# Measuring the Effect of Conversational Aspects on Machine Translation Quality

**Marlies van der Wees** **Arianna Bisazza** **Christof Monz**
Informatics Institute, University of Amsterdam
{m.e.vanderwees,a.bisazza,c.monz}@uva.nl

## Abstract

Research in statistical machine translation (SMT) is largely driven by formal translation tasks, while translating informal text is much more challenging. In this paper we focus on SMT for the informal genre of dialogues, which has rarely been addressed to date. Concretely, we investigate the effect of dialogue acts, speakers, gender, and text register on SMT quality when translating fictional dialogues. We first create and release a corpus of multilingual movie dialogues annotated with these four dialogue-specific aspects. When measuring translation performance for each of these variables, we find that BLEU fluctuations between their categories are often significantly larger than randomly expected. Following this finding, we hypothesize and show that SMT of fictional dialogues benefits from adaptation towards dialogue acts and registers. Finally, we find that male speakers are harder to translate and use more vulgar language than female speakers, and that vulgarity is often not preserved during translation.

## 1 Introduction

Research in statistical machine translation (SMT) has mostly been driven by formal translation tasks. These are, however, not representative for the abundance of informal data emerging on the Internet, for which state-of-the-art SMT systems perform markedly worse (van der Wees et al., 2015a). Recent years have therefore shown an increasing effort in improving SMT for informal text, for example by normalizing noisy text to more formal text (Bertoldi et al., 2010; Banerjee et al., 2012; Ling et al., 2013a), or by enhancing formal training data with user-generated data (Banerjee et al., 2011; Jehl et al., 2012; Ling et al., 2013b).

In this paper we focus on SMT for dialogues, an informal genre that involves, by definition, multiple speakers, and is thus noticeably different from formal text (Fernández, 2014). Formal text is typically written by a single writer with a clear intention (e.g., informing or persuading), and moreover has been editorially controlled according to standards of language use. In dialogues, on the other hand, different *speakers* have different intentions and language use, affected, for example, by their *gender*. Such variations are reflected by *register*, a term referring to socio-situational language variation (Lee, 2001), and *dialogue acts*, functional actions such as questions or answers (Bunt, 1979).

While these and other dialogue-specific aspects have been analyzed in dialogue research (Schlangen, 2005; Fernández, 2014), their impact on SMT has hardly been studied. A likely explanation is the lack of adequate evaluation data, i.e., parallel conversations annotated with dialogue-specific variabels. In this paper, we take a first step towards investigating the effect of dialogue acts, speakers, gender, and register on SMT performance by measuring their respective impact on annotated dialogues from movie subtitles.

Since movie dialogues are fictional, we can only consider them as an approximation of real face-to-face conversation. However, several corpus-based studies have shown that, while movie dialogues differ from natural spoken dialogues in terms of spontaneity—they exhibit fewer incomplete utterances, hesitations, and repetitions—, they do not differ to a great extent in terms of linguistic features and main





*Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2571–2581, Osaka, Japan, December 11-17 2016.

messages (Forchini, 2009; Forchini, 2012; Dose, 2013). In fact, movie dialogues approximate real face-to-face conversation more than for example SMS and chat, in which media constraints influence the flow of a conversation (Whittaker, 2003; Brennan and Lockridge, 2006). Finally, Danescu-Niculescu-Mizil and Lee (2011) have shown that certain psycholinguistic and gender-specific aspects of language are also observed in fictional dialogues, indicating that conclusions drawn from experiments on fictional dialogues generalize at least partially to real spoken conversations.

The structure and contributions of this paper are as follows: First, in Section 2 we annotate multilingual movie dialogues with four dialogue-specific variables; dialogue acts, speakers, gender, and register level, and we release these annotated corpora. In Section 3 we describe our approach to measure how SMT quality is affected by each of the four dialogue-specific aspects. Next, in Section 4 we use our annotated benchmarks to show that (i) performance fluctuations among the studied dialogue dimensions are larger than randomly expected, (ii) male speakers are harder to translate than female speakers and use more vulgar language, and (iii) vulgarity is often not preserved during translation. Finally, in Section 5 we investigate and confirm the hypothesis that SMT of fictional dialogues benefits from adaptation towards various dialogue acts and registers, indicating that apart from domain adaptation, adaptation to other variables should be considered to improve SMT quality for fictional, and potentially real, dialogues.

## 2 Corpus construction and annotation

To measure the effect of dialogue acts, speakers, gender, and register on SMT performance we need a multilingual dialogue corpus in which utterances are annotated with each of these dialogue aspects. Unfortunately, existing corpora are limited to the English language (Janin et al., 2003; McCowan et al., 2005; Danescu-Niculescu-Mizil and Lee, 2011; Banchs, 2012; Walker et al., 2012) or contain only some of the required annotations (Wang et al., 2016). We therefore first automatically annotate multilingual movie dialogues with the above dialogue dimensions for five language pairs: Arabic-English, Chinese-English, Dutch-English, German-English, and Spanish-English.

