, there are four very similar descriptions, which makes these corpora less diverse overall. For a fair comparison, we treat Flickr30K and MS COCO as collections of five similar corpora, compute MSTTR for each of these, and report the average.
4. **Attributive adjectives** Drieman shows that spoken language contains fewer attributive adjectives than written language. We use SpaCy's tagger and parser to determine if an adjective is attributive or not. We consider a token to be an attributive adjective if its part-of-speech tag is ADJ, and it has an *amod* dependency relation with a head that is either tagged as NOUN or PROPN. In other words: if it's an adjective modifying a noun.
5. **Adverbs** We count all tokens with the ADV part-of-speech tag. The literature shows mixed results for the use of adverbs: Harrell (1957) studied children's production of stories, and found *fewer* adverbs in spoken than in written language, while Chafe and Danielewicz (1987) show that adverbs are used *more* in conversation and letters, and less in lectures and academic writing. They explain this pattern by arguing that the key variable is not *modality* but *involvement*. Whenever people are more involved with their audience or their environment, they also tend to use more locative or temporal adverbials. And whenever they are more *detached* (talking about more abstract ideas), they tend to use fewer adverbs.
6. **Prepositions** Chafe and Danielewicz (1987) show that prepositions are used more in (academic) written language. We count all tokens with the ADP part-of-speech tag.

The metrics below are computed by matching the tokenized descriptions with different sets of words.

7. **Consciousness-of-projection terms** DeVito (1966) defines these as: "words which indicate that the observed is in part a function of the observer." He shows that these words are more frequently used in speech than in writing. Since DeVito does not provide a list of the terms used in his work, we compiled our own list containing the following words: *apparently, appear, appears, certainly, clearly, definitely, likely, may, maybe, might, obviously, perhaps, possibly, presumably, probably, seem, seemed, seemingly, seems, surely*. The consciousness-of-projection terms contain Biber's (1988) set of *possibility modals* and *seem and appear*.
8. **Self-reference terms** DeVito (1966) also shows that self-reference terms (first-person pronouns and phrases like *the author*) are used more in spoken than in written language. We only use *I, me, my* as self-reference terms, since phrases like *the author* are not relevant in this domain.

[8] Hunspell is the spell checker from LibreOffice, which has a powerful hyphenation function. See: https://hunspell.github.io for more details. We use the Pyphen library (https:/github.com/Kozea/Pyphen) as an interface.

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

Table 2: Dutch terms that were used for each feature.

9. **Positive allness terms** DeVito (1966) shows that spoken language contains more 'allness terms' than written language. For DeVito, these include both positive (*all, every, always*) and negative (*none, never*) terms. Following more recent work, which also focuses explicitly on negations, we decided to distinguish between the two. As *positive* allness terms, we use the words *all, each* and *every*.
10. **Negations** Biber et al. (1999, Chapter 3) show that spoken language contains more negations than written language. For the *negative* allness terms, we focus on explicit, non-affixal negations: *n't, neither, never, no, nobody, none, nor, not, nothing, nowhere*. (Using the terminology from Tottie (1980).)
11. **Pseudo-quantifiers** While DeVito (1966) did not find any significant differences in the use of exact numerals, between spoken and written language, he did find such differences in the usage of terms like *many*, that are "loosely indicative of amount or size." We use the following terms: *few, lots, many, much, plenty, some* and *a lot*.

Table 2 shows the Dutch terms used for each feature. For all features except average token length, average description length, and MSTTR, we report the average number of occurrences per description, and per 1000 tokens. We also compute the Propositional Idea Density (PID) for the spoken and written descriptions. PID corresponds to the average number of propositional ideas per word in a text (Turner and Greene, 1977). According to Turner and Greene's annotation scheme, sentence (1a) breaks down into the five ideas expressed in (1b).[9] Because the nine words in (1a) express five ideas, the PID for this sentence is $5/9 = 0.56$.

(1) a. The old gray mare has a very large nose

b. HAS(MARE, NOSE), OLD(MARE), GRAY(MARE), LARGE(NOSE), VERY(LARGE)(NOSE)

We expect that written language has a higher PID than spoken language. In other words: that spoken language uses more words to convey the same amount of information. This hypothesis is based on the idea that written language is edited or condensed to convey as much information as possible. For example, Chafe and Danielewicz (1987) show that nominalizations (e.g. *categorization, development*) occur more often in written language. They argue that the spoken alternatives for nominalizations are often much longer: several clauses instead of one. Another example comes from Ravid and Berman (2006), who show that written narratives contain relatively more *propositional* content ("events, descriptions, and interpretations") and less *ancillary* content ("nonnovel, nonreferential, or nonnarrative"). Spoken narratives are said contain more ancillary content for communicative purposes. We use existing tools to measure idea density. For English, we use the Computerized Propositional Idea Density Rater (Brown et al., 2008).[10] For Dutch, we use the tool developed by Marckx (2017).

