Human-Informed Speakers and Interpreters Analysis in the WAW Corpus and an Automatic Method for Calculating Interpreters' Décalage

Irina Temnikova[1], Ahmed Abdelali[2], Souhila Djabri[3] and Samy Hedaya[4]
[1]Freelancer, Sofia, Bulgaria
[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
[3]University of Alicante, Spain
[4]Translation and Interpretation Institute, HBKU, Doha, Qatar
.
[1]irina.temnikova@gmail.com, [2]aabdelali@hbku.edu.qa,
[3]sd89@alu.ua.es, [4]SHedaya@hbku.edu.qa

Abstract

This article presents a multi-faceted analysis of a subset of interpreted conference speeches from the WAW corpus for the English-Arabic language pair. We analyze several speakers and interpreters variables via manual annotation and automatic methods. We propose a new automatic method for calculating interpreters' décalage (ear-voice span) based on Automatic Speech Recognition (ASR) and automatic alignment of named entities and content words between speaker and interpreter. The method is evaluated by two human annotators who have expertise in interpreting and Interpreting Studies and shows highly satisfactory results, accompanied with a high inter-annotator agreement. We provide insights about the relations of speakers' variables, interpreters' variables and décalage and discuss them from Interpreting Studies and interpreting practice point of view. We had interesting findings about interpreters behavior which need to be extended to a large number of conference sessions in our future research.

1 Introduction

A key characteristics which speech-to-speech machine translation systems strive to have is a good trade-off between accuracy of translation and low latency (Waibel and Fuegen, 2012; Bangalore et al., 2012). **Latency** is defined as the delay between the input speech and the delivered translation (Niehues et al., 2016) and roughly corresponds to interpreter's **décalage** in human interpreting.

While a number of engineering approaches are being proposed to reduce latency by in the same time maintaining good automatic speech translation quality (Waibel and Fuegen, 2012; Bangalore et al., 2012; Sridhar et al., 2013b; Schmid and Garside, 2005), few approaches are getting explicitly inspired by human interpreting, by learning from the strategies which interpreters employ in order to produce good quality translation (Niehues et al., 2016; He et al., 2015; Sridhar et al., 2013a).

In line with this area of research, starting with an initial objective to boost a speech machine translation system working with English/Arabic language pair (Dalvi et al., 2017) we conduct experiments on a subset of sessions from the WAW corpus (Abdelali et al., 2018) - a corpus of simultaneously interpreted conference speeches, to get informed about interpreters' behaviour and learn which strategies interpreters employ to maintain good output accuracy while in the same time not exceeding their delay from the speaker. Our task is complex, as we want to find a way in which human expertise in interpreting can boost the performance of speech machine translation systems.

With this article, we are enriching our previous research (Temnikova et al., 2017; Abdelali et al., 2018) and run an extensive multilateral analysis on a subset of WAW corpus interpreted sessions, before extending to a large number of sessions. The aim of this article is to test how much and what information we can extract by a combined manual (expert) and automatic analysis and also to propose a new automatic method for décalage calculation. We present the results of a manual evaluation run by two human experts on the points of reference generated by our décalage method.

Knowing that the strategies applied by interpreters and their décalage (including décalage as a sign of cognitive challenges and as a strategy) depend on source input characteristics, and that décalage can subsequently influence other interpreters' variables (Lee, 2002), we analyze: 1)

Temnikova, I, Orăsan, C., Corpas Pastor, G., and Mitkov, R. (eds.) (2019) *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, September 5 - 6, pages 105–115.
https://doi.org/10.26615/issn.2683-0078.2019_013

the source speech characteristics of several conference sessions (including the presence of noise and other interruptions), 2) several output variables of interpreters (such as décalage, average interpreters' output speed, number of hesitations, repetitions and false starts) and we interpret our findings using the rich knowledge of a practitioner interpreter with background in Interpreting Studies. We address all these issues with a combination of automatic methods and manual (expert) annotations of both speech recordings and speakers' and interpreters' transcripts. We link our new findings with the manually annotated interpreting strategies in the same subset of conference sessions by two human annotators (Abdelali et al., 2018; Temnikova et al., 2017), see Section 3.

The rest of the article is structured as follows: Section 2 presents some of the relevant related work; Section 3 introduces the data and the general methodology; Sections 4 and 5 present the analysis of source speeches (both manual annotation and automatic analysis of fluency indicators and external conditions tags); Sections 6 and 8 discuss the analysis of interpreter variables (décalage and fluency indicators) and present our automatic décalage calculation method; Section 7 shows an approximate analysis of speakers input rate and interpreters delivery rate (speaking speed). Section 9 provides the overall results discussion and Section 10 concludes the article.

2 Related Work

Interpreting corpora are used as a resource for research in both Interpreting Studies (IS) (Bendazzoli and Sandrelli, 2009; Russo et al., 2018; Defrancq, 2015) and in Machine Translation (MT) (Paulik and Waibel, 2009; Shimizu et al., 2013; Sridhar et al., 2013a). Due to the different aims and available tools, the methods used for research in these two fields are somewhat different. As we come from the MT research perspective (but get inspired by IS), the related works which are the closest to us are He et al. (2016) and Sridhar et al. (2013a). He et al. (2016) run a corpus analysis on a parallel corpus of translated and simultaneously interpreted text for the Japanese-English language pair. They use a machine learning classifier (differently from us) in order to classify interpreters' strategies in the text. The strategies that they examine are segmentation, passivization, generalization, and summarization (similar to us). Sridhar et al. (2013a) performs a corpus analysis of the EPIC corpus (Bendazzoli and Sandrelli, 2005) investigating interpreters strategies and behaviour for the English-Spanish language pair. They analyze features such as décalage, compression (somewhat corresponding to our *summarizing* and *omissions*), hesitations, some discourse features (e.g. analysis of the use of pronouns). Their paper makes an overview of the whole corpus for these features, without linking the features as potentially causing one another and without entering in details and analyzing specific sessions, as our paper does.