For this annotation process we build on two main resources: (i) the OpenSubtitles corpus (Lison and Tiedemann, 2016), containing non-professionally translated subtitles, collected from `www.opensubtitles.org` and cross-lingually aligned using time information, bilingual lexicons, and cognates (Tiedemann, 2008); and (ii) the Internet Movie Script Database (IMSDb)[1], containing English movie scripts with speakers, utterances, and context (e.g., change of scenes). Using these and a number of additional resources we create our annotated corpora as follows:

1. First we collect **speaker-utterance pairs** from IMSDb scripts, based on their respective indentation sizes. We then use the Champollion sentence aligner (Ma, 2006) monolingually to align English subtitles to the English script, and follow the OpenSubtitles alignment links to align foreign subtitles to the English subtitles-script bitext. Finally, we discard the script text from the resulting 'tritext', yielding multilingual speaker-annotated dialogue corpora. Table 1a shows statistics on the average number of speakers and main characters (i.e., speakers with at least 20 utterances) per movie.

2. We learn each speaker's **gender** based on their name's occurrence in a number of online name databases and a list of gender-revealing tags such as 'aunt', 'boy', or 'grandma'. Annotations are available for ∼58% of the utterances, and the average female-to-male ratio is 1:1.7, see Table 1b.

3. We heuristically detect the **dialogue act** of each source-language utterance, considering *questions*, *exclamations* and *declaratives*. Distributions of these dialogue acts differ between language pairs, see Table 1c, but make up on average 28%, 9%, and 63% of the corpora, respectively.

4. We define a **register** label, based on the fraction of colloquial and vulgar expressions in an utterance, according to meta-information from an online dictionary[2]. We consider three register levels: *vulgar*, *colloquial*, and *neutral*, comprising circa 10%, 68%, and 22% of the corpora, respectively. Distribution statistics per language pair are shown in Table 1d.

[1] `www.imsdb.com`
[2] `www.dict.cc`

| | a) Avg. speaker statistics | | b) Gender statistics | | | c) Dialogue act statistics | | | d) Register statistics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Lang. pair | #Speakers | #Main chars | %M | %F | %Unk. | %Ques. | %Excl. | %Decl. | %Vulg. | %Coll. | %Neut. |
| AR → EN | 44.6 | 5.6 | 36.2 | 20.4 | 43.4 | 29.8 | 6.0 | 64.2 | 10.8 | 67.1 | 22.1 |
| DE ↔ EN | 47.3 | 5.9 | 36.1 | 19.8 | 44.1 | 27.2 | 12.2 | 60.6 | 10.3 | 67.7 | 22.0 |
| ES ↔ EN | 42.5 | 6.0 | 37.2 | 22.9 | 39.9 | 28.4 | 11.2 | 60.4 | 10.6 | 67.7 | 21.7 |
| NL ↔ EN | 43.8 | 6.0 | 36.5 | 22.3 | 41.2 | 29.0 | 4.2 | 66.8 | 10.1 | 68.3 | 21.6 |
| ZH → EN | 42.3 | 5.6 | 36.9 | 21.1 | 42.0 | 28.1 | 10.6 | 61.3 | 10.7 | 68.9 | 20.4 |

Table 1: Annotation distributions of the dialogue benchmarks. a) Main characters are speakers with 20 or more utterances. b) Uncertain gender annotations are labeled 'unknown'. c) Annotated dialogue acts: questions, exclamations, declaratives. d) Register levels: vulgar, colloquial, neutral.

**Post-processing and annotation quality.** The above described alignment and annotation process is done fully automatically, making it prone to errors. We therefore increase the alignment and annotation quality of our corpus by taking a number of measures: First, *before* running the Champollion aligner, we remove context information such as '[moaning]', '[clapping]' or '[chuckles]', which is most prevalent in, but not limited to, subtitles created for hearing-impaired people. In addition, we remove subtitle-specific tokens indicating continuation of a sentence on the next screen or switches in speaker turns, yielding more fluent and less fragmented utterances.

Next, *after* running the alignment process, we favor high-quality alignments by selecting only movies or movie versions (OpenSubtitles typically contains several alternative versions for a single movie (Tiedemann, 2016)) that meet the following criteria; (i) sentence lengths between both language pairs are sufficiently close, (ii) the number of sentences for which ambiguous speakers have been aligned does not exceed a given threshold, (iii) the letter distribution is sufficiently similar to the average distribution of the language. By enforcing these quality standards, we respectively reduce the number of OpenSubtitles alignment errors, Champollion alignment errors, and OCR errors. Finally, we remove utterances with ambiguous speaker labels as these are caused by erroneous Champollion alignments.