Because this is an exploratory study, we will only report descriptive statistics. These allow us to formulate hypotheses about the differences between spoken versus written image descriptions. We can test these hypotheses in a future study with spoken and written descriptions for the same images, collected in the same controlled setting.

5 Results

This section presents an overview of the different metrics for the Dutch and the English data. We first present the English results, followed by the Dutch results, and end with a summary of our main findings.

			TokLen		DescLen		Attributives		Adverbs		Prepositions	
Name	#Desc	#Tok	Syll	Char	Syll	Tok	Desc	‰	Desc	‰	Desc	‰
MS COCO	414,113	4,348,698	1.29	4.01	13.53	10.50	0.64	60.97	0.16	14.99	1.75	166.37
Flickr30K	145,000	1,787,693	1.30	4.11	16.05	12.33	0.97	78.81	0.15	12.20	1.91	154.78
Places	229,388	4,765,891	1.26	4.08	26.27	20.78	1.39	66.77	0.97	46.67	3.06	147.14

		Consciousness		Self-reference		Allness		Negations		PseudoQuant	
Name	MSTTR	Desc	‰	Desc	‰	Desc	‰	Desc	‰	Desc	‰
MS COCO	0.32	0.00	0.22	0.00	0.09	0.02	1.56	0.00	0.42	0.06	6.01
Flickr30K	0.38	0.01	0.63	0.00	0.09	0.02	1.30	0.00	0.35	0.04	2.88
Places	0.34	0.08	4.07	0.05	2.49	0.07	3.22	0.06	2.88	0.24	11.66

Table 3: Results for our analysis of MS COCO, Flickr30K (both written), and the Places Audio Caption Corpus (spoken). For the top table, columns correspond to: number of descriptions, number of tokens, average token length (in syllables), average description length (in syllables, in tokens), features 4-6 (per description, per 1000 tokens). For the bottom table, columns correspond to the mean-segmental type-token ratio, followed by features 7-11 (per description, per 1000 tokens).

5.1 English results

Table 3 shows the results for the English descriptions. We immediately see that, in line with the literature, spoken image descriptions are almost twice as long as their written counterparts. With almost half the number of descriptions of MS COCO, the Places dataset has significantly more tokens. Based on the literature, we might also expect spoken descriptions to use shorter words than written descriptions. This is indeed the case when we look at syllable length, but when we look at the number of characters, tokens in the MS COCO dataset have a shorter average length. We conclude there is no clear difference in token length between spoken and written image descriptions.

MSTTR. We next look at the richness of the vocabulary used by the crowd workers. Following Drieman's work, we expected that written descriptions would have a higher type-token ratio than spoken descriptions. This expectation is not borne out by the data. The MSTTR score for the Places data falls between the scores for MS COCO and Flickr30K. A possible explanation for this result is that spoken language is typically produced without any preparation, which leads speakers to 'fall back' on a more basic vocabulary. But with the Places dataset, participants could think of a description before they pressed the 'record' button, alleviating cognitive constraints on language production.

Adjectives and prepositions. For the remaining features, we report the average number of occurrences per description, as well as per 1000 tokens. Based on Drieman's work, we thought that attributive adjectives might occur more in written descriptions, but when we look at Table 3, we find a mixed result: spoken descriptions contain *more* attributive adjectives per description, but *fewer* attributive adjectives per 1000 tokens than the written descriptions in the Flickr30K dataset. This is possible because the spoken descriptions are longer than the written ones. We conclude that there is no clear difference between written and spoken descriptions in the use of attributive adjectives. We draw the same conclusion for the use of prepositions.

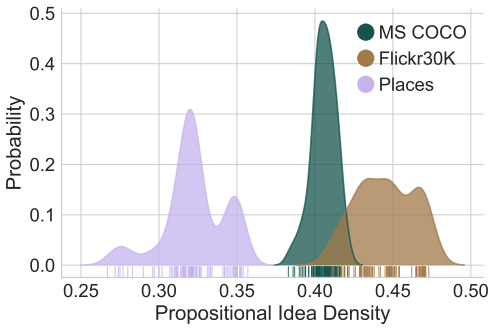

Figure 2: Distribution of the Propositional Idea Density scores for each of the three datasets, computed over $3 \cdot 100$ sets of 1000 descriptions. The lines on the x-axis show individual scores.