Calculation of Décalage Most of the Interpreting Studies approaches for calculating interpreters décalage involve manual input: there can be manual adding of tags while using software to display aligned segments and play speaker's and interpreter's recordings (Defrancq, 2015; Lee, 2002). Some researchers use the EXMARaLDA platform[1]. Although humans can usually make deeper choices than machines, manual methods take a lot of efforts. The automatic approaches include Sridhar et al. (2013a) and Ono et al. (2008).

The most important issue in calculating interpreters décalage is deciding on the measurement units and points of reference (Defrancq, 2015; Timarová et al., 2011). Measuring units can be words or seconds. The points of reference vary: e.g. end of a speaker's and start of interpreter's content word (Ono et al., 2008), words with literal translation (Oléron and Nanpon, 1965), every 5 seconds, beginning of segments where at least one interpreter omitted more than 15 words, beginning of sentence, units of meaning (Podhajská, 2008) and "segments correspondence based on content, instead only on simple lexical equivalence" (Barik, 1973). Our measurement units are seconds, and reference points are selected aligned words (see Section 6.1). Our method differs from Sridhar et al. (2013a) as we removed the stop words and used content words and Named Entities (for NEs we also differ from Ono et al., 2008). We also differ from both approaches as we run evaluation of our décalage method reference points alignment with two expert human annotators.

3 General Methodology

Data Selection - The WAW Corpus: For all experiments and analysis we used the recordings

[1] https://exmaralda.org

and transcripts of conference speeches and of interpreters from the WAW corpus for the source language English and target Arabic. **The WAW corpus** is a conference interpreting corpus collected from three conferences which took place in Qatar in 2013-2014: WISE 2013 (World Innovation Summit for Education), ARC'14 (Qatar Foundation's Annual Research and Development Conference), and WISH 2013 (World Innovation Summit for Health). Most speeches (133) have as source language English, target Arabic, with very few (7) having source language Arabic and target English. The WAW corpus was collected in order to train the QCRI's[2] speech-to-speech machine translation system. It is composed of the recordings of both the conference speeches and interpreters (collected from interpreters' booths), their transcripts (obtained from transcription agencies), and the translations of the transcripts into the opposite language. The transcripts were manually annotated with tags[3]. For more details see Abdelali et al. (2018). The WAW corpus currently contains information such as: recordings length in seconds, interpreters' gender, topics, length of transcripts in words, number of tags in each transcript (both for speakers and interpreters). The corpus does not contain the names, nor any personal information about speakers and interpreters, the number of speakers or interpreters per session, prosody annotation. It has not been Part-of-Speech (POS) tagged nor syntactically parsed. We do not also know any details about the way conference interpreting was organized, e.g. if interpreters were given the speeches to get prepared before interpreting took place. In total there were 12 interpreters, some of which interpreted more than one speech. See Figure 1 for number of speeches per interpreter and average session duration.

In Abdelali et al. (2018) and Temnikova et al. (2017) a subset of source and target transcripts were manually annotated for some interpreting strategies (as sequences of words): additions, omissions, self-corrections, and summarizing. The results were showing omissions as highest number of strategies (Korpal, 2012), followed by additions (see Figure 5).

Data Selection - Speeches Used in this Paper: The interpreted conference sessions analyzed in this paper are all for the English-Arabic language

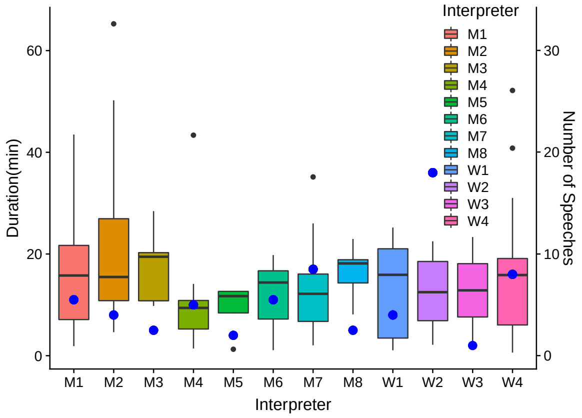

Figure 1: Average session length and number of speeches (●) per interpreter in the WAW Corpus.

direction. The majority of experiments (except for the speed comparison for the whole corpus) analyze **five** interpreted conference sessions, which were a subset of the sessions which we used in our previous research (Abdelali et al., 2018; Temnikova et al., 2017). Out of the 5 sessions two were from the same male interpreter (M7) and three from two female interpreters (W2 and W4). Male and female interpreters were selected in order to be able to analyze potential gender differences. Table 2 shows the duration in minutes of these sessions and the speakers and interpreters transcripts lengths in words. The selection criteria were the following:

1. M7, W2, and W4 were the interpreters, which had the highest numbers of sessions interpreted (see the blue dots in Figure 1).

