DiaBLa: A Corpus of Bilingual Spontaneous Written Dialogues for Machine Translation

Rachel Bawden, Eric Bilinski, Thomas Lavergne, Sophie Rosset

DiaBLa: A Corpus of Bilingual Spontaneous Written Dialogues for Machine Translation

Rachel Bawden[1*] **Eric Bilinski**[2] **Thomas Lavergne** [2,3] **Sophie Rosset**[2]

[1]School of Informatics, University of Edinburgh, Scotland
[2]LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[3]Univ. Paris-Sud, Orsay, France

rachel.bawden@ed.ac.uk lastname@limsi.fr

Abstract

We present a new English-French dataset for the evaluation of Machine Translation (MT) for informal, written bilingual dialogue. The test set contains 144 spontaneous dialogues (5,700+ sentences) between native English and French speakers, mediated by one of two neural MT systems in a range of role-play settings. The dialogues are accompanied by fine-grained sentence-level judgments of MT quality, produced by the dialogue participants themselves, as well as by manually normalised versions and reference translations produced *a posteriori*. The motivation for the corpus is two-fold: to provide (i) a unique resource for evaluating MT models, and (ii) a corpus for the analysis of MT-mediated communication. We provide an initial analysis of the corpus to confirm that the participants' judgments reveal perceptible differences in MT quality between the two MT systems used.

1 Introduction

The use of Machine Translation (MT) to translate everyday, written exchanges is becoming increasingly commonplace; translation tools now regularly appear on chat applications and social networking sites to enable cross-lingual communication. MT systems must therefore be able to handle a wide variety of topics, styles and vocabulary. Importantly, the translation of dialogue requires translating sentences coherently with respect to the conversational flow so that all aspects of the exchange, including speaker intent, attitude and style, are correctly communicated (Bawden, 2018).

It is important to have realistic data to evaluate MT models and to guide future MT research for informal, written exchanges. In this article, we present DiaBLa (*Dialogue BiLingue* 'Bilingual Dialogue'), a new dataset of English-French spontaneous written dialogues mediated by MT,[1] obtained by crowdsourcing, covering a range of dialogue topics and annotated with fine-grained human judgments of MT quality. To our knowledge, this is the first corpus of its kind. Our data collection protocol is designed to encourage speakers of two languages to interact, using role-play scenarios to provide conversation material. Sentence-level human judgments of translation quality are provided by the participants themselves while they are actively engaged in dialogue. The result is a rich bilingual test corpus of 144 dialogues, which are annotated with sentence-level MT quality evaluations and human reference translations.

We begin by reviewing related work in corpus development, focusing particularly on informal written texts and spontaneous bilingual conversations (Section 1.1). We discuss the potential of the corpus and the collection method for MT research in Section 2, both for MT evaluation and for the study of language behaviour in informal dialogues. In Section 3 we describe the data collection protocol and interface. We describe basic characteristics of the corpus in Section 4. This includes a description of the annotation layers (normalised versions, human reference translations and human MT quality judgments) and examples. We illustrate the usefulness of the human evaluation by providing a comparison and analysis of the MT systems used (Section 4.4). We compare two different types of MT system, a baseline model and a mildly context-aware model, based on a quantitative analysis of the human judgements. We also provide examples of challenging phenomena for translation and include a preliminary analysis of a dialogue-level phenomenon,

[*] This work was started while the first author was a PhD student at LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay.

[1] Participants write and receive messages in their respective native language thanks to MT systems translating between the two languages.

namely consistent use of formal and informal person pronouns (Section 4.5). Finally, we provide plans for future work on the corpus in Section 5. The corpus, interface, scripts and participation guidelines are freely available under a CC BY-SA 3.0 licence.[2]

1.1 Related work

A number of parallel corpora of informal texts do exist. However they either cover different domains or are not designed with the same aim in mind. OpenSubtitles (Lison and Tiedemann, 2016; Lison et al, 2018) is a large-scale corpus of film subtitles, from a variety of domains, making for very heterogeneous content. However the conversations are scripted rather than being spontaneous and are translations of monolingual texts, rather than being bilingual conversations, and are subject to additional constraints such as sentence length due to the fact that they are subtitles. The MSLT corpus (Federmann and Lewis, 2016) is designed as a bilingual corpus, and is based on oral dialogues produced by bilingual speakers, who understand the other speaker's original utterances. This means that it is not possible to analyse the impact that using MT has on the interaction between participants. Other bilingual task-orientated corpora exist, for example BTEC (Basic Travel Expression Corpus; Takezawa et al 2002), SLDB (Spoken Language DataBase; Morimoto et al 1994) and the Field Experiment Data of Takezawa et al (2007), which is the most similar corpus to our own in that it contains MT-mediated dialogue. However these corpora are restricted to the travel/hospitality domains, therefore not allowing the same variety of conversation topic as our corpus. Human judgments for the overall quality are provided for the third corpus (Field Experiment Data), but only at a very coarse-grained level. Feedback about the participants' perception of MT quality is therefore of a limited nature in terms of MT evaluation, since sentence-level evaluations are not provided. Similarly, the multilingual dialogue corpora from the Verbmobil project (Wahlster, 2000) provide mediated interactions between speakers of two languages. However, the topic of conversation is also of a limited nature, centred around scheduling meetings. Furthermore, the corpora are not freely available.

2 Motivation

The main aim of our corpus is to serve as a test set to evaluate MT models in an informal setting in which communication is mediated by MT systems. However, the corpus can also be of interest for studying the type of language used in written dialogues, as well as the way in which human interaction is affected by use of MT as a mediation tool. We develop these two motivations here, starting with the corpus' utility for MT evaluation (Section 2.1) and then discussing the corpus' potential for the analysis of MT-assisted interaction (Section 2.2).

2.1 MT evaluation

The corpus is useful for MT evaluation in three ways: as (i) a test set for automatically evaluating new models, (ii) a challenge set for manual evaluation, and (iii) a validation of the effectiveness of the protocol to collect new dialogues and to compare new translation models in the future.

Test set for automatic evaluation The test set provides an example of spontaneously produced written utterances in an unscripted setting, with high quality human reference translations. It could be particularly useful for evaluating contextual MT models due to the dialogic nature of the utterances, the need to take into account previous MT outputs and the presence of metadata concerning both the dialogue scenario and the speakers involved. While it is true that the quality of the MT systems will influence the dialogue (in terms of translation errors), the dataset is a useful resource as an example of a scenario in which MT has been used for mediation. MT systems will continue to make errors and have not reached the level of human translators (particularly concerning aspects such as style, politeness and formality). It is therefore important to know how to handle errors when they arise, regardless of the system that has produced them, and to study how users of the systems change their language behaviour depending on the limitations of the systems.

Challenge set for manual evaluation It can be used as a challenge set for manual evaluation. The sentence-level human judgments provided can be used as an indicator as to which sentences were the most challenging for MT. Manual evaluation of new translations of our test set can then be guided towards those sentences whose translations are marked as *poor*, to provide an informed idea of the quality of the new

[2] `https://github.com/rbawden/DiaBLa-dataset` and `https://github.com/rbawden/diabla-chat-interface`

models on these difficult examples, and therefore to encourage development for particularly challenging phenomena.

Validation of the protocol for the collection of human judgments of MT quality Human evaluation remains the most accurate form of MT evaluation, especially for understanding which aspects of language pose difficulties for translation. While hand-crafted examples and challenge sets provide the means to test particular phenomena (King and Falkedal, 1990; Isabelle et al, 2017), it is also important to observe and evaluate the quality of translation on spontaneously produced texts. Our corpus provides this opportunity, as it contains spontaneous productions by human participants and is richly annotated for MT quality by its end users. In Section 4.4, we provide a preliminary comparative evaluation of the two MT systems, in order to show the utility of the human judgments collected. This same collection method can be applied to new MT models for a similar evaluation.