Adverbs and other features. We observe that spoken descriptions contain more adverbs than written ones; three times more adverb tokens than MS COCO, and almost four

[9]This example was taken from (Brown et al., 2008).

[10]We use CPIDR version 3.2.3738.41169 on OS X 10.13.2, using Wine version 1.8-rc4.

			TokLen		DescLen		Attributives		Adverbs		Prepositions	
Name	#Desc	#Tok	Syll	Char	Syll	Tok	Desc	‰	Desc	‰	Desc	‰
Written	5,070	52,548	1.47	4.60	15.22	10.36	0.52	50.37	0.22	21.56	1.91	184.03
Spoken	4,604	57,805	1.49	4.58	18.70	12.56	0.50	39.51	0.67	52.76	1.83	144.69

		Consciousness		Self-reference		Allness		Negations		PseudoQuant	
Name	MSTTR	Desc	‰	Desc	‰	Desc	‰	Desc	‰	Desc	‰
Written	0.39	0.01	0.84	0.00	0.04	0	0.04	0.00	0.21	0.02	1.69
Spoken	0.37	0.03	2.22	0.02	1.53	0	0.33	0.01	0.79	0.06	4.78

Table 4: Results for our analysis of the Dutch spoken and written descriptions.

times more than Flickr30K. The same holds for consciousness-of-projection terms, self-reference terms, positive allness terms, negations, and pseudo-quantifiers: all these kinds of terms are used more often in spoken than in written image descriptions.

Propositional idea density. Figure 2 shows the distribution of Propositional Idea Density scores for each of the three datasets, visualized using Kernel Density Estimation. We computed the PID scores over 100 samples of 1000 descriptions for each dataset. We observe that the spoken descriptions have a lower PID than both written datasets, confirming the hypothesis that spoken descriptions use more words to convey the same amount of propositional information. Of course, the extra-propositional information may be useful as well, e.g. to convey pragmatic messages. Future research should look into whether users prefer the spoken or the written variant.

5.2 Dutch results

Table 4 shows the results for the Dutch descriptions. As with the English descriptions, we observe that the spoken descriptions are longer than their written counterparts, albeit to a lesser extent. Whereas the English spoken descriptions were almost twice as long as the written descriptions, the Dutch spoken descriptions are only two tokens longer on average.

Token length and MSTTR. We do not find any major differences in terms of token length or mean-segmental type-token ratios. The spoken descriptions *are* slightly less diverse, but not by a large margin. Unlike the English spoken data, the participants for the Dutch spoken data did not have any time to prepare, since the experiment immediately started recording as the picture was presented. We hypothesize that the differences that Drieman found might have been due to the length of the spoken and written samples, and that with a description spanning multiple sentences, speakers are perhaps more likely to repeat themselves, leading to less diversity in their descriptions.[11]

Adjectives and prepositions. In contrast to the English descriptions, we do observe a difference in the use of attributive adjectives between spoken and written descriptions. Written description contain slightly more attributive adjectives per description (even though written descriptions are shorter on average), and significantly more attributive adjectives per 1000 tokens. We also find that written descriptions contain more prepositions than spoken descriptions. These findings are in line with Drieman's original results.

Adverbs and other features. We find that spoken descriptions contain more than twice as many adverbs than written descriptions, mirroring the results for English. And, just like in English, we find that spoken descriptions also contain more negations, pseudo-quantifiers, and consciousness-of-projection, self-reference, and allness terms.

Propositional idea density. We also computed the Propositional Idea Density for both written and spoken descriptions, but we found little difference between the two: 0.44 for written descriptions versus 0.46 for their spoken counterparts. This is a far cry from the highly contrastive results we found for English. We conclude that there is no clear difference for Dutch spoken and written descriptions, though we should note that Marckx (2017) translated the rules to compute propositional idea density

[11] We did use normalized rather than raw spoken descriptions in our analysis, but the entire corpus of spoken Dutch descriptions contains only 139 repetitions/false starts, which is unlikely to have a strong effect over 57K+ tokens.

from English to Dutch. It may be the case that the Dutch PID rater overlooked linguistic constructions for communicating propositional ideas that only exist in Dutch.

5.3 Summary and discussion of our findings

Looking at the results for both Dutch and English, we have found that: (1) Spoken descriptions are likely to be longer than written descriptions and, in English, seem to have a lower propositional information density than written descriptions. (2) Spoken descriptions contain more adverbs than written descriptions. (3) Spoken descriptions contain more pseudo-quantifiers and allness terms. (4) Speakers have a bigger tendency to "show themselves" in their descriptions than writers, who are less *involved* (in the sense of Chafe and Danielewicz 1987). We can see this in the use of more consciousness-of-projection and self-reference terms. Akinnaso (1982) calls this *egocentric language*, indicating "that the observed is in part a function of the observer" (p. 102). It has been shown that negations in image descriptions often reflect the author's expectations about the image they are describing (van Miltenburg et al., 2016).