Despite efforts to improve alignment quality, our corpus still contains some incorrect alignments. To quantify these, and to verify the correctness of the automatic annotations, we manually inspect randomly selected fragments across different language pairs and movies. Based on evaluation of a sample of 120 utterances, we estimate a final alignment accuracy of 92.5%. In addition, Table 2 shows confusion matrices for manual versus automatic annotation of gender, dialogue acts, and register for the 120 selected utterances. With the overall annotation agreement per variable ranging from 85% to 97.5%, we find that our automatic annotation strategies are very accurate. Disagreement between manual and automatic annotations occurs mostly for speakers labeled with unknown gender, and between the register levels colloquial and neutral, indicating that these categories can benefit from more advanced annotation methods. For example, to better distinguish colloquial and neutral register levels, one could exploit sentence length or language model perplexity.

| Gender Annot. | | Automatic M | F | U | Total |
|---|---|---|---|---|---|
| Manual | **M** | 42 | 0 | 8 | 50 |
| | **F** | 1 | 27 | 3 | 31 |
| | **U** | 2 | 2 | 35 | 39 |
| Total | | 45 | 29 | 46 | 120 |

| Dial.act Annot. | | Automatic Q | E | D | Total |
|---|---|---|---|---|---|
| Manual | **Q** | 28 | 0 | 0 | 28 |
| | **E** | 1 | 8 | 0 | 9 |
| | **D** | 1 | 1 | 81 | 83 |
| Total | | 30 | 9 | 81 | 120 |

| Register Annot. | | Automatic V | C | N | Total |
|---|---|---|---|---|---|
| Manual | **V** | 9 | 0 | 0 | 9 |
| | **C** | 0 | 74 | 6 | 80 |
| | **N** | 0 | 12 | 19 | 31 |
| Total | | 9 | 86 | 25 | 120 |

Table 2: Confusion matrices for manual and automatic annotation of gender (left; M=male, F=female, U=unknown), dialogue acts (center; Q=questions, E=exclamations, D=declaratives), and register levels (right; V=vulgar, C=colloquial, N=neutral).

*a) Original German-English OpenSubtitles alignment*

| **German subtitles** | **English subtitles** |
|---|---|
| Erstklassig! | Classic. |
| Bilanz der Werbekampagne... minus 347 Pfund. | Profit from major sales push, minus £347. |
| Soll ich... dir einen Cappuccino holen? | Shall I go and get you a cappuccino? // You know, ease the pain a bit. // Yeah. |
| Als Seelentröster? | |
| Ja. | Yeah. |
| Lieber nur einen halben. | Better make it a half. |
| Mehr kann ich mir nicht leisten. | All I can afford. |
| Logisch. | Get your logic. |
| Demi-Cappu. // Kommt sofort. | Demi-cappu coming right up. |

*b) Annotated German-English dialogue*

| **German utterance** | **English utterance** | **Annotations** |
|---|---|---|
| Erstklassig! Bilanz der Werbekampagne minus 347 Pfund. | Classic. Profit from major sales push, minus £347. | William, M, neutral, exclamation |
| Soll ich dir einen Cappuccino holen? | Shall I go and get you a cappuccino? You know, ease the pain a bit. | Martin, M, coll., question |
| Ja. Lieber nur einen halben. Mehr kann ich mir nicht leisten. | Yeah. Better make it a half. All I can afford. | William, M, coll., declarative |
| Logisch. Demi-Cappu. Kommt sofort. | Get your logic. Demi-cappu coming right up. | Martin, M, coll., declarative |

Table 3: Example dialogue from Notting Hill; a) original sentences in the OpenSubtitles corpus, where // indicates a sentence boundary in many-to-one or one-to-many alignments, and b) final annotated utterances generated in our annotation pipeline, annotated with speaker (William, Martin), gender (M=male, F=female), register level (neutral, colloquial, vulgar) and dialogue act (declarative, exclamation, question). Note that sentences pairs from the original corpus are often merged in the Champollion alignment process, and that erroneous OpenSubtitles alignments are not corrected.

Table 3 shows an example dialogue with annotations and its original form in the OpenSubtitles corpus. Note that our annotated corpora differ from OpenSubtitles since only actual dialogues are included (i.e., no context), many erroneously aligned sentence pairs have been removed, and utterances are longer and less fragmented. The latter is a result of the Champollion alignment process. Since sentences in the IMSDb scripts are typically longer than those in OpenSubtitles, Champollion regularly enforces one-to-many alignments. Following the OpenSubtitles-internal alignment links then yields a large number of many-to-many alignments in which subtitles get merged into longer utterances.

Finally, table 4a lists the statistics of the benchmarks which we use in this paper and make available for download[3]. While the remainder of this paper uses the annotated fictional dialogues to analyze the impact of dialogue-specific aspects on SMT, we believe that our data set may also help to advance dialogue research—today largely confined to the English language—in a multilingual scenario.