2.2 MT-assisted interaction

As MT systems are becoming more common online, it is important for them to take into account the type of language that may be used and the way in which user behaviour may affect the system's translation quality. Non-canonical syntactic structures, spelling and typing errors, text mimicking speech, including pauses and reformulations, must be taken into account if MT systems are to be used for successful communication in more informal environments. The language used in our corpus is relatively clean in terms of spelling. However participants are encouraged to be natural with their language, and therefore a fruitful direction would be in the analysis of the type of language used. Another interesting aspect of human-MT interaction would be to study how users themselves adapt to using such a tool during the dialogues. How do they deal with translation errors, particularly those that make the dialogue incoherent? Do they adjust their language over time, and how do they indicate when they have not understood correctly? An interesting line of research would be to use the corpus to study users' communication strategies, for example by studying breakdowns in communication as in (Higashinaka et al, 2016).

3 Data collection and protocol

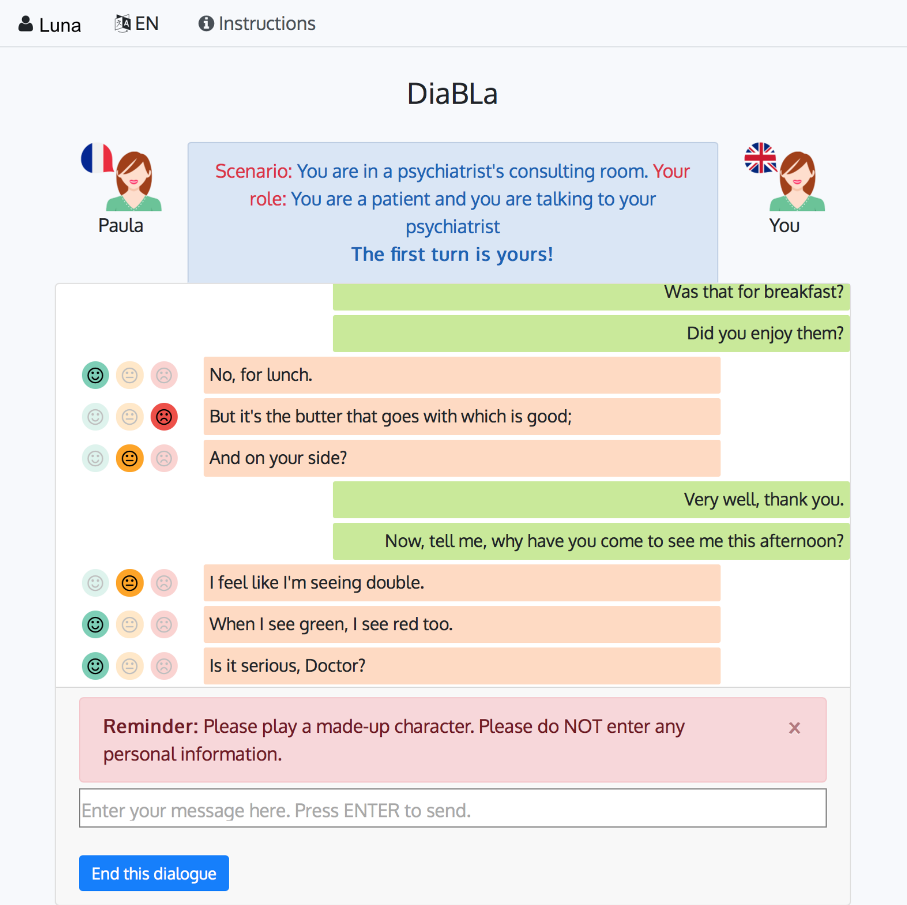
We collected the dialogues via a dedicated web interface (shown in Figure 1) allowing participants to register, log on and chat. Each dialogue involves two speakers, a native French speaker and a native English speaker. Each writes in their native language and the dialogue is mediated by two MT systems, one translating French utterances into English and the other translating English utterances into French.

Participants Participants were adult volunteers recruited by word of mouth and social media. They participated free of charge, motivated by the fun of taking part in fictional role-play. They provided basic information: age bracket, gender, English and French language ability, other languages spoken and whether they work in research or Natural Language Processing (NLP) (see Table 1 for basic statistics). Participants were evenly distributed between males and females and came from both NLP and non-NLP backgrounds. Although age ranges were the same for English and French speakers, the modal age brackets differed (55-64 for English and 25-34 for French).[3]

	EN	FR	All
Total number	37	38	75
#Researchers	7	17	24
#Experience in NLP	6	14	20
#Female/#Male	21/16	16/22	37/38
Min. age	18-24	18-24	18-24
Max. age	65-74	65-74	65-74
Median age	55-64	25-34	35-44
Modal age	55-64	25-34	25-34

Table 1: Some basic characteristics of the dialogue participants, recorded and distributed with the corpus.

[3] This was a direct consequence of the availability of volunteers. It is possible that the effect of age-specific language differences are attenuated by the fact that users played fictional characters. However, the effect of age on the language used could be studied further, since speaker age is provided with the corpus.



Luna EN Instructions

DiaBLa

Paula

Scenario: You are in a psychiatrist's consulting room. Your role: You are a patient and you are talking to your psychiatrist

The first turn is yours!

You

Was that for breakfast?

Did you enjoy them?

No, for lunch.

But it's the butter that goes with which is good;

And on your side?

Very well, thank you.

Now, tell me, why have you come to see me this afternoon?

I feel like I'm seeing double.

When I see green, I see red too.

Is it serious, Doctor?

Reminder: Please play a made-up character. Please do NOT enter any personal information. ×

Enter your message here. Press ENTER to send.

End this dialogue

Figure 1: The dialogue interface as viewed by the English participant (note that this example shows the middle of a dialogue rather than the beginning). The English participant's original utterances are shown on the right side in green. The French participant's utterances (machine translated into English) are on the left in peach colour. The quality of these machine translations has been evaluated by the English speaker using the smiley notation (Cf. Section 4.4). Note that the finer-grained evaluation is currently hidden in this example. See Figure 3 for the more detailed view of sentence-level evaluation.

Scenarios To provide inspiration and to encourage a wide variety of different utterances, a role-play scenario is given at the start of each dialogue and roles are randomly assigned to the speakers. We designed twelve scenarios, shown in Figure 2. The scenarios were chosen to reflect a range of everyday scenarios that people can relate to but that do not restrict the dialogue to be of a formulaic nature. The first turn is assigned randomly to one of the speakers to get the dialogue started. This information is indicated at the top of the dialogue screen in the participants' native languages. A minimum number of 15 sentences per speaker is recommended, and participants are informed once this threshold is reached, although they can continue for longer.[4] Participants are told to play fictional characters and not to use personal details. We nevertheless anonymise the corpus prior to distribution to remove usernames mentioned in the text.

Evaluation method The participants evaluate each other's translated sentences from a monolingual point of view. The choice to use the participants to provide the MT evaluation is an important part of our protocol: we can collect judgments on the fly, facilitating the evaluation process, and it importantly means that the evaluation is performed from the point of view of participants actively engaged in dialogue. Although some errors may go unnoticed (e.g. a word choice error that nevertheless makes sense in context), many errors can be detected this way through judgments about coherence and understanding of the dialogue flow. Having information about perceived mistakes could also be important for identifying those mistakes that are unperceived.

MT quality is evaluated twice, (i) during the dialogue and (ii) at the end of the dialogue. Evaluations are saved for later and not shown to the other participant.