2. There was a large difference in the number of annotated interpreting strategies in these transcripts (see Figure 5): in M7-T2 the interpreter employed the highest number of strategies, compared to all interpreter-transcript pairs, while in M7-T1 the interpreter employed the lowest number of strategies.

3. Similarly, W2-T2 had the lowest strategies employed by a female interpreter, while W4-T1 had the highest total number of strategies. W2-T1 was added to create a comparison between two very different sessions of the same interpreter as for M7.

The topics and conferences of the five selected recordings were: education conference WISE'14 (interpreter M7 and W4), topics - general edu-

[2] Qatar Computing Research Institute (QCRI).

[3] Transcription instructions with tags definition: http://tiny.cc/WAWTranscriptionTags.

cation (M7-T1), MOOCs (M7-T2), online education (W4-T1) and the general conference ARC'14 (W2, topics: W2-T2 - energy and environment, W2-T1 - traffic road accident).

Human Annotators: In Sections 4 (manual analysis of source speeches) and 6.1 (manual evaluation of décalage) we have relied on two annotators (A1 and A2), who both had research experience in Interpreting Studies. In addition, A1 completed studies in translation strategies and A2 has practitioner experience as a simultaneous interpreter and a degree in Interpreting Studies. Both annotators have advanced knowledge of English and native Arabic. We also consulted A3, who is a practitioner conference interpreter in Qatar with English and Arabic as source languages.

Methods Overview: The **source speech characteristics that we analyze are:** 1) environment conditions: noise, music, quality of sound 2) speakers variables: number of speakers, topics, speech intelligibility, (dis)fluency, accent, input rate, technicality of the topic. We have selected these variables in line with the IS state-of-the-art research, e.g. (Moser-Mercer, 1996; Pio, 2003; Plevoets and Defrancq, 2016; Fernández, 2015; Cecot, 2001). The **interpreters variables which we analyze are:** number of hesitations, false starts, repetitions, strategies used, delivery rate, décalage.

We use automatic methods for calculating the number of tags in the transcripts, to compute the speaking speed of speakers and interpreters, and for computing décalage. We use manual methods for evaluating the clarity and challenges in source recordings, for expert feedback on interpreters behaviour, and for manual evaluation of the décalage method. We compare all these new findings with our previous results of manual annotation of interpreting strategies (see Figure 5 from our previous article).

4 Analysis of Source Speeches - Manual Analysis

Method and Settings: The manual analysis of source speeches consisted in both annotators listening to the five recordings and entering values for several criteria and free text comments in an Excel spreadsheet form. The criteria (with available values) included sound quality (*very good, good, bad*), speech intelligibility (*clear, medium, difficult to understand*), (dis-)fluency (*fluent, not fluent*), number of topics, speakers' accent (*strong foreign accent, accent, no accent*), speakers' speed (*normal, fast, slow* - as perceived by the annotator), number of speakers, topic technicality of the source recording (*very technical, somewhat technical, very few technical words, not technical*).

Results: The manual analysis results are available online [4]. The cells in green show the points in which both annotators agreed. As we are aware that some of these criteria are not concretely defined, we run an objective automatic analysis (see Section 5). The feedback of A1 and A2 was that: M7-T1 and W2-T2 consisted in a conference presentation (with or without the session chair recorded), and W2-T1, W4-T1, and M7-T2 were panels; W2 were two women interpreters, who changed; in M7-T1 the speaker was reading and the interpreter was prepared; in W2-T2 the interpreter applied anticipation. As it can be seen from the online form, there is difference between the two annotators. What they mostly agree about is speech intelligibility, (dis)fluency, number of topics, number of speakers, topic technicality of the source speech, and a bit on speaker's speed. Specifically, M7-T1 had 1 speaker, M7-T2 was a panel with 8 speakers, W2-T1 had 6 or 7 speakers, W2-T2 had 2 speakers (one moderator), and W4-T1 had 6 speakers and was the only speech recording to have 3-4 topics.

5 Analysis of Source Speeches - Tags Analysis

Method and Settings: In order to complement the analysis in Section 4 with more objective numerical results, we counted the number of tags in the source recordings transcripts which were manually annotated by professional transcribers. In order to make the results comparable, we normalized the tags numbers per transcript length (divided per number of words) and then multiplied by 1000 to get a higher (but still comparable) numbers. Table 1 shows the tags and their definitions.

Our hypothesis is, as described by state-of-the-art research, that the presence of at least some of these tags may create challenges for interpreters (e.g. if the speakers make false starts [FALSE], hesitate [HES], repeat or correct themselves [REP] or if there is noise and music). *Unidentifiable* is an important tag, as if a word or phrase is not under-

[4] A1 and A2 manual speakers analysis is available at: http://tiny.cc/WAWManualSpeakerAnalysis

Tag	[FALSE]	[REP]	[INTER]	[HES]	[INTERJ]	[BREATH]
Meaning	False start	Repetition or Correction	Interruption	Hesitation	Interjection	Breathing
Tag	[LAUGH]	[APPLAUSE]	[MUSIC]	[NOISE]	[NE]	[UNK]
Meaning	Laugh	Applause	Music	Noise	Named Entity	Unidentifiable

Table 1: Tags annotated in WAW transcripts.

standable by transcribers it may also be such for interpreters. We also counted the Named Entities (NEs), as they correspond to names of people, locations and organizations and interpreters are usually supposed to render them correctly.

Results: We displayed only those tags whose value is above 0. Figure 2 shows the amount of tags per source recording which interpreters had to deal with (here we refer to recordings as "interpreter-transcript pair" for consistency). As it can be seen, the source recording with most tags was interpreted by W2 (W2-T1), the second one was M7-T2, while the source recording corresponding to W4-T1 had nearly no source speech tags at all.

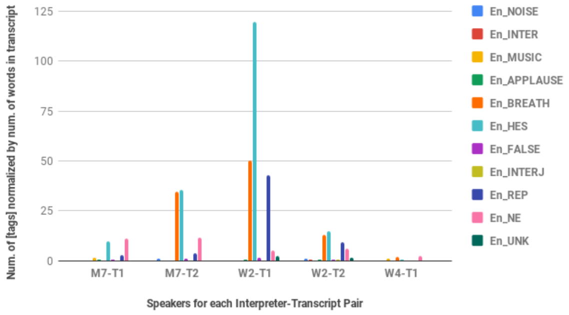

Figure 2: Number of transcription tags for the source speech of each session (normalized by number of words in the source transcript).

6 Analysis of Interpreters Décalage

In this section we propose a new automatic method for calculating the décalage of interpreter from speaker(s).

Gillies (2018) defines décalage as "*the time difference between what the speaker says and its reproduction by the interpreter in the target language*".

We want to be able to access interpreters' décalage in the WAW corpus for two reasons: 1) to determine when and how often in our data longer décalage is used as a strategy (Cecot, 2001; Moser-Mercer, 1997) and 2) to take it into account when analyzing the potential cognitive difficulties of interpreters. In fact longer décalage is generally to be avoided by interpreters as they should then keep more information in short-term memory and accuracy may significantly decrease (Lee, 2002). This is especially valid for interpreting between languages with highly different syntactic structures (Lee, 2002; Barik, 1975; Gile, 1997) such as English and Arabic (Bassam et al., 2014; Badr et al., 2009). Thus keeping décalage short can also be considered as an interpreting strategy.

Although décalage is an important feature and we wanted to implement it previously, we had a number of obstacles before being able to build this method. The biggest challenges were related to aligning source speech transcripts and interpreters transcripts. In fact interpreters transfer meaning and can completely restructure speaker's speech, make omissions, add words, and use completely different words than the standard translation equivalents.

Also, the alignment needed to be done at word level, which turned out to be very cumbersome and tedious to be performed manually; hence resorting to automatic alignment methods was a better option. This task had to include building or acquiring Automatic Speech Recognition (ASR) systems for both English and Arabic languages, to be able to automatically recognize words and mark them with their appropriate time-stamps.

6.1 Analysis of Interpreters: Décalage - Method and Evaluation

Transcripts alignment: The source speech and interpreters' transcripts were aligned by time and words-anchors were extracted using a bilingual dictionary. The anchors are Named Entities (NEs) and words that carry meaning (content words) - as opposed to frequent and functional words. We obtained the content words and NEs from the output of the part-of-speech (POS) taggers. To carry the alignment, we force-aligned the transcripts using our in-house ASR system (Khurana and Ali, 2016). The result of this process produced a transcript where each word is tagged with its offset time and duration.

POS tagging: Next, we used the part-of-speech tagger module of Farasa (Darwish et al., 2017) to POS tag the Arabic transcripts, and the Stanford POS tagger (Toutanova et al., 2003) for English. Additionally, we acquired a bilingual dictionary that was used for the alignment. The dictionary contains around 20k entries.

Computation of décalage: We compute décalage as the time between when the speaker pronounces a specific named entity (NE) or a content word and when the interpreter pronounces it (or its correspondent) using the onset reference. This time difference reflects the delay between when the interpreter hears a concept and when he is able to produce its correspondence in the target language.

Limitations of the Automatic Décalage Estimation Method: There might be instances in which the approach would not capture this lapse and the availability of these indicators could vary, based on the strategies that the interpreter choose to use. For example, the interpreter might choose to use a pronoun to replace a NE or a concept that was mentioned earlier (e.g. in cases of *summarizing* or *omission*). This will impact the number of anchors that are available for assessment and their alignment. Another inherent issue related to the source and target language pair is when the sentences are reordered differently between the source and target languages. We hypothesize that this would not be a major concern as this additional décalage could be shared across all transcripts/interpreters with the same language pair; but it might impact the comparison with other language pairs.

Décalage Method Evaluation: In order to test if our décalage calculation method is giving correct results, we run manual evaluation with our two annotators A1 and A2. Décalage was run on 16 interpreter-transcript pairs (two per interpreter, with two male interpreters - M7 and M1 and two female interpreters - W2 and W4), resulting in a total of 874 aligned décalage anchor word pairs. We selected semi-randomly from them 20 snippets of 10 consecutive lines (a total of 199). The snippets contained a representative variety of issues: named entities (person names, organizations, countries), content words (nouns, adjectives, verbs, adverbs), function words (like determiners and pronouns), several words which speakers repeated. The annotators had to label each aligned word pair by providing a label among: *Valid*, *Invalid*, *Somewhat valid* and *I don't know*. Annotators were informed to not look for correct word translations only (as interpreters transfer sense), but to also check if the two words are equivalent in terms of being a part of groups of words, in which the speaker and interpreter talk about the same. We then compared their results and run inter-annotator agreement comparison. The evaluation showed that A1 marked **193 (96.98%) pairs as Valid**, 0 as Invalid, 3 as Somewhat valid, and 3 as "I don't know". A2 labeled **185 (92.96%) pairs as Valid**, 14 as Invalid, 0 as Somewhat valid and 0 as "I don't know". In terms of inter-annotator agreement, **the annotators agreed on <u>182</u> out of 199 pairs (both labeled as *Valid*)**; 11 had the combination Invalid (A2)-Valid (A1); 3 were Invalid (A2)-Somewhat valid (A1) and 3 - Valid (A2)-"I don't know" (A1) [5].

6.2 Analysis of Interpreters: Décalage - Results

Figure 3 shows the anchor-based décalages for the two sessions of the male interpreter M7, while Figure 4 - for the female interpreters W2 and W4. The dots are the single décalages per anchored pair, the line is the average décalage over time, and the width of the grey shaded area indicates the variation.

It is clear from Figure 3 that the décalages in M7-T1 are mostly small – Median of 3.630 secs and Mean of 4.235 secs (in light green); while in M7-T2 (in light blue) the dots are much more spread around and there are many more instances in which the décalage (delay) is high and has a Median of 5.250 secs and a Mean of 5.838 secs.

Figure 4, shows one session of interpreter W4 (W4-T1) and the two sessions of interpreter W2 (W2-T1 and W2-T2). While W2's décalage in T2 looks consistent (constant) across the whole session with a Median of 3.880 secs and a Mean of 4.874 secs (light blue line), W4 starts with a lower décalage but there is a significant increase in the delay as time passes (pink line). Something similar with a much steeper increase in Figure 4 can be observed for W2 in T1 (W2-T1, light green line), for which the ending décalage is approximately 14 seconds vs 2 seconds in the beginning.

[5] We run Cohen's kappa, but received a surprisingly low IAA (0.132), despite an agreement of 93% between the annotators. This result turned out to be a Cohen's kappa known paradox (Yarnold, 2016).

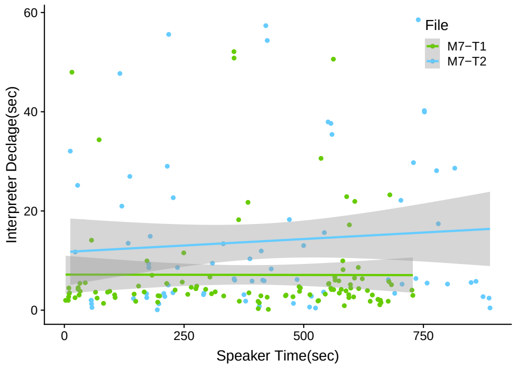

Figure 3: Comparison of interpreter's décalage between the two sessions of male interpreter (M7).

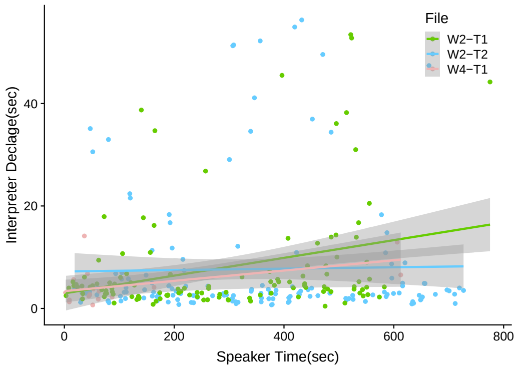

Figure 4: Comparison of interpreter's décalage between the sessions of female interpreters.

As shown above, while the process can be fully automated, challenges that are related to the domain and the availability of an ASR system that can provide the feeds are a major issue. Additionally, the accuracy of the lexicons is the weakest link of the proposed approach. The availability of this type of resource is strictly dependent on the language pair as well as on the domain. On the other hand, efforts by volunteers carrying the task of manually curating these resources and expanding them is a solution and a warranty for the approach.

7 Analysis of Speakers' and Interpreters' Speaking Speed

Method and Settings: As the manual annotation of speakers' speed in source recordings in Section 4 did not show much agreement between annotators (also because no objective definition was given), we wanted to complement our analysis with a more objective numerical approach. In this section we present an approximative calculation of average speaking speed per session of both speakers (input rate) in source language recordings and interpreters. We do that by dividing the number of words in each transcript by the length of recordings in minutes. We do this first for the five speeches under consideration, and then in order to validate our approach and get general observations - for all the WAW corpus En-Ar speeches.

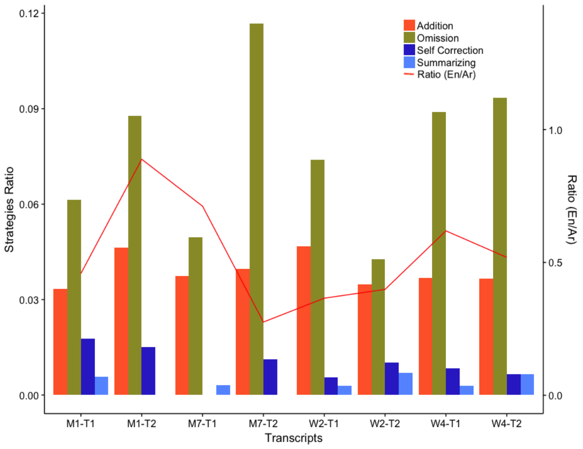

Figure 5: Annotated Strategies normalized by the transcript length in words for each session.

We realize that this is an approximative measure, as 1) speaking speed could vary during the session and 2) there are sessions with several speakers and/or interpreters. In future work we will use décalage's anchor points to calculate speaking speed in a more accurate way.

Results: Table 2 shows the results for the 5 sessions. The highest conference speakers' input rates (see column "En (words)") are in descending order for M7-T2, W4-T1, and W2-T2 (which were also indicated by A1 as *fast* speakers). The source speed of W2-T1 is nearly the same as for W2-T2, and M7-T1 is clearly the lowest speed. For matters of conformity with related work, we have converted the source input rate (speakers speed) into words/minute. According to (Pochhacker, 2015), an input rate of 100-120 words per minute is considered as "comfortable speech rate" (Pochhacker, 2015) and 150-180 words per minute is too high. Thus, the source input rates in M7-T2 and W4-T1 were exceptionally high, while in M7-T1 - near the ideal range. In terms of interpreters (see column "Ar (words)") M7-T2 has the lowest average speed and M7-T1 - the highest. This shows large variability of the same interpreter (M7). In addition to this, M7-T1 and M7-T2 exhibit the opposite correlation between speaker's and interpreter's speed: among the 5 speeches M7-T1 has the lowest speaker's speed and the highest interpreter's speed (also close to speaker's speed); M7-T2 has the highest speaker's speed and the lowest interpreter's speed. In terms of difference between speakers' speed and interpreter's speed M7-T2 has the highest value of 108.94 and the lowest difference value is 2.2 for M7-T1 (which means that in average the interpreter is moving almost at the same speed as speaker). It can be also seen that in M7-T1 speaker's (En) and interpreter's (Ar) number of words is nearly the same (differently from the other 4 recordings). According to A2's feedback in Section 4 in M7-T1 the speaker is reading (no spontaneous speech element) and the interpreter seems well prepared (according to both annotators the interpreter rendered correctly all statistical details), and thus most probably had the speech beforehand.

In order to have a wider picture of what our approximate speed calculation method generates, Figure 6 shows the approximate speaking speed results for all source and interpreters recordings in the WAW corpus for the interpreting direction En-Ar. Clearly there is a repeated general tendency across all speeches with the speed of interpreters being generally lower (around 1/2 from the speed of the source language speaker(s)).

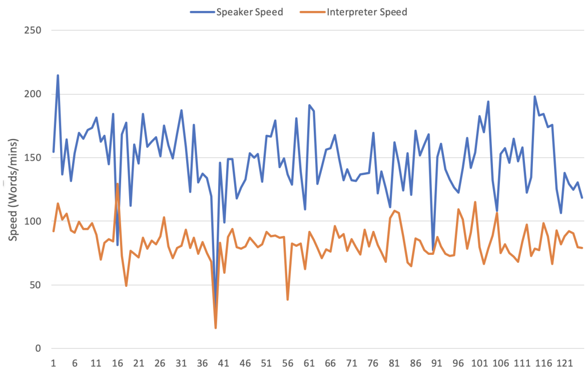

Figure 6: Speakers and Interpreters average speed for English into Arabic.

8 Analysis of Interpreters - Tags

Method and Settings: Similarly to speakers, we counted the number of tags in the interpreters (target language) recordings transcripts which were manually annotated by professional transcribers during transcription. We applied exactly the same method which we used for speakers (described in Section 5). We analyzed the same tags as in Table 1. We base our analysis on the assumption that hesitations, repetitions and false starts in interpreters' transcripts may show that the interpreter is challenged (Cecot, 2001). For example, it is known that hesitation pauses and other disfluencies of interpreters can be caused by difficulties in syntactic and lexical planning of discourse (Cecot, 2001). For matters of consistency we analyze all the available tags.

Results: Figure 7 shows the distribution of tags per interpreter-transcript pair. As in Figure 2, only existing tags are displayed. Clearly W2-T1 and W4-T1 have the highest number of tags. W2-T1 has an exceptionally high number of hesitations and W4-T1 has an exceptionally high number of breathing annotated. The lowest number is for M7-T2 which has only some [NOISE] tags.

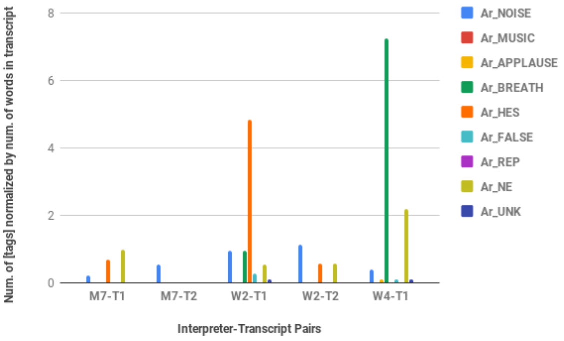

Figure 7: Number of transcription tags for the interpreter's output for each session (normalized by number of words in the interpreter's transcript).

9 Discussion

Analysis of interpreters: W2 interpreted the highest number of sessions (see Figure 1). Next are M7 and W4. The average session length for W4 is higher than of M7 and slightly higher than W2. In M7-T1 and M7-T2, speaker's input rate and interpreter's speaking speed confirm the large difference in strategies used by M7 (see Figure 5). Also, the highest input rates in M7-T2 and W4-T1 could explain the largest number of omissions in these two interpreters' sessions.

Analysis of speaker-interpreter combinations:

M7-T1 - 1 speaker (speaker reading and interpreter prepared), second shortest duration. Had a low number of annotated strategies (additions, omissions and summarizing), constant décalage from speaker of in average 3-4 seconds.

M7-T2 - panel. Had a relatively high number of

	Duration (sec)	En (words)	Ar (words)	En words/min	Ar words/min	Diff.
M7-T1	742.2	1341	1315	108.4	106.2	2.2
M7-T2	907.8	2656	1007	**175.54**	66.6	**108.94**
W2-T1	859.2	1959	1448	136.79	100.8	35.99
W2-T2	731.8	1678	1137	137.58	93	44.58
W4-T1	1043.5	2737	1423	157.37	81.6	75.77

Table 2: Speakers and interpreters speed (rounded) in the 5 analyzed speeches.

speakers' [BREATH] and [HES]. Interpreter had no tags, except for some [NOISE]. Had the highest number of annotated omissions (see Figure 5) and had also additions and self-corrections. This is the session with highest input rate and the interpreter with lowest speed. Interpreter must have skipped a lot (and used some generalizations according to A1) to maintain low speed. As we have seen in Figure 3 his décalage is higher and is increasing with the duration of the session. According to A2 the interpreter seems to be using silence and pauses to keep décalage lower.

W2-T1 - panel (2 interpreters), the speaker had a high number of hesitations [HES], breathing [BREATH] and repetitions [REP] (see Figure 2). The interpreter had a relatively high number of [HES]. So, there was a high number of hesitations in both speakers and interpreter (compare Figures 2 and 7). Had a much higher number of omissions and higher number of additions than W2-T2. We see a steep increase in décalage which ends with over 14 seconds.

W2-T2 1 speaker (2 interpreters), shortest recording duration. Interpreter applied anticipation. Very technical speech, speakers talked with lower voice. The lowest number of strategies (but all 4 are used). Had a nearly constant décalage (a bit increasing towards the end) of in average 3-4 seconds.

W4-T1 - panel, longest duration. The interpreter had a high number of [BREATH] and the highest number of NEs, which visibly does not correspond to the number of NEs in the speakers' transcript. Further analysis of the [NE] tag is necessary. Has a large number of omissions annotated. Décalage is also increasing, but not so steep as for W2-T1. Also here, the speakers' average input rate (according to our calculations) is high.

10 Conclusions and Future Research

Our aim was to test what amount and quality of insights we can gather from the WAW corpus with our new methods - a combination of automatic approaches and interpreters expertise. We presented an automatic décalage method which was tested on the English-Arabic language pair and showed to have high evaluation results from two expert human annotators.

We analyzed in detail five conference sessions (as they had interpreting strategies manually annotated) and provided general observations about multiple interpreters. We discovered that the dependence between speakers' variables (e.g. input rate and hesitations) and interpreters variables (e.g. décalage and strategies used) is very complex.

We found that: 1) manual expert analysis of an experienced researcher with interpreting and Interpreting Studies background enormously enriches automatic analysis findings; 2) the data existing in our corpus, accompanied by the new automatic décalage method provides rich insights.

Our analysis showed that among the issues that create challenges for interpreters and may generate increasing décalage and a higher amount of used strategies are: 1) large number of speakers; 2) spontaneous speech (as in question-answering sessions and panels vs prepared presentations or reading); 3) speakers' hesitations and repetitions; 4) high speakers input rate (see especially W4-T1 and M7-T2). We also found out that interpreters have much lower speaking speed than speakers' input rate, which adds to our previous and current observations that interpreters usually generate much fewer words.

As future work we need to run our experiments on a larger number of conference sessions to get general observations, to deepen our analysis of input rate and interpreters' delivery rate and test our methods on other corpora and language pairs.

11 Acknowledgements

We would like to thank the HIT-IT 2019 reviewers for their comments and Katsiaryna Panasuyk (A3) for her feedback.

References

Ahmed Abdelali, Irina Temnikova, Samy Hedaya, and Stephan Vogel. 2018. The WAW Corpus: The First Corpus of Interpreted Speeches and their Translations for English and Arabic. In LREC 2018. (ELRA), Miyazaki, Japan.

Ibrahim Badr, Rabih Zbib, and James Glass. 2009. Syntactic phrase reordering for english-to-arabic statistical machine translation. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pages 86–93.

Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez. 2012. Real-time incremental speech-to-speech translation of dialogs. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pages 437–445.

Henri C Barik. 1973. Simultaneous interpretation: Temporal and quantitative data. Language and speech 16(3):237–270.

Henri C Barik. 1975. Simultaneous interpretation: Qualitative and linguistic data. Language and speech 18(3):272–297.

Hammo Bassam, Moubaiddin Asma, Obeid Nadim, and Tuffaha Abeer. 2014. Formal description of arabic syntactic structure in the framework of the government and binding theory. Computación y Sistemas 18(3):611–625.

Claudio Bendazzoli and Annalisa Sandrelli. 2005. An approach to corpus-based interpreting studies: Developing EPIC (european parliament interpreting corpus). In MuTra 2005–Challenges of Multidimensional Translation: Conference Proceedings. pages 1–12.

Claudio Bendazzoli and Annalisa Sandrelli. 2009. Corpus-based interpreting studies: Early work and future prospects. Tradumàtica: traducció i tecnologies de la informació i la comunicació 1(7).

Michela Cecot. 2001. Pauses in simultaneous interpretation: A contrastive analysis of professional interpreters performances. The interpreters newsletter 11:63–85.

Fahim Dalvi, Yifan Zhang, Sameer Khurana, Nadir Durrani, Hassan Sajjad, Ahmed Abdelali, Hamdy Mubarak, Ahmed Ali, and Stephan Vogel. 2017. QCRI live speech translation system. EACL 2017 page 61.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, and Mohamed Eldesouki. 2017. Arabic pos tagging: Dont abandon feature engineering just yet. In Proceedings of the Third Arabic Natural Language Processing Workshop. pages 130–137.

Bart Defrancq. 2015. Corpus-based research into the presumed effects of short EVS. Interpreting 17(1):26–45.

Emilia Iglesias Fernández. 2015. Making sense of interpreting difficulty through corpus-based observation. Interpreting Quality: A Look Around and Ahead 19:35.

Daniel Gile. 1997. Conference interpreting as a cognitive management problem. Applied Psychology-London-Sage- 3:196–214.

Andrew Gillies. 2018. Conference Interpreting: A Students Practice Book. Routledge.

He He, Jordan L Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In HLT-NAACL. pages 971–976.

He He, Alvin Grissom II, John Morgan, Jordan Boyd-Graber, and Hal Daumé III. 2015. Syntax-based rewriting for simultaneous machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pages 55–64.

Sameer Khurana and Ahmed Ali. 2016. QCRI advanced transcription system (QATS) for the Arabic multi-dialect broadcast media recognition: MGB-2 challenge. In Spoken Language Technology Workshop (SLT), 2016 IEEE.

Paweł Korpal. 2012. Omission in simultaneous interpreting as a deliberate act, Intercultural Studies Group Universitat Rovira i Virgili, chapter 9, pages 103–111.

Tae-Hyung Lee. 2002. Ear voice span in english into korean simultaneous interpretation. Meta: Journal des traducteurs/Meta: Translators' Journal 47(4):596–606.

Barbara Moser-Mercer. 1996. Quality in interpreting: Some methodological issues. LINT.

Barbara Moser-Mercer. 1997. Beyond curiosity: Can interpreting research meet the challenge? In G.M. Fountain J.H. Danks, G.M. Shreve and M.K. McBeath, editors, Cognitive Processes in Translation and Interpreting. Sage Publications, London, pages 176–195.

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2016. Dynamic transcription for low-latency speech translation. In Interspeech. pages 2513–2517.

Pierre Oléron and Hubert Nanpon. 1965. Recherches sur la traduction simultanée. Journal de psychologie normale et pathologique .

Takahiro Ono, Hitomi Tohyama, and Shigeki Matsubara. 2008. Construction and analysis of word-level time-aligned simultaneous interpretation corpus. In LREC 2008.

Matthias Paulik and Alex Waibel. 2009. Automatic translation from parallel speech: Simultaneous interpretation as mt training data. In Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on. IEEE, pages 496–501.

Sonia Pio. 2003. The relation between st delivery rate and quality in simultaneous interpretation. The Interpreters Newsletter 12:69–100.

Koen Plevoets and Bart Defrancq. 2016. The effect of informational load on disfluencies in interpreting. Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association 11(2):202–224.

Franz Pochhacker. 2015. Routledge encyclopedia of interpreting studies. Routledge.

Květa Podhajská. 2008. Time lag in simultaneous interpretation from english into czech and its dependence on text type. Folia Translatologica 10:87–110.

Mariachiara Russo, Claudio Bendazzoli, Bart Defrancq, et al. 2018. Making way in corpus-based interpreting studies. Springer.

Philipp H Schmid and Adrian Garside. 2005. Method and apparatus for reducing latency in speech-based applications. US Patent 6,961,694.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Constructing a speech translation system using simultaneous interpretation data. In Proceedings of International Workshop on Spoken Language Translation (IWSLT).

Vivek Kumar Rangarajan Sridhar, John Chen, and Srinivas Bangalore. 2013a. Corpus analysis of simultaneous interpretation data for improving real time speech translation. In INTERSPEECH. pages 3468–3472.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013b. Segmentation strategies for streaming speech translation. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pages 230–238.

Irina Temnikova, Ahmed Abdelali, Samy Hedaya, Stephan Vogel, and Aishah Al Daher. 2017. Interpreting strategies annotation in the waw corpus. RANLP 2017 page 36.

Sárka Timarová, Barbara Dragsted, and Inge G Hansen. 2011. Time lag in translation and interpreting: A methodological exploration. Methods and strategies in process research: Integrative approaches in translation studies pages 121–146.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for computational Linguistics, pages 173–180.

Alexander Waibel and Christian Fuegen. 2012. Simultaneous translation of open domain lectures and speeches. US Patent 8,090,570.

Paul R Yarnold. 2016. Oda vs. π and κ: paradoxes of kappa. chance (PAC; 0= no inter-rater agreement, 100= perfect agreement) 2:7.