## 3 Measuring dialogue effects on SMT

In this section we measure the effect of dialogue dimensions on SMT performance of fictional dialogues. To this end, we quantify BLEU (Papineni et al., 2002) fluctuations between differences in dialogue acts, speakers, gender, and register, and we determine whether the observed fluctuations are larger than randomly expected.

[3] http://ilps.science.uva.nl/resources/movie-dialogues

| | a) Evaluation data | | b) Training data | |
|---|---|---|---|---|
| Languages | #Movies | #Utterances | #Lines | #EN tokens |
| AR → EN | 187 | 94K | 12.3M | 122M |
| DE ↔ EN | 220 | 123K | 9.7M | 85M |
| ES ↔ EN | 161 | 87K | 16.3M | 156M |
| NL ↔ EN | 238 | 129K | 17.9M | 171M |
| ZH → EN | 211 | 107K | 6.3M | 59M |

Table 4: Specifications of parallel training and evaluation data. Training data consists of OpenSubtitles corpora, evaluation data consists of speaker-annotated dialogues.

### 3.1 Basic experimental setup

We run our experiments using an in-house phrase-based SMT system similar to Moses (Koehn et al., 2007), with features including lexicalized reordering, linear distortion with limit 5, and lexical weighting. Our systems are trained on 59M–171M tokens (depending on the language pair, see Table 4b) of unannotated OpenSubtitles corpora. We use Kneser-Ney smoothed 5-gram language models (500M–1.7B tokens, depending on the language pair) that linearly interpolate OpenSubtitles with various LDC and WMT corpora using weights optimized on a held-out set of OpenSubtitles data. Systems are tuned using pairwise ranking optimization (PRO) (Hopkins and May, 2011) on a different held-out OpenSubtitles set. The resulting systems are thus at all levels adapted to the movie dialogues translation task rather than the general domain.

### 3.2 Approximate randomization testing

When translating dialogues, we naturally observe *some* BLEU variations across categories such as different dialogue acts or speakers. An important question is whether the observed differences are to be expected (the null hypothesis), or whether they are indicators that one category is truly harder to translate than another (the alternative hypothesis). We test this hypothesis with an approximate randomization approach (Edgington, 1969; Noreen, 1989).

While approximate randomization (also known as approximate permutation) is often used to compare the mean and variance of *two* groups, it can be adapted to our setting with multiple categories. To this end, we compute BLEU for each of the categories in a dialogue variable (e.g., vulgar, colloquial, and neutral utterances for the dialogue variable of register level). Next, we randomly permute category labels over utterances, following the original distribution of utterances per category, and we recompute BLEU for the randomized labels.

As our test statistic of interest, we define and measure the *mean absolute BLEU difference*, which captures BLEU fluctuations between categories:

$$\text{MBD} = \frac{2}{|S|^2 - |S|} \sum_{i=1}^{|S|} \sum_{j=i+1}^{|S|} \mid \text{BLEU}_i - \text{BLEU}_j \mid \quad (1)$$

Here $S$ is the set of categories for a given dialogue variable (e.g., $S_{\text{register}} = \{\text{vulgar, colloquial, neutral}\}$), and $\text{BLEU}_i$ the BLEU score for category $i$. Each pair of categories $(i, j)$ is compared exactly once in terms of BLEU scores. Note that MBD is a specific instance of *mean absolute difference* (MD) or *Gini mean absolute difference* (GMD), a measure of statistical dispersion which has shown to be superior to other common statistical dispersion measures such as variance, standard deviation and interquartile range (Yitzhaki, 2003).

Next, we compute the p-value by counting how often (in a total of 1,000 permutations) we observe an MBD value that is at least as extreme as the one observed for the real categories. If for a given dialogue variable $p \leq 0.05$ or $p \leq 0.01$, we conclude that this variable has a weakly or strongly significant impact on SMT quality, respectively.

For dialogue acts, gender and register we permute labels over the entire benchmark. For speakers we only permute labels within each movie since inter-movie variations in BLEU are affected by many other factors (e.g., script writers, translators, movie genre) which distract from the impact of speakers. In addition, when computing speaker-specific BLEU, we only include main characters (i.e., speakers with at least 20 utterances) to avoid BLEU's instability on small documents .