[4] Participants can write several sentences in one utterance. However these are split into sentences and displayed as individual sentences. This makes the sentence-level evaluation more straightforward. It also potentially encourages participants to write sentence by sentence rather than to huge monologues, therefore encouraging more turn sharing in the dialogue.

1. You are both lost in a forest.
Roles: N/A

2. You are chefs preparing a meal.
Role 1: You are the head chef and you are talking to your subordinate.
Role 2: You are the subordinate chef and you are talking to the head chef.

3. You are in a classroom.
Role 1: You are the teacher and you are talking to a student.
Role 2: You are the student and you are talking to your teacher.

4. You are feeding the ducks by the pond.
Roles: N/A

5. You are both organising a party.
Role 1: It's your party.
Role 2: It's their party.

6. You are both stuck in a lift at work.
Role 1: You are an employee and you are with your boss.
Role 2: You are the boss and are with an employee.

7. You are in a retirement home.
Role 1: You are visiting and talking to an old friend.
Role 2: You are a resident and you are talking with an old friend who is visiting you.

8. You are in a bar.
Role 1: You are the bartender and talking to a customer.
Role 2: You are a customer and are talking to the bartender.

9. You are in an aeroplane.
Role 1: You are scared and are speaking to the person sitting next to you.
Role 2: You are speaking to the person next to you, who is scared.

10. You are at home in the evening.
Role 1: You are telling your spouse about the awful day you had.
Role 2: You are listening to your spouse telling you about the awful day they had.

11. You are in a psychiatrist's consulting room.
Role 1: You are the psychiatrist and are with your patient.
Role 2: You are a patient and you are talking to your psychiatrist.

12. You are on holiday by the pool.
Role 1: You are trying to relax and the other person wants to do something else.
Role 2: You want to do something else and the other person is trying to relax.

Figure 2: The twelve scenarios and speaker roles. Equivalent descriptions were presented in French to the French participants. Each scenario was presented six times for each MT model to ensure a balanced corpus.

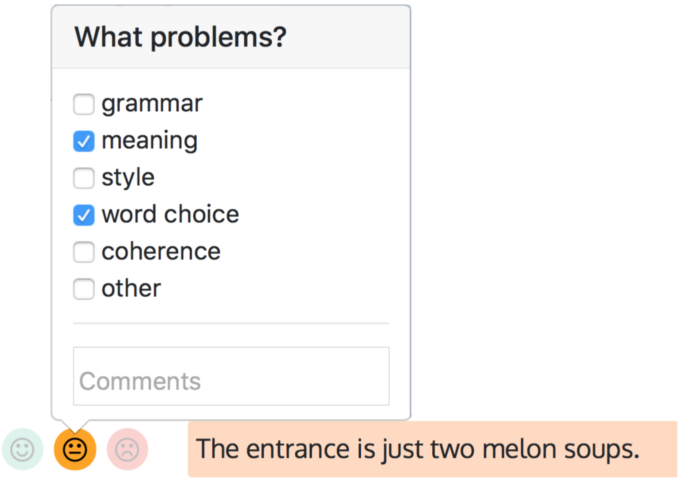
Participants evaluate each translated sentence during the dialogue by first selecting an overall translation quality (*perfect*, *medium* or *poor*) using a smiley notation system. If they select either *medium* or *poor*, they are prompted to indicate which types of errors they think occur in the translation: *grammar*, *meaning*, *style*, *word choice*, *coherence*[5] and *other*[6] (see Figure 3 for an example). Note that several problems can be indicated for the same sentence. If the participants wish, they can also write a free comment providing additional information or suggesting corrections. This annotation schema was designed to strike the right balance between providing fine-grained evaluations (finer grained than many existing datasets such as the Field Experiment Data of Takezawa et al (2007)), without making the task tedious or overly complex for the participants, which could also impact the naturalness of the dialogue. The participants can update their previous evaluations at any time during the dialogue any number of times. This may occur if they change their mind about their evaluation, for example if new utterances in the dialogue make it clear that their previous evaluation was not correct. The entire history of the evaluations (including the times of updates) is recorded so that changes in the perception of errors are documented.

Once the dialogue is finished, participants give overall feedback of the MT quality. They are asked to rank the quality of the translations in terms of *grammaticality*, *meaning*, *style*, *word choice* and *coherence* on a five-point scale (*excellent*, *good*, *average*, *poor* and *very poor*), and to indicate whether any particular aspects of the translation or of the interface were problematic. Finally, they indicate whether they would use such a system to communicate with a speaker of another language.

Before starting to interact, participants were given instructions (with examples) on how to evaluate MT

[5] Defined as a lack of consistency with previous utterances or the context.

[6] See Appendix B.5 for examples provided to participants of the error types.



What problems?

☐ grammar
☑ meaning
☐ style
☑ word choice
☐ coherence
☐ other

Comments

The entrance is just two melon soups.

Figure 3: The sentence-level evaluation form. The original French sentence was *L'entrée c'est juste deux soupes de melon.* "The starter is just two melon soups."

quality and these instructions remained available during the dialogues (cf. Appendix B). There is expected to be a certain degree of variation in participants evaluation. This subjectivity, inevitable with any human evaluation, is interesting, as it gives an indication of the variance of the tolerance for errors, and which types of errors were considered most detrimental.

MT systems We compare the quality of two MT model types (see Section 4.4). Within a dialogue, the same model type is used for both language directions, and each model is used an equal number of times for a given scenario, and therefore for the the same number of dialogues. Both models are neural encoder-decoder models with attention (Bahdanau et al, 2015), implemented using Nematus (Sennrich et al, 2017). The first model (BASELINE) is trained to translate sentences in isolation. The second (2TO2), is trained to translate sentences in the context of the previous sentence, as in (Tiedemann and Scherrer, 2017) and (Bawden et al, 2018). This is done by concatenating each sentence with its previous sentence, separated by a special token, and translating both sentences at once. In a post-processing step, only the current sentence is kept. Note that if the previous sentence is written by the same speaker as the current sentence, then the original previous sentence is prepended. If the previous sentence is written by the other speaker (in the opposite language), then the MT output of the previous sentence is prepended to the current sentence. This means that the previous sentence is always in the same language as the current sentence, and also corresponds to the context seen by the current speaker, as illustrated in Examples 1 and 2.[7]

(1) **Same speaker as previous sentence:**

$$\overbrace{\text{You would not believe the day I've had.}}^{\text{Original English sentence}}\ \text{<CONCAT>}\ \overbrace{\text{I never want to go back to work again!}}^{\text{Current English sentence}}$$

(2) **Different speaker from previous sentence:**

$$\overbrace{\text{We could go to town for a while}}^{\text{MT translation of French sentence}}\ \text{<CONCAT>}\ \overbrace{\text{I've seen the town and I'm not impressed}}^{\text{Current English sentence}}$$

Training Data and MT setup The systems are trained using the OpenSubtitles2016 parallel corpus (Lison and Tiedemann, 2016). The data is cleaned, tokenised and truecased using the Moses toolkit (Koehn et al, 2007) and tokens are split into subword units using BPE (Sennrich et al, 2016b). The data is then filtered to exclude poorly aligned or truncated sentences,[8] resulting in a training set of 24,140,225 sentences. Hyper-parameters are given in App. A. During the dialogues, the participants' text is first split into sentences and preprocessed in the same way as the training data. Translation is performed using MARIAN for fast CPU decoding (Junczys-Dowmunt et al, 2018).

[7] Note that these sentences would be preprocessed but are shown here in their raw form to aid clarity.