Negative findings. Some of the 'negative' findings (where, unlike earlier work, we find no difference between spoken and written language) may be explained in functionalist terms. For example, token length may not be a function of spoken versus written language, but rather of *register*; abstract or formal language tends to use longer words than concrete or informal language (Feng et al., 2011). Another explanation comes from the fact that Drieman (1962a) used *paintings* as a stimuli, which also come with a particular vocabulary, whereas MS COCO, Flickr30K, and the Places Audio Caption corpus use real-life, everyday photographs, which may not elicit the same kind of expert language.[12]

Limitations. There are two main threats to the validity of this study. First, the English descriptions from the Places Audio Caption Corpus have been automatically transcribed, with a 20% word error rate. This means that there may be biases in which words are recognized from the audio and which are not. McKoskey and Boley (2000) show that short words and function words are often confused by Automatic Speech Recognition (ASR) systems with other function words, brief periods of silence, background noise, or filled pauses. Fosler-Lussier and Morgan (1999) also show that infrequent words are less likely to be recognized correctly. See Goldwater et al. (2010) for a survey and further analysis of ASR errors.

Second, for both comparisons (Dutch and English), the images differ between the spoken and the written descriptions. The use of different images may have an effect on the length of the descriptions, for example if one set of images is more complex, or contains more objects per image than another. The length difference in English between spoken and written descriptions is much larger than in Dutch, making the impressive average of 21 tokens per description looking more like a peculiarity of the Places dataset, than a general property of spoken descriptions. Using different sets of images may also influence the word distribution, though perhaps more in terms of content words, than in the way participants report their observations. It seems to us that the use of self-reference terms, for example, should be relatively independent from the contents of the images. Finally, this study is limited in scope. We have only focused on descriptions in Dutch and English, two Germanic languages. It is an open question whether the differences we have found also hold up in other languages.

6 Conclusion and future research

We performed an exploratory study to find differences between spoken and written image descriptions in both Dutch and English. We found four main differences, summarized in the previous section. Where should we go from here? We offer three directions to consider.

[12] Here is a written sample from Drieman's study:

> "A landscape with rushes —could be the Biesbos. The subject sounds charming, rural: fisherman near the edge of the rushes, — open water, in the background trees and a churchtower. Yet, the impression made by the actual painting is much less charming. There is a suggestion of movement among the rushes in the foreground and also in the clouds — a movement from left to right — investing the calm subject with something ominous and troubled. The bright patches in the sky, too, enhance this impression. Especially the bright horizon, on the right and beyond the churchtower, recalls, by contrast with the dark patches, elsewhere, an atmosphere of thunder."

Controlled replication. As Akinnaso (1982) notes, Drieman's study carefully controlled for (1) the topic of the descriptions; (2) the circumstances in which participants were asked to provide the descriptions; and (3) participants' background and level of linguistic knowledge. Changing any of these factors between the written and spoken condition makes the resulting data less comparable. Because we used existing datasets, we were not able to control for these. Although we believe that our main findings *should* hold up, the only way to know for sure is to carry out a follow-up study. The benefit of this exploratory study is that we have compiled a freely available set of tools to analyze spoken versus written language, and we have narrowed down the potential differences between spoken and written descriptions to four main differences. We can now also begin to study how potential users feel about these differences.

Qualitative analysis. We have limited ourselves to a quantitative analysis of the differences between spoken and written language. The key reason for this is that we do not have any parallel sets of spoken and written descriptions for the same images. However, we would still like to emphasize the importance of looking at individual images and individual descriptions in more detail, as manual inspection may reveal interesting examples and phenomena that are glossed over by automated metrics.

What do users want? Having found differences between spoken and written language, we should now ask ourselves: what kind of descriptions would users of image description technology prefer? Research on this topic goes back to user studies of ALT-text on the internet. For example, Petrie et al. (2005) asked a group of blind people about the type of content they would like to be described. They found that there is no single answer to this question, because descriptions are context dependent. But generally speaking, blind users like to know about objects, buildings, and people; activities; the use of color; the purpose of the image; the emotion and atmosphere; and the location where the picture was taken. Gella and Mitchell (2016) asked a panel of visually impaired users about automatic image captioning, and also found that users want to hear about humor and emotional content (besides concrete, literal content). While these studies are important for our understanding of the needs of blind users, they only focus on *what* should be described, and not so much on *how* images should be described, which is still an open question. Possibly the most interesting feature to explore in the context of this paper is the use of subjective language. Furthermore, the datasets discussed in this paper all use pictures from Flickr, or unspecified images from the web. But Gella and Mitchell found that blind users would also like to have image description technology for personal, news, and social media images. It is unclear how these should be described, and whether these kinds of images would elicit similar differences between spoken and written descriptions.