| | a) BLEU per dialogue act | | | | b) Speaker-ssMBD | c) BLEU per gender | | | d) BLEU per register | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Languages | Quest. | Excl. | Decl. | MBD | | Male | Female | MBD | Vulg. | Coll. | Neut. | MBD |
| AR → EN | 23.1 | 20.3 | 19.3 | 2.5▲ | 14.9% | 20.4 | 22.0 | 1.6▲ | 17.2 | 21.3 | 19.7 | 2.8▲ |
| DE → EN | 24.0 | 20.9 | 21.4 | 2.0▲ | 22.3% | 21.4 | 22.7 | 1.2▲ | 17.7 | 21.9 | 24.7 | 4.6▲ |
| ES → EN | 28.9 | 25.7 | 26.8 | 2.1▲ | 18.0% | 26.7 | 28.0 | 1.3▲ | 24.4 | 27.4 | 28.8 | 2.9▲ |
| NL → EN | 26.7 | 29.0 | 23.7 | 3.5▲ | 23.9% | 23.9 | 26.8 | 2.9▲ | 20.8 | 24.8 | 26.7 | 4.0▲ |
| ZH → EN | 15.3 | 15.2 | 12.9 | 1.6▲ | 16.6% | 12.8 | 14.4 | 1.6▲ | 10.9 | 13.4 | 13.1 | 1.7▲ |
| EN → DE | 17.7 | 18.5 | 16.5 | 1.3▲ | 15.9% | 16.7 | 17.6 | 0.9▲ | 13.6 | 16.6 | 19.9 | 4.2▲ |
| EN → ES | 16.8 | 16.5 | 21.5 | 3.3▲ | 13.7% | 18.9 | 19.7 | 0.8▲ | 17.2 | 19.2 | 21.0 | 2.6▲ |
| EN → NL | 25.5 | 22.7 | 24.6 | 1.8▲ | 20.6% | 23.9 | 26.5 | 2.6▲ | 21.4 | 24.6 | 26.3 | 3.3▲ |

Table 5: BLEU for dialogue acts, speakers, gender, and register, translated using baseline SMT trained and tuned on OpenSubtitles corpora. MBD: mean absolute BLEU difference, see Equation (1), all statistically significant at $p \leq 0.01$ (▲). ssMBD: percentage of movies with statistically significant speaker-MBD at $p \leq 0.05$.

## 4 Results

In this section we discuss the observed BLEU fluctuations (see Table 5) for our four dialogue variables of interest, guided by Spanish-to-English and English-to-German examples in Table 6, to which we provide pointers (EX#) in the text.

### 4.1 The effect of dialogue acts on SMT quality

As shown in Table 5a, there are substantial performance fluctuations between dialogue acts for all language pairs. However, there is no consistent pattern between different languages. For instance, we observe punctuation errors (EX1) for ES↔EN, and verb drop (EX2) and wrong word order (EX3) for EN→DE. This makes it particularly interesting to further investigate how dialogue acts can be exploited to improve translation quality of (fictional) dialogues. Improving SMT for the dialogue acts under consideration resembles cross-lingual question answering (Tiedemann, 2009). However, when considering finer dialogue act granularities, it may be profitable to exploit context information, which is not used in our current SMT setup.

### 4.2 The effect of speakers on SMT quality

In Table 5b we report the percentage of movies per language pair for which the observed MBD is statistically significant at $p \leq 0.05$, which is 18.6% on average. Since there are too many speakers to report individual BLEU scores, we randomly select 100 German-English movies, and compute for each of these $\Delta$MBD as the difference between MBD for real speakers and the average MBD for randomized labels:

$$\Delta\text{MBD} = \text{MBD}_{real} - \overline{\text{MBD}}_{random} \quad (2)$$

Figure 1 shows that inter-speaker BLEU fluctuations among real speakers are often larger than inter-speaker BLEU fluctuations among randomized speaker tags. These findings suggest that, while domain adaptation is an established task in SMT, conversational SMT may benefit—at least for the fraction of movies with statistically significant speaker differences—from a fine-grained adaptation at the speaker

| Annotations | ES source | ES→EN SMT output | EN reference |
|---|---|---|---|
| S1, M, neutral, declarative | se acabaron los días de olvidar, han empezado los de recordar. | the days are over, have begun to forget them to remember. | the days of forgetting are over. the days of remembering have begun. |
| S2, F, colloquial, question | ¿sabes qué pareces cuando hablas así? | you know what you look like when you talk like that? | know when you go on what you sound like? |
| S1, M, vulgar, declarative | un *j***do*(EX4) hombre sensato. | a sensible man.(EX6) | i sound like a sensible *f***ing* man. |
| S2, F, colloquial, excl. | un pato. ¡cuac, cuac! | a duck. [quack, quack!(EX1) | you sound like a duck. quack, quack |

| Annotations | EN source | EN→DE SMT output | DE reference |
|---|---|---|---|
| S1, M, neutral, declarative | the days of forgettin' are over. the days of remembering have begun. | die tage von vergisst sind vorbei. die tage von an(EX2) haben begonnen. | ja, aber jetzt kommen die tage des erinnerns. |
| S2, F, colloquial, question | know when you go on what you sound like? | weiß, wenn du auf(EX2) was du klingst wie?(EX3) | weißt du, wie du klingst? |
| S1, M, vulgar, declarative | i sound like a sensible *f***ing* man. | ich klinge wie ein vernünftig *verd***ter* mann. | wie ein vernünftiger(EX5) mann. |
| S2, F, colloquial, declarative | you sound like a duck. quack, quack | du klingst wie eine ente. quak, quak | nein, wie eine ente! quak, quak, quak! |

Table 6: Censored ES→EN (top) and EN→DE (bottom) translation examples of an annotated dialogue, originating from Pulp Fiction and involving two speakers: Pumpkin (S1, M=male) and Honeybunny (S2, F=female). Examples of phenomena marked with (EX#) are discussed in Sections 4.1–4.4.

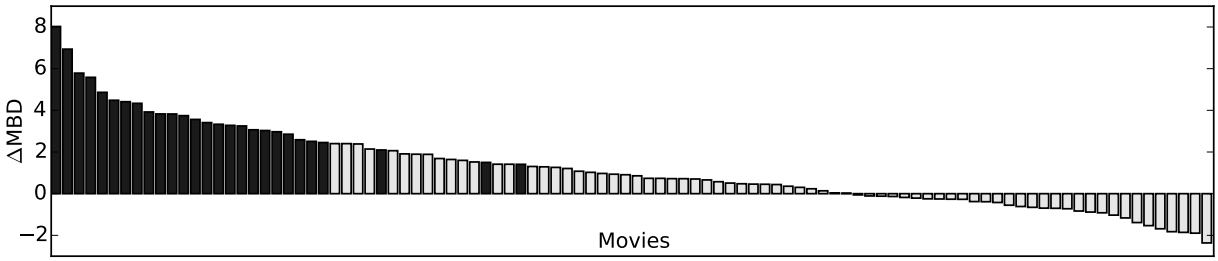


Figure 1: ΔMBD (see Section 4.2) for 100 randomly selected German-English benchmark movies. Black bars indicate movies with statistical significant positive ΔMBD at $p \leq 0.05$.

level, as proven successful in speech recognition research (Shinoda, 2011), and related to recent work on personalizing machine translation (Mirkin et al., 2015; Mirkin and Meunier, 2015).

Finally, since our analysis is carried out on fictional dialogues, it may be worth investigating to what extent BLEU scores fluctuate between actors or script writers rather than only characters, however this requires additional annotation.

### 4.3 The effect of gender on SMT quality

Table 5c shows that BLEU scores per gender follow a similar pattern in all language pairs. Male speakers are significantly harder to translate than female speakers, despite the fact that male speakers are likely better represented in the parallel OpenSubtitles data, based on the male-to-female ratio in our evaluation sets. However, we find that female utterances are better covered by the language model, with perplexity values on average 8% higher for males than females. This finding is consistent with recent work by Wang et al. (2016), who show that SMT can benefit from gender-adapted language models but do not provide gender-specific BLEU scores. Gender differences in movie dialogues have also been reported by Danescu-Niculescu-Mizil and Lee (2011), who show that characters adapt their language easier to females than to males.

However, we have to be careful drawing conclusions about the impact of gender on real spoken dialogues from our observations on fictional dialogues. Since the vast majority of movie scripts are written by men (Lauzen, 2016), our findings reflect differences between language of male and female characters as perceived by male writers. The observed BLEU differences might therefore be based on stereotypes rather than real gender differences. On the other hand, a tremendously large body of work has studied gender and language or discourse (Tannen, 1994; Wodak, 1997; Holmes and Meyerhoff, 2008, among others), indicating that these concepts are closely intertwined. To the best of our knowledge, our work is the first to study the impact of gender on SMT, albeit for scripted dialogues, and we believe that meta-information about a speaker's gender is also a potential source to customize SMT for real dialogues.

### 4.4 The effect of register on SMT quality

The results per register (Table 5d) show that SMT quality is worst for vulgar utterances and generally best for neutral sentences. While consistent with previous findings that informal language is hard to translate (van der Wees et al., 2015a), this observation cannot solely be attributed to poor model coverage, since colloquial and vulgar language are well-covered in our OpenSubtitles-trained systems.

When manually inspecting human translations for vulgar expressions, we find that these vary from literal (EX4) to very nuanced (EX5) translations, yielding inconsistent SMT output. We also observe that vulgarity is often not preserved in (both human and machine) translation (EX6): A comparison of vulgarity scores shows that, while vulgarity sometimes increases, the number of vulgar utterances in the SMT output is on average 35% lower than in the reference set.

Finally, since poor SMT quality is observed for both male characters and vulgar language, we hypothesize that the two might co-occur. Indeed, the average vulgarity scores for males are 64% higher than for females, which may in part explain the observed SMT quality between genders.

## 5 Preliminary adaptation towards dialogue variables

We observed that BLEU scores significantly fluctuate between differences along dialogue dimensions. This finding suggests that SMT for fictional dialogues may benefit from adaptation towards different categories along these dimensions. To verify this hypothesis we run a number of adaptation experiments, in which we adapt our baseline SMT systems towards different dialogue acts and different registers—two dialogue aspects which can be computed straightforwardly for the unannotated training corpora.

We adapt our systems at two levels: First, we create category-specific language models by interpolating our general movie dialogue language model with a language model trained on only the most relevant subset of the bitext's target side. We determine relevant sentences by applying the same annotation guidelines that were used for annotation of the benchmarks (Section 2). Second, we tune our systems on held-out sets selected according to the same criteria, thus comprising category-specific data.

We run adaptation experiments for the language pairs with the largest observed MBD; Dutch-English, English-Spanish, and Arabic-English for dialogue acts, and German-English, English-German, and Dutch-English for register. Note that the aim of our adaptation experiments is to verify whether SMT performance can benefit from a simple adaptation approach at the fine-grained level of different dialogue-specific aspects, rather than presenting a novel SMT adaptation approach.

The results of our adaptation experiments are shown in Table 7. The first observation we can make is that the adapted systems result in substantially lower mean absolute BLEU differences (MBD) for both dialogue dimensions—dialogue act and register level—for all language pairs except Arabic-English. This means that most of the adapted systems generate translations of more uniform quality with a lower degree of fluctuation in BLEU. Further, the BLEU scores for the individual categories of both dialogue dimensions show that the lower MBD scores are due to statistically significant improvements for most of the dialogue acts and registers. The only case where our simple adaptation method causes a statistically significant drop in BLEU is for the translation of questions from Dutch into English. Vulgar and colloquial language profit particularly well from language model adaptation, while results for question and exclamation marks are more variable between language pairs. Finally, we would like to emphasize that we do not claim that the simple adaption method used here constitutes the best adaptation approach

| | MBD | | Questions | | | Exclamations | | | Declaratives | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Language pair | Base. | Adapt. | Base. | Adapt. | Diff. | Base. | Adapt. | Diff. | Base. | Adapt. | Diff. |
| NL → EN | 3.5 | 3.2 | 26.7 | 26.5 | –0.2$^{\triangledown}$ | 29.0 | 28.9 | –0.1 | 23.7 | 24.1 | +0.4$^{\blacktriangle}$ |
| EN → ES | 3.3 | 2.8 | 16.8 | 17.5 | +0.7$^{\blacktriangle}$ | 16.5 | 17.3 | +0.8$^{\blacktriangle}$ | 21.5 | 21.5 | 0.0 |
| AR → EN | 2.5 | 2.6 | 23.1 | 24.3 | +1.2$^{\blacktriangle}$ | 20.3 | 20.4 | +0.1 | 19.3 | 20.9 | +1.6$^{\blacktriangle}$ |

| | MBD | | Vulgar | | | Colloquial | | | Neutral | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Language pair | Base. | Adapt. | Base. | Adapt. | Diff. | Base. | Adapt. | Diff. | Base. | Adapt. | Diff. |
| DE → EN | 4.6 | 4.2 | 17.7 | 18.4 | +0.7$^{\blacktriangle}$ | 21.9 | 22.7 | +0.8$^{\blacktriangle}$ | 24.7 | 24.7 | 0.0 |
| EN → DE | 4.2 | 3.8 | 13.6 | 14.6 | +1.0$^{\blacktriangle}$ | 16.6 | 17.8 | +1.2$^{\blacktriangle}$ | 19.9 | 20.3 | +0.4$^{\blacktriangle}$ |
| NL → EN | 4.0 | 2.8 | 20.8 | 23.1 | +2.3$^{\blacktriangle}$ | 24.8 | 25.6 | +0.8$^{\blacktriangle}$ | 26.7 | 27.3 | +0.6$^{\blacktriangle}$ |

Table 7: Results of adaptation experiments. Top: adaptation towards dialogue acts for the 3 language pairs with the largest mean absolute BLEU difference (MBD, see Equation (1)) between dialogue acts. Bottom: adaptation towards registers for the 3 language pairs with the largest MBD between register levels. Statistical significance against the baseline at $p \leq 0.05$ ($\triangle$/$\triangledown$) and $p \leq 0.01$ ($\blacktriangle$/$\blacktriangledown$) is measured using approximate randomization (Riezler and Maxwell, 2005).

for dialogue-specific phenomena, but rather that already a simple adaptation approach can benefit from our dialogue-specific annotations.

## 6 Conclusions and implications

While SMT research has mostly been driven by formal translation tasks, very little work has been reported on SMT for informal genres such as dialogues, a genre that differs substantially from formal text and thus poses different translation challenges. Following the previous finding that genre and topic affect SMT differently (van der Wees et al., 2015b), we have in this paper analyzed the impact of dialogue-specific aspects in SMT for fictional dialogues. We created and released a movie-dialogue benchmark in which utterances are annotated with dialogue acts, speakers, gender, and register, and we studied the effect of these four variables on SMT performance.

Our analysis shows that BLEU fluctuations for all variables are often significantly larger than randomly expected. When looking at specific dialogue aspects, we found that the register level has a significant impact on translation quality, with translations of vulgar utterances being of substantially lower quality than neutral or even colloquial utterances for all language pairs under consideration. Similarly we found large variations in translation quality between different dialogue acts, although we did not detect a consistent pattern between different languages; e.g., questions can be more difficult to translate than exclamations for one language pair, while the reverse is true for another language pair.

These findings suggest that conversational SMT may benefit from adaptation at fine-grained levels. We tested and confirmed this hypothesis in a series of simple adaptation experiments.

Finally, we found that male speakers are harder to translate and use more vulgar language than female speakers, and that vulgarity is often not preserved during translation. While our analyses are carried out on fictional dialogues, we believe that our findings generalize at least partially to other types of dialogues, and are thus valuable for advancing conversational SMT.

## Acknowledgements

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.213. We thank Raquel Fernández for sharing valuable insights from dialogue research, Tobias Schnabel for providing feedback on the approximate randomization approach, and the anonymous reviewers for their thoughtful comments.

## References


Rafael E. Banchs. 2012. Movie-DiC: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 203–207.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2011. Domain adaptation in statistical machine translation of user-forum data using component level mixture modelling. In *Proceedings of the XIII Machine Translation Summit*, pages 285–292.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2012. Domain adaptation in SMT of user-generated forum content guided by OOV word reduction: Normalization and/or supplementary data. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 169–176.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2010. Statistical machine translation of texts with misspelled words. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 412–419.

S.E. Brennan and Calion B. Lockridge. 2006. Computer-mediated communication: A cognitive science approach. In *Encyclopedia of language and linguistics,*, pages 775–780. Elsevier Ltd., Oxford, UK.

Harry Bunt. 1979. Conversational principles in question-answer dialogues. In *Zur Theorie der Frage*, pages 119–141. Narr Verlag.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87.

Stefanie Dose. 2013. Flipping the script: A corpus of american television series (CATS) for corpus-based language learning and teaching. *Corpus Linguistics and Variation in English: Focus on Non-native English*, 13.

Eugene S. Edgington. 1969. Approximate randomization tests. *The Journal of Psychology*, 72(2):143–149.

Raquel Fernández. 2014. Dialogue. In *The Oxford Handbook of Computational Linguistics (2 ed.)*. Oxford University Press.

Pierfranca Forchini. 2009. Spontaneity reloaded: American face-to-face and movie conversation compared. In *Corpus Linguistics*.

Pierfranca Forchini. 2012. *Movie language revisited. Evidence from multi-dimensional analysis and corpora*. Peter Lang, Internationaler Verlag der Wissenschaften.

Janet Holmes and Miriam Meyerhoff. 2008. *The handbook of language and gender*, volume 25. John Wiley & Sons.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI meeting corpus. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, pages I–364. IEEE.

Laura Jehl, Felix Hieber, and Stefan Riezler. 2012. Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 410–421.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

Martha M. Lauzen. 2016. The celluloid ceiling: Behind-the-scenes employment of women on the top 100, 250, and 500 films of 2015. Technical report, Center for the study of women in television and film.

David Y.W. Lee. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3):37–72, September.

Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. 2013a. Paraphrasing 4 microblog normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 73–84.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013b. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 176–186.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation*, pages 489–492.

Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. 2005. The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64.

David Schlangen. 2005. Modelling dialogue: Challenges and approaches. *Künstliche Intelligenz*, 3/05:23–28.

Koichi Shinoda. 2011. Speaker adaptation techniques for automatic speech recognition. In *Proceedings of APSIPA ASC*.

Deborah Tannen. 1994. *Gender and discourse*. Oxford University Press.

Jörg Tiedemann. 2008. Synchronizing translated movie subtitles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1902–1906.

Jörg Tiedemann. 2009. Translating questions for cross-lingual QA. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 112–119.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 3518–3522.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2015a. Five shades of noise: Analyzing machine translation errors in user-generated text. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 28–37.

Marlies van der Wees, Arianna Bisazza, Wouter Weerkamp, and Christof Monz. 2015b. What's in a domain? Analyzing genre and topic differences in statistical machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 560–566.

Marilyn A Walker, Grace I Lin, and Jennifer Sawyer. 2012. An annotated corpus of film dialogue for learning and characterizing character style. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1373–1378.

Longyue Wang, Xiaojun Zhang, Zhaopeng Tuy, Andy Way, and Qun Liu. 2016. Automatic construction of discourse corpora for dialogue translation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2748–2754.

Steve Whittaker. 2003. Theories and methods in mediated communication. In *The Handbook of Discourse Processes*, pages 243–286. Erlbaum.

Ruth Wodak. 1997. *Gender and discourse*. Sage.

Shlomo Yitzhaki. 2003. Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron*, 61(2):285–316.