[8] We automatically align the subtitles on the word level using FASTALIGN Dyer et al (2013) using the *grow-diag-final-and* heuristic. We filter out those sentences for which fewer than 80% of either source or target tokens are aligned with one or more words of the sentence in the other language.

4 Corpus characteristics

Language direction	EN→FR	FR→EN	All
#Turns			
Total	1,067	1,089	2,156
Mean per dialogue	7.4	7.6	15.0
#Sentences			
Total	2,865	2,883	5,748
Mean per dialogue	19.9	20.0	39.9
Min. / max. per dialogue	5 / 42	5 / 60	10 / 102
Mean per turn	2.7	2.6	2.7
Min. / max. per turn	1 / 9	1 / 10	1 / 10
#Tokens (original messages)			
Total	27,817	29,058	56,875
Total unique	3,588	4,244	-
Mean per dialogue	193.2	201.8	395.0
Mean per sentence	9.7	10.1	9.9
#Tokens (MT versions)			
Total	28,745	27,510	56,255
Total unique	3,698	3,141	-
Mean per dialogue	199.6	191.0	390.7
Mean per sentence	10.0	9.5	9.8
#Tokens (reference translations)			
Total	30,093	27,014	57,107
Total unique	4,361	3,556	-
Mean per dialogue	209.0	187.6	396.6
Mean per sentence	10.5	9.4	9.9

Table 2: Characteristics of the resulting corpus in terms of dialogue length and sentence length for both original and translated utterances.

Table 2 shows the basic characteristics of the 144 dialogues. 75.7% of dialogues contain more than 35 sentences and the average sentence length is 9.9 tokens, slightly longer than the translations. An extract of dialogue, carried out in scenario 10 (cf. Figure 2), is given in Figure 4, providing an example of the type of language used by the participants. The language used is colloquial and contains a number of fixed expressions (e.g. *get off your intellectual high-horse*, *Mr Fancy pants*), which can prove difficult for MT, as is the case here. The systems are sometimes robust enough to handle spelling and grammatical errors (e.g. *qui ne penses* 'who think$_{2.sg}$' instead of *qui ne pense* 'who thinks$_{3.sg}$'' and *rality* instead of *reality*, translated into French as *ralité* instead of *réalité*, conserving the spelling error in translation, as a result of subword segmentation). The dialogues also contain cultural references, such as to films and actors. In many cases named entities are well conserved, although sometimes cause problems, for example *Marcel Carné* translated as *Marcel Carborn*. The explanation is that *Carné* is segmented into two subwords, *Car* and *né* 'born' during pre-processing, which are separately translated into English before being concatenated back together in a post-processing step.

4.1 Normalised versions

Although participants were encouraged to use their best spelling and grammar, such errors did occur (missing or repeated words, typographical errors, inconsistent use of punctuation). We provide manually normalised versions of sentences containing errors. The aim of this normalisation is to provide information about the presence of errors (useful for studying their impact on translation), and for providing a basis for the human reference translations, as we do not attempt to reproduce errors in the translations. Corrections are kept to a minimum (i.e non-canonical use of language was not corrected if linked to the colloquial use of language), and therefore in practice are limited to the addition of capital letters at the beginning of sentences and fullstops at the end of sentences and typographical error correction only when the correct form can easily be guessed from the context.[9]

[9] When a wrong but attested word is used, but it is made explicit that this is intentional, we do not correct the word. E.g. the use of 'deserts' instead of 'desserts' in 'The **deserts** are in the refrigerator. I said deserts for fun, I meant desserts!'

EN	Orig: Why do get the impression this is the first time you've mentioned anything going wrong between us? MT: Pourquoi pensez-vous que c'est la première fois que vous dites que quelque chose va mal entre nous ? Ref: Pourquoi est-ce que j'ai l'impression que c'est la première fois que tu mentionnes quelque chose qui ne va pas entre nous ? Eval: ☹ '…not able to make the difference between TU and VOUS…'
FR	Orig: Je suis comme Louis Jouvet dans les enfants du Paradis, quand il dit à Arletty qu'il a besoin de changer d'air. MT: I'm like Louis Jouvet in the children of Heaven, when he tells Arletty he needs a change of air. Ref: I'm like Louis Jouvet in Children of Paradise, when he says to Arletty that he needs a change of air. Eval: 🙂
FR	Orig: Tu te rappelles ce film magnifique de **Marcel Carné**, nous l'avons vu en français MT: You remember that beautiful **Marcel Carborn**, we saw it in French. Ref: Do you remember that superb film by Marcel Carné? We saw it in French. Eval: 😐 (MEDIUM) "'that"…could probably do without it'
EN	Orig: i have no idea what you're talking about. MT: Je ne sais pas de quoi tu parles. Ref: Je n'ai aucune idée de quoi tu parles. Eval: 🙂
EN	Orig: But, i'm going to go with you on this one and assume that it has just slipped my mind. MT: Mais je vais y aller avec toi sur ce coup-là et assumer que ça m'a échappé. Ref: Mais, je vais essayer de te suivre et supposer que ça m'est juste sorti de la tête. Eval: 🙂
EN	orig:so, yeah, sure, i remember your stupid film. MT: Donc, oui, bien sûr, je me souviens de ton stupide film. Ref: Donc, oui, bien sûr, je me rappelle de ton stupide film. Eval: 🙂
EN	Orig: it probably sucked. MT: C'était probablement naze. Ref: Il était probablement pourri. Eval: 🙂
	…
FR	Orig: J'ai besoin de stimulation intellectuelle, émotionnelle, pas de vivre à côté d'un geek attardé et boutonneux qui ne **penses** que à ses jeux vidéos et à ses mangas. MT: I need intellectual stimulation, emotional stimulation, not to live next to a retarded and pimply-faced geek who **thinks** only of his video games and its Manga. Ref: I need intellectual and emotional stimulation, not living next to a retarded, pimply geek who only thinks of his video games and mangas. Eval: 🙂
EN	Orig: well look at you, **Mr Fancy pants**! MT: Eh bien, regardez-moi ça, Mr. Ref: Eh bien, regarde-toi, M. le snobinard ! Eval: 😐 'the Mr isn't useful'
	…
EN	Orig: how about you get off your **intellectual high-horse** and get back to **rality** for a bit MT: Que diriez-vous de sortir de votre cheval intellectuel et de retourner à la **ralité** un moment ? Ref: Que penses-tu de redescendre un peu de ton piédestal intellectuel et de revenir un peu à la réalité. Eval: 😐 (MEANING) 'cheval intellectuel doesn't exists'

Figure 4: A dialogue extract from scenario number 10: 'You are at home in the evening…'. Shown are the original utterances ("Orig:"), the machine translated versions that were shown to the other participant ("MT:"), the reference translations produced *a posteriori* ("Ref:") and some details from the participant evaluation produced during the dialogue ("Eval:"). The MT outputs were produced by the baseline model in this dialogue.

4.2 Machine translations

Each sentence is translated automatically into the other language for the other participant. As previously mentioned, a single type of MT system (BASELINE or 2TO2) is used for all sentences within a dialogue. The use of two different systems is relevant to our analysis of the human evaluations produced by dialogue participants (Section 4.4). The choice of MT system does of course affect the quality of the MT. However, even as techniques advance, the corpus will remain relevant and useful as a test set and for analysing human language behaviour in this setup independently of this choice. The only limiting factor is having an MT model of sufficient quality for the participants to understand each other enough to communicate basic ideas and to effectively exchange on misunderstandings or provide correctly translated reformulations. Our MT models both largely surpass this criterion, as indicated by the positive feedback from participants.

4.3 Human reference translations

In order for the corpus to be used as a test set for future MT models, we also produce human reference translations for each language direction. Translators were native speakers of the target language, with very good to bilingual command of the source language, and all translations were further verified and corrected where necessary by a bilingual speaker.

Particular attention was paid to producing natural, spontaneous translations. The translators did not have access to the machine translated versions of the sentences they were translating to avoid any bias towards the MT models or the training data. However, they could see the machine translated sentences of the opposite language direction. This was important to ensure that utterances were manually translated in the context in which they were originally produced (as the speaker would have seen the dialogue) and to ensure cohesive translations (e.g. for discursive phenomena, such as anaphora and lexical repetition). Spelling mistakes and other typographical irregularities (e.g. missing punctuation and capital letters) were not transferred to the translations; the translations are therefore clean (as if no typographical errors had been present). This choice was made because reproducing the same error type when translating is not always possible and so could depend on an arbitrary decision.

4.3.1 Translation difficulties

The setup and the informal nature of the dialogues posed some unique challenges for translation, both for MT and also more fundamentally for translation in general. We mention a selection of these here to illustrate the complexity of deciding how to translate, where there is not a simple correct translation. We list four of these problems here, with real examples from the corpus.

Informal nature of text Unlike more formal texts such as those from news and parliamentary domains that are typically used in MT, the dialogues are spontaneous productions that contain many colloquialisms and examples of idiomatic speech for which translation equivalents can be difficult to find. We chose idiomatic equivalents based on communicative intention, rather than producing more literal reference translations, as for instance shown in Example 3.

(3) Well look at you, **Mr Fancy pants!**
*Eh bien, regarde-toi, **M. le snobinard** !*

Language-specific ambiguity An ambiguity in one language that does not hold in the other can sometimes lead to seemingly nonsensical utterances. This is a theoretical translation problem and not one that can be solved satisfactorily; either extra explanatory information is added concerning the structure of the other language, sacrificing the spontaneous and natural aspect of the dialogue, or the translation is translated without this extra information, which could leave the receiver of the message confused about the relevance of the utterance. In Example 4, the French speaker corrects his use of the male version of the word ‘patient’ to the female version (*patiente*). Given that this gender distinction is not one that is made in English, the choice is either to specify the gender of the patient in the English translation, which would be considered odd by the English speaker, or to translate both words as the same word *patient* as done here, which also leads to the translation of the second translation appearing incoherent. Although this is the only example of this kind in our corpus, this phenomenon could easily reoccur in other gender marked nouns. In Example 5, the word *ice-cream* is automatically translated into its French equivalent *glace*, which also has a second meaning of ‘mirror’. The French speaker picks up on this ambiguity concerning the word *glace* ‘ice cream or mirror’ and asks for clarification. However, the distinction that is made, between the *glace* which is eaten and the one that you look into is not one that will resonate with the English speaker, for whom no ambiguity exists. The effect of this is that the translation of the French speaker’s utterances appear nonsensical to the

English speaker. In our reference translations, we translated these utterances as best as possible, despite resulting incoherences in the target language, without adding additional information.

(4) FR: D'ailleurs il est l'heure de **mon patient**$_{\text{male}}$ suivant.
*Ref: Besides, it's time for my next **patient**.*
FR: Ou plutôt, de **ma patiente**$_{\text{female}}$ suivante, d'ailleurs.
*Ref: Or **patient** should I say.*

(5) EN: I can't stop thinking about **ice-cream**. . .
*MT: quoi que je fasse , je ne peux pas m'empêcher de penser à **une glace**. . .*
FR: Pensez vous en permanence à la **glace** qui se mange ?
*Ref: Do you always think about **ice cream** that's eaten?*
FR: ou bien à une **glace** pour se regarder ?
*Ref: Or about a **mirror** to look into?*

Self-correction Mistakes in the original utterance may be corrected by the speaker in subsequent utterances (e.g. poor choice of word, spelling mistake). In theory, we would want the translation of the erroneous utterance to reflect what was initially intended by the speaker, where it can be safely guessed (i.e. a model that is robust to noisy source sentences). However this means that a second utterance whose purpose is simply to correct the first would be superfluous if the translation of the first sentence does not contain an error. An example of a spelling error and correction is given in Example 6. As with the ambiguity cases, we choose to provide reference translations that correct the errors when they are evident rather than inject artificial mistakes into translations as the types of errors injected would be decided arbitrarily.

(6) EN: . . . to do **uor** jobs. . .
*. . . pour faire **notre** travail. . .*
EN: Typo: . . . to do **our** jobs. . . .
*Typo: . . . pour faire **notre** travail. . .*

Self-corrections were infrequent, with only 2 instances of explicit indication as in Example 6 (using the word *typo*), but it is a situation that is likely to arise in such written dialogue scenarios.

Meta-discussions Although not common, mistranslations by the MT models occasionally led to coherence problems in the dialogue that led to meta-discussions about what was originally intended. These meta-discussions may or may not contain elements of the mistranslation, which must then be translated back into the other language. In Example 7, the utterance *Ou à la limite thaï* was poorly translated by the MT system as *the Thai limit* rather than *at a push Thai*. The resulting mistranslation repeated by the English speaker in *What do you mean by **the Thai limit?***. Ideally, we would want to indicate to the original French speaker that there was a translation problem and therefore not translate into French using the original expression *à la limite thaï*. We therefore choose to translate the English sentence as it would most likely have been understood by the English speaker, resulting in a French translation that is different from the original term used.

(7) Tu connais un restau indonésien?
Do you know an Indonesian restaurant?
Ou **à la limite thaï** ?
*Or at a push Thai? (MT: Or **the Thai limit**)*
What do you mean by **the Thai limit?**
*Qu'est-ce que tu veux dire par **la limite thaïlandaise** ?*

Of the translation difficulties mentioned here, meta-discussions are the most frequent. For example, there are 9 instances of *I mean* and 26 instances of *do you mean* related to meta-discussions in the original English sentences.

4.4 Human judgments of MT quality

As described in Section 3, participants evaluated the translations from a monolingual (target language) perspective. We provide a preliminary analysis of these judgments to show that they can be used to distinguish the MT quality of different models and that such a protocol is useful for comparing MT models in a setting in which they are used to mediate human communication.

Overall MT quality Although an in-depth linguistic analysis is beyond the scope of this paper, we look here at overall trends in evaluation.[10] Figure 5 illustrates the overall differences between the translation of the two models compared according to the three sentence-level quality labels *perfect*, *medium* and *poor*.

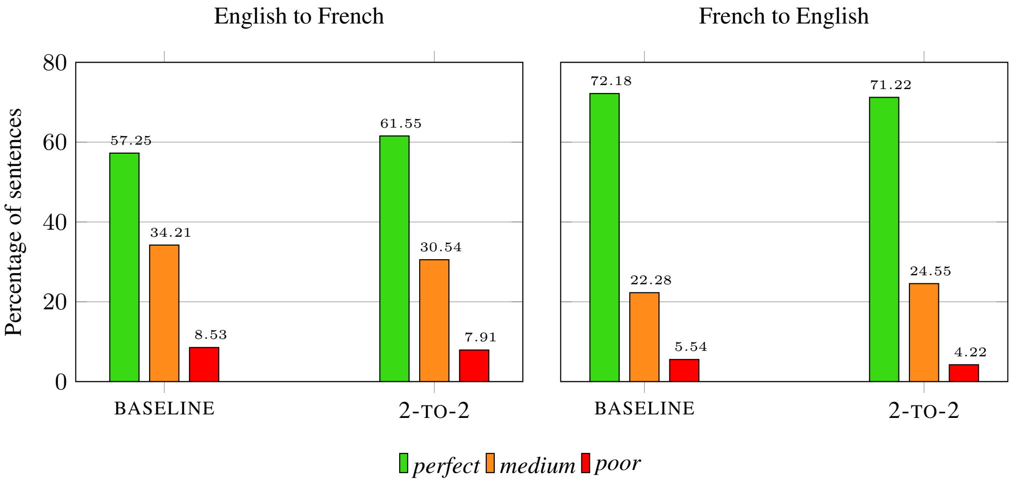


Figure 5: Percentage of sentences for each language direction and model type marked as *perfect/medium/poor* by participants.

The results show unsurprisingly that MT quality is dependent on the language pair; translation into English is perceived as better than into French, approximately half of all EN→FR sentences being annotated as *medium* or *poor*.[11] There is little difference in perceived quality between the BASELINE and 2TO2 models for FR→EN. This contrasts with EN→FR, for which the number of sentences marked as *perfect* is higher by +4 absolute percentage points for 2TO2 than for BASELINE. An automatic evaluation with BLEU[12] (Papineni et al, 2002) shows that the contextual model scores mildly better than the baseline, particularly for EN→FR. We retranslate all sentences with both models and compare the outputs to the reference translations: for FR→EN, the 2TO2 model scores 31.34 (compared to 31.15 for the BASELINE), and for EN→FR, 2TO2 scores 31.60 (compared to 30.99 for the BASELINE). These scores reflect the tendencies seen by the human evaluations: a smaller relative difference between the two models for FR→EN and a greater difference for EN→FR. For both language directions, 2TO2 results in a higher BLEU score, which is reflected in the smaller percentage of sentences perceived as *poor* compared to BASELINE. As for the quality difference remarked between the two language directions by the human participants, the BLEU scores cannot offer such insights, since they cannot be compared across languages and on different sets of sentences.

Types of errors encountered The comparison of the breakdown of problem types for each model and language direction is shown in Figure 6.

The few number of problems classed as *other* indicates that our categorisation of MT errors was sufficiently well chosen. The most salient errors for all language directions and models are related to *word choice*, especially when translating into French, with approximately 16% of sentences deemed to contain a word choice error. As with the overall evaluations, there are few differences between BASELINE and 2TO2 for FR→EN, but some differences can be seen for EN→FR: 2TO2 models perform better. There are fewer errors in most problem types, except *word choice*. However, the only error types that are statistically significant (according to a Fisher exact significance test (Fisher, 1922), based on the presence or not of each error type for each model type) are *style* ($p \leq 0.1$) and *coherence*($p \leq 0.01$). There is a notably difference in the the lower frequency of *coherence*-related errors for 2TO2. Coherence errors also appear to be less serious, as there is a lower percentage of translations labelled as *poor* as opposed to *medium*. These results are encouraging, as they show that our data collection method is a viable way to collect human judgments, and that such judgments can reveal fine-grained differences in MT systems, even when evaluating on different sentence sets.

[10] Given that the evaluation is performed by different participants on different subsets of sentences, comparisons should only be made when strong tendencies occur. The evaluations nevertheless provide a rich source of information about how MT mistakes are perceived by different native speakers.

[11] It should be noted that the *medium* category is often when any kind of problem is noticed in the sentence, however minor and whatever the nature of the error.

[12] Calculated using Moses' `multi-bleu-detok.perl`.

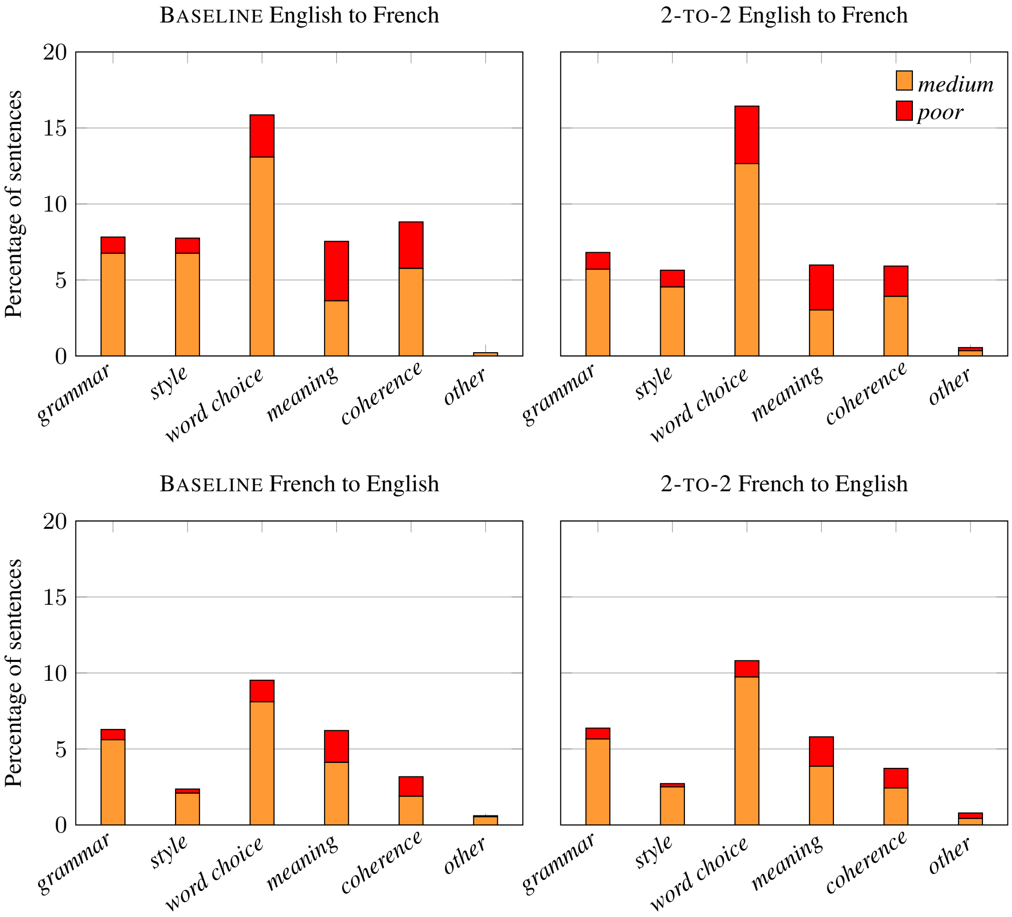


Figure 6: Percentage of all sentences for each language direction and model type marked as containing each problem type (a sentence can have several problems). Bars are cumulative and show the percentage for sentences marked as *medium* (orange) and *poor* (red).

Global participant feedback In spite of the errors, the translation quality is in general good, especially into English, and participant feedback is excellent concerning intelligibility and dialogue flow. As well as sentence-level judgments, participants indicated overall MT quality once the dialogue was complete. Participants indicated that they would use such a system to communicate with a speaker of another language 89% of the time. In 81% of dialogues, grammaticality was marked as either *good* or *excellent*. Coherence, style and meaning were all indicated as being *good* or *excellent* between 76% and 79% of the time. As a confirmation of the sentence-level evaluations, word choice was the most problematic error type, indicated in only 56% of dialogues as being *good* or *excellent* (40% of dialogues had *average* word choice, leaving a very small percentage in which it was perceived as *poor*). There were few differences seen between the two model types for these coarse-grained evaluations. One notable difference was seen for *style* for EN→FR, where 2TO2 scores better than the BASELINE. For BASELINE, style is marked as *average* for more dialogues than 2TO2 (38% vs. 18%) and as *good* for fewer dialogues (46% vs. 65%).

Examples of errors and the challenge faced for the MT of bilingual dialogues It is important to understand the potential problems that using MT for real-life conversations can present, particularly in terms of misunderstandings resulting in a negative social impact. Some of these problems can be flagged up in our corpus, and we will hope that these can be analysed further in the future.

Certain MT errors are easily identifiable, such as in Example 8, where an item has been repeated until the maximum sentence length is complete. However the most damaging mistakes are those that go undetected. Example 9 is an illustration of how a mistranslation can have quite serious consequences. In this example the translation has the opposite meaning of the intended utterance and the English speaker therefore understands something inappropriate from the French speaker's original utterance.

(8) EN: If I check out some cocktail recipes and I'll buy all the mixers, fruits, mint, lemon, lime etc.
MT: Si je prends des recettes de cocktail et que j'achète toutes les mixers, fruits, menthe, citron, . . .
FR: Alors, je trouve que ça fait beaucoup de citron, mais sinon pas de problème.
Ref: Well, I think that's quite a lot of lemon, but otherwise no problem.

(9) FR: De toute façon ton patron a toujours été **antipathique** !
MT: Anyway, your boss has always been ***fond of you.***
Ref: In any case your boss has always been ***unpleasant****!*
EN: Really? I mean he's twice my age. And that would be inappropriate. . .

4.5 Focus on a dialogue-level phenomenon

We study one specific discursive phenomenon: the consistent use of French pronouns *tu* and *vous*. The French translation of singular *you* is ambiguous between *tu* (informal) and *vous* (formal). Their inconsistent use was one of the most commented problems by French speakers, and a strategy for controlling this choice has been suggested for this reason (Sennrich et al, 2016a). Neither of our models explicitly handles this choice, although 2TO2 does take into account pairs of consecutive sentences, and therefore could be expected to have more consistent use of the pronouns across neighbouring sentences. As a proxy for their ability to account for lexical cohesion, we look at the two models' ability to ensure consistent translation of the pronouns across consecutive sentences. For each model, we take translated sentences in which *tu* or *vous* appear, and for which the previous sentence also contains either *tu* or *vous*. By comparing the number of times the current sentence contains the same pronoun as the previous sentence, we can estimate the degree of translation consistency for this particular aspect. The results are shown in Table 3 for both MT models and also for all reference translations. The reference translations show a very high level of consistency in the use of pronouns in consecutive sentences, showing that in most scenarios we expect the same pronoun to be used by both speakers. For the comparison of the two MT models, although the absolute figures are too low to provide statistical significance, we can see a general trend that the 2TO2 model does show greater consistency in the use of the pronouns over the baseline model, with +10% in the consistency use of *tu* and +6% in the consistent use of *vous*.

	BASELINE		2TO2		REFERENCE	
Prev. \ Curr.	*tu*	*vous*	*tu*	*vous*	*tu*	*vous*
tu	**52**	32	**56**	24	**124**	3
vous	28	**30**	23	**28**	6	**181**

Table 3: For each model, the number of times each model translates using *tu* and *vous* in the current sentence and either of the forms *tu* and *vous* also appears in the previous sentence.

5 Conclusion and future work

The corpus presented in this article is an original and useful resource for the evaluation of the MT of dialogues and dialogue in general. It provides a basis for many future research studies, in terms of the interaction between MT and humans: how good communication can be when using MT systems, how MT systems must adapt to real-life human behaviour and how humans handle communication errors. The corpus was designed to reflect three main characteristics:

1. MT-mediated dialogues
2. a wide range of non-scripted topics
3. fine-grained sentence-level human judgments of MT quality.

Although there exist corpora that satisfy some of these characteristics, our corpus is unique in displaying all three. As described previously, the pre-existing MT-mediated dialogue corpora are typically restricted to specific scenarios, often focused on a structured task. The relatively unrestricted nature of our corpus allows participants to be freer in their language. We also design the dialogues such that the participants do not have access to the original source sentences of the other participant, testing in a realistic setting how well MT models can be used for mediation.[13] This makes it possible to collect human judgments of MT quality that are based on the participants' ability to spot errors in the continuity of a dialogue situation. The pre-existing corpora do not contain such fine-grained judgments of quality. Certain corpora contain judgements (e.g. the Field Experiment Data of Takezawa et al (2007)), but only on a global level rather than sentence by sentence. The rich feedback on the translation quality of the MT models evaluated in our corpus makes it a useful tool for both evaluation and analysis of this setup.

We compared the use of two different model types in the collection of dialogues: a baseline NMT model that translated sentences independently of each other, and a lightly contextual model that translated sentences in the context of the previous sentence. Our analysis of translations produced by the models based on the sentence-level human quality judgments for each of the two MT models revealed interesting differences between the two model types. Whereas there was little difference seen between the models for FR→EN, where there were deemed to be fewer errors overall by both models, greater differences could be seen for EN→FR. There were notably 4% more sentences considered to be of perfect quality and there was a reduction in the number and severity of coherence errors, indicating that the contextual nature of the model could be improving translation coherence.

We intend to further extend the English-French corpus in future work and annotate it with discourse-level information, which will pave the way for future phenomenon-specific evaluation: how they are handled by different MT systems and evaluated by the participants. In this direction, we manually annotated anaphoric phenomena in 27 dialogues (anaphoric pronouns, event coreference, possessives, etc.). Despite the small size of this sample, it already displays interesting characteristics, which could provide a strong basis for future work. Anaphoric references are common in the sample annotated: 250 anaphoric pronouns, 34 possessive pronouns, and 117 instances of event coreference. Their incorrect translation was often a cause of communication problems (see Example 10 in which the French pronoun *il* is poorly translated as *he* rather than the inanimate version *it*), the impact of which will be investigated further.

(10) FR: Je peux m'allonger sur ce **canapé**?
*MT: Can I lie on this **couch**?*
*Ref: Can I lie down on the **sofa**?*
FR: Je ne veux pas déranger, **il** a l'air propre et neuf
*MT: I don't want to bother, **he** looks clean and new…*
*Ref: I don't want to be any bother. **It** looks clean and new…*

The protocol for data and MT judgment collection presented provides a useful framework for future evaluation of MT quality. Our preliminary analyses of sentence-level human judgments show that the evaluation procedure is viable, and we have observed interesting differences between the two types of MT model used in our experiments, providing complementary to automatic evaluation metrics. The same protocol could also be applied for new MT models and could also be extended to include more scenarios. It could also be extended to other language pairs, as part of a larger, international effort.

References

Bahdanau D, Cho K, Bengio Y (2015) Neural Machine Translation by Jointly Learning to Align and Translate. In: Proceedings of the 3rd International Conference on Learning Representations, ICLR'15

Bawden R (2018) Going beyond the sentence: Contextual Machine Translation of Dialogue. PhD thesis, Université Paris-Saclay

Bawden R, Sennrich R, Birch A, Haddow B (2018) Evaluating Discourse Phenomena in Neural Machine Translation. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies, New Orleans, Louisiana, USA, NAACL-HLT'18, pp 1304–1313

Dyer C, Chahuneau V, Smith NA (2013) A simple, fast, and effective reparameterization of IBM Model 2. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, NAACL-HLT'13, pp 644–648

Federmann C, Lewis W (2016) Microsoft Speech Language Translation (MSLT) Corpus: The IWSLT 2016 release for English, French and German. Microsoft Research

Fisher RA (1922) On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society pp 87–94

[13] Contrary to some efforts that ensure perfect communication between participants.

Higashinaka R, Funakoshi K, Kobayashi Y, Inaba M (2016) The Dialogue Breakdown Detection Challenge: Task Description, Datasets, and Evaluation Metrics. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, Portorož, Slovenia, LREC'16, pp 3146–3150

Isabelle P, Cherry C, Foster G (2017) A Challenge Set Approach to Evaluating Machine Translation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, EMNLP'17, pp 2476–2486

Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, Hoang H, Heafield K, Neckermann T, Seide F, Germann U, Aji AF, Bogoychev N, Martins AFT, Birch A (2018) Marian: Fast Neural Machine Translation in C++. arXiv:180400344 [cs] ArXiv: 1804.00344

King M, Falkedal K (1990) Using test suites in evaluation of machine translation systems. In: Proceedings of the 1990 Conference on Computational Linguistics, Helsinki, Finland, COLING'90, pp 211–216

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, ACL'07, pp 177–180

Lison P, Tiedemann J (2016) OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: Proceedings of the 10th Language Resources and Evaluation Conference, Portorož, Slovenia, LREC'16, pp 923–929

Lison P, Tiedemann J, Kouylekov M (2018) OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan

Morimoto T, Uratani N, Takezawa T, Furuse O, Sobashima Y, Iida H, Nakamura A, Sagisaka Y, Higuchi N, Yamazaki Y (1994) A Speech and Language Database for Speech Translation Research. In: Proceedings of the 3rd International Conference on Spoken Language Processing, Yokohama, Japan, ICSLP'94, pp 1791–1794

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, ACL'02, pp 311–318

Sennrich R, Haddow B, Birch A (2016a) Controlling Politeness in Neural Machine Translation via Side Constraints. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, USA, NAACL-HLT'16, pp 35–40

Sennrich R, Haddow B, Birch A (2016b) Neural Machine Translation of Rare Words with Subword Units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, ACL'16, pp 1715–1725

Sennrich R, Firat O, Cho K, Birch A, Haddow B, Hitschler J, Junczys-Dowmunt M, Läubli S, Valerio A, Barone M, Mokry J, Nădejde M (2017) Nematus: a Toolkit for Neural Machine Translation. In: Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, EACL'17, pp 65–68

Takezawa T, Sumita E, Sugaya F, Yamamoto H, Yamamoto S (2002) Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain, LREC'02, pp 147–152

Takezawa T, Kikui G, Mizushima M, Sumita E (2007) Multilingual Spoken Language Corpus Development for Communication Research. Computational Linguistics and Chinese Language Processing 12(3):303–324

Tiedemann J, Scherrer Y (2017) Neural Machine Translation with Extended Context. In: Proceedings of the 3rd Workshop on Discourse in Machine Translation, Copenhagen, Denmark, DISCOMT'17, pp 82–92

Wahlster W (2000) Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System. In: Verbmobil: Foundations of Speech-to-Speech Translation, Springer, Berlin, Heidelberg, pp 3–21

Appendices

A MT Hyper-parameters

Hyper-parameters for both MT models:

- 90,000 joint BPE operations, and threshold = 50
- Filtering out of parallel sentences in which fewer than 80% of tokens are aligned (after running FastAlign (Dyer et al, 2013))
- Embedding layer dimension = 512, hidden layer dimension = 1024, batch size = 80, tied decoder embeddings and layer normalisation, maximum sentence length = 50

B Participation guidelines

The following guidelines were presented to all participants, and were available during the dialogue if needed. A French translation was presented to French-speaking participants.

DiaBLa Instructions

You will be participating in an improvised written dialogue with another user. You will each write in your own native language (English or French). Don't worry - you do not need to be able to speak or understand the other language. Machine translation systems will translate all of the other person's sentences into your language. You will also evaluate the translations from a monolingual point of view (i.e. is the sentence grammatical? Does it make sense? Was the word choice ok? Is it stylistically appropriate? Is it coherent with respect to previous sentences?)

Please read all instructions carefully before continuing!

B.1 Signing up and logging in

You must first register (we require some basic information – see FAQ). Log in using the email address you registered with. Choose a username and the language you are going to speak in, which must be your mother tongue. You will be talking to real people. To increase your chances of finding someone to chat to, fill in this spreadsheet with your availabilities. Or try your luck and log in straight away!

B.2 Dialoguing

Once logged in, invite someone by clicking on their username or wait for someone to invite you. You can accept or refuse an invitation to dialogue. If the request is accepted, you will be taken to the dialogue screen. One of you is assigned the first turn, and after that, you are free to dialogue as you please. You will be presented with a setting (at the top of the chat box) in which the dialogue will take place. E.g. "You are in a forest" and your role. Now improvise a dialogue in the setting provided, as in improvised drama or role-play. I.e. play a made-up character and not yourself. The dialogues are to be like written drama transcriptions, rather than chat messages. We recommend writing at least 15 sentences each (you will receive a message when this happens). You can write more, but it is even more useful for us to have more dialogues rather than fewer very long ones.

☹ **Please do not use:**

- emoticons or SMS speech
- your partner's username, your own username or personal details

☺ **Please do use:**

- your best spelling, grammar and punctuation
- the correct gender of you and your partner (for instance when using pronouns)
- your imagination! You can refer to imaginary objects/people around you

B.3 Evaluation

You will evaluate the other person's translated utterances by selecting one of the smileys:

- green smiley face: "perfect"
- orange neutral face: "ok but not perfect"
- sad red face: "poor"

When you select a smiley, you will be prompted to indicate what is wrong with the translation: grammar, meaning, word choice, style, coherence, plus any extra comments to make your evaluation clearer. See FAQ for some examples.

B.4 Purpose

We will be using the dialogues to evaluate the machine translation systems and how easy communication was. The dialogues will be used to create a corpus of dialogues, which will be freely distributed for research purposes, and also used to analyse the machine translation models. Be natural, spontaneous and creative! However, please avoid making your sentences purposefully difficult in order to trick the machine translation system. Thank you!

B.5 FAQ

What if I don't understand what my partner says?
As in real life, speak about the problem with your partner. Say that you don't understand and try to continue the dialogue as best as possible.

When evaluating, what do the error types correspond to?

- Grammatical: the sentence is agrammatical. A few examples: (i) number disagreement: "The boy are there.", (ii) Missing articles: "I want dog.", (iii) Wrong use of tenses, (iv) Gender disagreement (for French), etc.
- Meaning: the sentence does not appear to make sense, e.g.: I was told by my avocado that a sentence was likely.
- Word choice: a poor word choice was made, e.g.: "I did you a chocolate cake" (instead of "I made you a chocolate cake."), "He took an attempt" (instead of "He made an attempt")
- Style: the level of formality is inconsistent or language usage is strange, although grammatically well-formed and understandable, e.g.: Strange/unnatural utterances, wrong level of formality: "What's up" in a job interview, etc.
- Coherence: Lack of consistency with previous utterances or the context: wrong pronoun used that refers to something previously mentioned, inconsistent use of "tu" and "vous" (for French), word choice is inconsistent with what was previously said (e.g. "I'm angry! – What do you mean by 'upset'?"), etc.

Why do you need personal information?
This enables us to evaluate whether certain aspects of the conversation (e.g. gender marking in French) are correctly translated or not. It allows us to analyse how machine translation systems react to the differences in language use, which depends for instance on the age of the user. The personal information that will be distributed in the resulting corpus is the following:

- Your age bracket (18-24, 25-34, 35-44, etc.)
- Your gender
- Your French and English ability
- The other languages you speak
- Whether or not you have studied/worked in Natural Language Processing or research

Why do you need to know speaker gender?

Speaker gender can be important in certain languages in terms of which words agree with the gender of the speaker (e.g. French). We therefore ask you to choose between male and female for practical reasons. If you do not identify with either gender, please choose the one by which you wish to be identified linguistically (i.e. would you prefer to be referred to as “he” or “she”?). The important thing is to be coherent when you dialogue in your use of gender.