There are many ways to describe an image. As Vedantam et al. (2015) show, we can collect 50 descriptions for an image and still find meaningful variation. But this variation is not random. By manipulating the image description task, we can identify the different factors influencing the description process. With this paper, we hope to have shown that modality is one of the factors causing descriptions to diverge.

7 Acknowledgements

We thank four anonymous reviewers for their comments and suggestions. Emiel van Miltenburg is supported via the 2013 NWO Spinoza grant awarded to Piek Vossen.

References

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

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, September 5-8. Association for Computational Linguistics.

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, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

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

Cati Brown, Tony Snodgrass, Susan J Kemper, Ruth Herman, and Michael A Covington. 2008. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior research methods*, 40(2):540–545.

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.

Wallace Chafe and Deborah Tannen. 1987. The relation between written and spoken language. *Annual Review of Anthropology*, 16(1):383–407.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622. Association for Computational Linguistics.

Joseph A. DeVito. 1966. Psychogrammatical factors in oral and written discourse by skilled communicators. *Speech Monographs*, 33(1):73–76.

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

Gerard HJ Drieman. 1962b. Differences between written and spoken language: An exploratory study, II. qualitative approach. *Acta Psychologica*, 20:78–100.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Shi Feng, Zhiqiang Cai, Scott A Crossley, and Danielle S McNamara. 2011. Simulating human ratings on word concreteness. In *FLAIRS Conference*.

Eric Fosler-Lussier and Nelson Morgan. 1999. Effects of speaking rate and word frequency on pronunciations in convertional speech. *Speech Communication*, 29(2-4):137–158.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July.

Spandana Gella and Margaret Mitchell. 2016. Residual multiple instance learning for visually impaired image descriptions. In *11th Women in Machine Learning Workshop, in conjunction with NIPS*.

Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. 2010. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181 – 200.

Michael Alexander Kirkwood Halliday. 1989. *Spoken and written language*. Language Education. Oxford University Press. Second edition.

Lester E Harrell. 1957. *A comparison of the development of oral and written language in school-age children*, volume 22 of *Monographs of the Society for Research in Child Development*. Wiley.

David Harwath and James Glass. 2017. Learning word-like units from joint audio-visual analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–517, Vancouver, Canada, July. Association for Computational Linguistics.

David Harwath, Antonio Torralba, and James Glass. 2016. Unsupervised learning of spoken language with visual context. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1858–1866. Curran Associates, Inc.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899, May.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

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

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 271–275. ACM.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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

Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *AAAI*, pages 3574–3580.

David McKoskey and Daniel Boley. 2000. Error analysis of automatic speech recognition using principal direction divisive partitioning. In Ramon López de Mántaras and Enric Plaza, editors, *Machine Learning: ECML 2000*, pages 263–270, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jim Miller and M. M. Jocelyne Fernandez-Vest. 2006. Spoken and written language. In Giuliano Bernini and Marcia L. Schwartz, editors, *Pragmatic organization of discourse in the languages of Europe*, Empirical approaches to language typology. EUROTYP. Berlin ; New York : Mouton de Gruyter.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790, Berlin, Germany, August. Association for Computational Linguistics.

Helen Petrie, Chandra Harrison, and Sundeep Dev. 2005. Describing images on the web: a survey of current practice and prospects for the future. *Proceedings of Human Computer Interaction International (HCII)*, 71.

Dorit Ravid and Ruth A. Berman. 2006. Information density in the development of spoken and written narratives in english and hebrew. *Discourse Processes*, 41(2):117–149.

Patricia Saylor. 2015. Spoke: A framework for building speech-enabled websites. Master's thesis, Massachusetts Institute of Technology.

Gunnel Tottie. 1980. Affixal and non-affixal negation in English: Two systems in (almost) complementary distribution. *Studia linguistica*, 34(2):101–123.

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.

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, August. Association for Computational Linguistics.

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, September. Association for Computational Linguistics.

Emiel van Miltenburg, Ákos Kádar, Ruud Koolen, and Emiel Krahmer. 2018. 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.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the vardial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain, April. Association for Computational Linguistics.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc.