Interpreting Strategies Annotation in the WAW Corpus

Irina Temnikova¹, Ahmed Abdelali¹, Samy Hedaya²,
Stephan Vogel¹, and Aishah Al Daher²

¹Qatar Computing Research Institute, HBKU, Doha, Qatar

²Translation and Interpreting Institute, HBKU, Doha, Qatar

^{1,2}{itemnikova, aabdelali, shedaya, svogel, aldaher}@hbku.edu.qa

Abstract

With the aim to teach our automatic speech-to-text translation system human interpreting strategies, our first step is to identify which interpreting strategies are most often used in the language pair of our interest (English-Arabic). In this article we run an automatic analysis of a corpus of parallel speeches and their human interpretations, and provide the results of manually annotating the human interpreting strategies in a sample of the corpus. We give a glimpse of the corpus, whose value surpasses the fact that it contains a high number of scientific speeches with their interpretations from English into Arabic, as it also provides rich information about the interpreters. We also discuss the *difficulties*, which we encountered on our way, as well as our *solutions* to them: our methodology for manual re-segmentation and alignment of parallel segments, the choice of annotation tool, and the annotation procedure. Our annotation findings explain the previously extracted specific statistical features of the interpreted corpus (compared with a translation one) as well as the quality of interpretation provided by different interpreters.

1 Introduction

As manual translation is slow, often repetitive, and requires a lot of cognitive efforts and the use of additional resources (e.g. dictionaries, encyclopedias, etc.), part of it is now done automatically. Thanks to these recent advances in technology, translation is done in a much faster and sometimes more accurate way. One of the automatic translation tools, Machine Translation (MT) (Hutchins

and Somers, 1992) in its present state is used (with pre- and post-editing) in many companies and public institutions.

Despite recent improvements in MT (e.g. Neural MT), automatic MT systems still lack the precision and fluency of human translators and interpreters (Shimizu et al., 2013), and are often criticized because of that. Due to this, we want to teach our in-house speech-to-text (S2T) machine translation system (Dalvi et al., 2017) the techniques human interpreters use.

Human interpreters run several heavy-load processes in parallel (e.g. processing speaker's input, translating and pronouncing the previously heard input, monitoring their own speech, and correcting previous errors (Kroll and De Groot, 2009). To overcome time and brain processing limitations, and the inability to go back and correct their own output, they use many strategies (Kroll and De Groot, 2009; Al-Khanji et al., 2000; Liontou, 1996).

Before learning which human interpreting strategies can improve our S2T system, we run a corpus analysis.

We use a corpus of transcripts of conference speeches, their simultaneous interpretations (performed by professional interpreters), and their manual translations. We first extract surface features (Section 4). Next, we manually annotate coarse-grained interpreting strategies (Section 6) in a sample of the corpus.

The rest of this article is structured as follows: Section 2 summarizes the related work; Section 3 presents our WAW corpus, Section 4 presents some corpus statistics; Section 5 describes the corpus segmentation and alignment methods; Section 6 provides the annotation procedure. Section 7 presents the annotation results, and Section 8 is the Conclusion.

2 Related Work

Before being able to identify which Interpreters' Strategies (IS) could benefit our speech-to-text translation system, we first studied the **existing work on Interpreting Strategies in Interpreting Studies**.

There is substantial research in this area (especially corpus-based), e.g. (Roderick, 2002; Shlesinger, 1998; Bartłomiejczyk, 2006; Lontou, 2012; Hu, 2016; Wang, 2012; Bendazzoli and Sandrelli, 2009; Lederer, 1978; Al-Khanji et al., 2000; Lontou, 1996; Tohyama and Matsubara, 2006).

The research outlines a number of strategies interpreters use in order to alleviate the working memory overload and time shortage. Although different researchers divide and classify them differently, the strategies can be roughly classified (Kalina, 1998; Al-Khanji et al., 2000) into:

1. **Comprehension strategies** (e.g. *preparation, gathering of topic information, terminology check-up, anticipation, chunking* a.k.a. *segmentation* or *salami-technique*),
2. **Target-text production strategies** (e.g. source-text conditioned strategies, such as *transcoding*; target-text conditioned strategies, such as *ear-voice span manipulation, expansion*, and compression or simplification techniques – such as: *passivization* and *omission*; *self-correction, decision for no-self-correction*),
3. **Other strategies** (e.g. buying time by pronouncing generic utterances or delaying the response, *self-monitoring*),
4. **Compensatory strategies** (e.g. *approximation, filtering, omissions, substitutions*).

Some researchers investigate language-pair-specific strategies, e.g. Tohyama and Matsubara (2006); Lontou (1996).

MT researchers' interest on applying interpreters skills to MT was driven by the advances in automatic Speech Translation (ST). Languages with different syntax and word order are a problem for real time and simultaneous ST. Paulik and Waibel (2009) exploited the availability of parallel recordings to leverage on the scarcity of parallel text between English and Spanish. In this way they achieved better than expected performance in MT. Their findings were a motivation to exploit the data produced by interpreters in order to

further improve MT. Shimizu et al. (2013) used information learned from simultaneous interpretation data to improve their MT system between Japanese and English, as these two languages have very different grammatical structure and word order. Results showed that emulating the simultaneous interpreters style helped both to improve the accuracy of the system while minimizing the delay before translation generation. Sridhar et al. (2013) made a corpus analysis for simultaneity, hesitations, compression, the lag between source language and target language words, and the use of deixis. He et al. (2016) also made corpus analysis to discover which strategies interpreters use and analysed segmentation, passivization, generalisation, and summarization. Finally, Hauenschild and Heizmann (1997) are a collection of papers from the time of VerbMobil, which contains MT papers inspired by translation and interpreting, as well as translation and interpreting papers, which contribute to MT.

3 The WAW Corpus

The corpus we use in our experiment is a corpus of recordings of speeches/lectures from conferences held in Doha, Qatar. The **WAW** corpus contains 521 recorded sessions (127h 10min 38sec) collected during talks at WISE 2013 (World Innovation Summit for Education)¹, ARC'14 (Qatar Foundation's Annual Research Conference, a general conference on many topics)², and WISH 2014 (World Innovation Summit for Health)³ research conferences in Doha, Qatar. Both speeches in English as a source language (subsequently translated into Modern Standard Arabic), and in Arabic as a source language (subsequently translated into English) are present. From the ethical point of view, all conference speakers signed a release form to transfer the ownership to the conferences organisers (Qatar Foundation). The names of interpreters are not published.

The speeches contained in the corpus have been:

1. **Interpreted by professional interpreters** hired by the conference organisers.
2. **Transcribed by professional transcribers,**

¹<http://www.wise-qatar.org/2013-summit-reinventing-education-life>.

²<http://marhaba.qa/qatar-foundations-annual-research-conference-arc14-calls-on-local-research-expertise/>.

³<http://www.wish-qatar.org>.

according to our guidelines.

3. Translated by professional translators.

The transcripts of both the original speakers and their interpretations have been translated by professional translators, according to our guidelines, into Arabic (if the original language was English) or into English (if the original language was Arabic).

95% of the source/original speeches were in English.

Figure 1 shows the resulting corpus composition. For each speech, we have two audios and four corresponding texts (Original transcript, Interpretation transcript, translation of the source language original transcript and the translation for the target language interpretation transcript).

Out of the 521 recorded sessions, 266 sessions (63h 48min 05sec) contain the complete pack of translations of both source speech transcripts and interpreted target language transcripts.

In total, there were 12 interpreters. According to information from the conference organisers, most of the interpreters were experienced, but we do not have any details about their level of proficiency or areas of expertise.

Figure 1: WAW Corpus Composition.

Out of this data, we have mainly used the original speakers' and interpreters' transcripts.

4 WAW Corpus' Assessment

Table 1 shows WAW’s size in number of files, time recorded, number of lines/segments, and number of words.

The different number of *Segments* between the Arabic transcripts and the English ones was due to the fact that the initial segmentations of the original speakers' audios and interpreters' audios, as received by the companies were done independently by different transcribers.

Figure 2 shows the different segmentation of the same original speech transcript and its interpretation (both audios with a length of around 50 sec).

00:00:00.000 → 00:00:04.941 great importance [BREATH] We had a 2008 end-of-life care strategy	00:00:00.111 → 00:00:06.125 عليها 2008) لدينا استراتيجية رعاية [تدري] إلى مرحلة الاحتضار في العام 2008 نريد أن نحرص على أن تتطرق إلى تلك الحالات.
00:00:04.941 → 00:00:07.024 which we now are [BREATH] in the process of refreshing,	00:00:06.258 → 00:00:30.511 [من] أن تحدث الدكتور [عمر شمس علي] عن كلغة مرحلة الاحتضار [تتفرغ] هناك وتد في البيت أو الموت في القطاع معين وفي بعض معينة.
00:00:07.072 → 00:00:12.087 [BREATH] [HES] and we are needing to ensure [BREATH] that we attack each of these areas.	00:00:30.510 → 00:00:40.411 [من] في أغلب الأحيان نحن نتخذ القرارات الخطأ وما تلحق القرارات أو الفوارت [تتخذ] [تتخذ] [إعادة: 17] لا نتمتع على المال.
00:00:12.087 → 00:00:16.699 [HES] [NEP:ER Dr. Ghali] referred to [BREATH] other issues about the cost of end-of-life care	00:00:40.411 → 00:00:52.313 [تتخذ] [تتخذ] [إعادة: 17] لا نتمتع على المال.
00:00:16.699 → 00:00:22.499 [BREATH] [HES] and that is also really critical in the choice [BREATH] as to the home [HES] death	00:00:52.313 → 00:00:52.313 [تتخذ] [تتخذ] [إعادة: 17] لا نتمتع على المال.
00:00:22.500 → 00:00:28.842 [BREATH] [HES] in an acute sector [BREATH] [HES] in an acute setting [HES] with expensive intervention occurring.	00:00:52.313 → 00:00:52.313 [تتخذ] [تتخذ] [إعادة: 17] لا نتمتع على المال.
00:00:28.893 → 00:00:32.239 [BREATH] too often we make [FALSE the wrong] the inappropriate choice.	00:00:52.313 → 00:00:52.313 [تتخذ] [تتخذ] [إعادة: 17] لا نتمتع على المال.
00:00:32.299 → 00:00:36.504 to make the right choice is not one to be dictated by money;	00:00:52.313 → 00:00:52.313 [تتخذ] [تتخذ] [إعادة: 17] لا نتمتع على المال.
00:00:36.504 → 00:00:43.591 [BREATH] it is to make a choice that reflects the real wishes as potentially expressed by the patient.	00:00:52.313 → 00:00:52.313 [تتخذ] [تتخذ] [إعادة: 17] لا نتمتع على المال.
00:00:43.591 → 00:00:44.502 thank you.	00:00:52.313 → 00:00:52.313 [تتخذ] [تتخذ] [إعادة: 17] لا نتمتع على المال.
00:00:44.502 → 00:00:47.393 [NOISE] thank you, thank you very much.	00:00:52.313 → 00:00:52.313 [تتخذ] [تتخذ] [إعادة: 17] لا نتمتع على المال.

Figure 2: Original Speaker’s vs Interpreter’s Transcripts Segmentation Differences.

The segmentation difference was one of the difficulties we encountered on our way to manually annotating the transcripts. Our solution is explained in Section 5.

The different number of Words between English and Arabic, which can be observed in Table 1 is another very interesting point. As Arabic is an agglutinative language, it has fewer words. Thus, it has been observed that the average ratio between English and Arabic words in the original texts and their translations is around 1.5 (Salameh et al., 2011). We have computed this ratio both for the transcripts (original vs interpreted, as visible in Figure 1 horizontally, i.e. A1 vs B1), and for the manual translations vs the transcripts (vertically, A1 vs B2 and B1 vs A2) to test if this ratio is confirmed in our cases. We have found that vertically, in A1 vs B2 and B1 vs A2, the average ratios are around 1.5, which confirms the previously observed, and that there is a very small divergence. However, Table 1 shows that horizontally, the ratio

Language	N. of files	Total Time	N. of Segments	N. of Words	N. of Words Translation
Arabic	133	31:54:33	9,555	159,657	198,588
English	133	31:54:33	26,824	289,109	224,296
Totals.	266	63:49:05	36,379	448,766	422,884

Table 1: WAW Corpus’ Size.

between English and Arabic words in interpreters’ vs original speakers’ transcripts is higher, around **1.8** ($289,109/159,657 = 1.81$). Our hypothesis is that interpreters add more words than translators do.

Figure 3 shows the horizontal (A1 vs B1) word ratios for each interpreter. Comparing to the results for manual translations, where there is less variety and all ratios tend to be close to 1.5, we see larger differences for some interpreters (longer colored rectangle). This shows that the same interpreter added a different number of words in Arabic vs. English in the different speeches he/she interpreted. E.g. interpreters I07 and I09 have the largest variety, while I01 and I08 have the smallest variety.

This word ratios higher variety further motivated our wish to have a more detailed look into the behaviour of single interpreters via manual annotation (see Sections 6 and 7).

Figure 3: Word Ratios between Original Speakers’ transcripts and Interpreters’ Transcripts, per Interpreter.

5 Document Segmentation and Alignment

As said earlier, **the original segmentation difference was one of the difficulties on our way**, as we wanted to align parallel segments so the annotator

annotates both the original and the interpretation segments in parallel, and later we learn automatically the segments correspondence.

Table 2 shows examples of files from the same session/lectures. E.g., File 2 has 286 English segments, and only 94 Arabic segments. The same happens with all the files.

For our annotation experiment, we worked on a corpus sample, composed by the parts of 4 files with a total of 7500 words, interpreted by 2 interpreters (I09 and I10). I10 did the most interpretations across the three conferences. This made us hypothesize that he/she was the most expert. I09 had the average words ratio closest to written translation (1.5) and had a high words ratio variety. We hypothesized that this interpreter could be a beginner.

First, we attempted to automatically re-align the texts using automatic alignment tools (Varga et al., 2006; Ma, 2010; Braune and Fraser, 2010). The results were unsatisfactory as interpreters change both the word order and lexical choice by paraphrasing or summarizing the original speaker, so alignment tools could not find enough parallel segments.

Manual re-segmentation and alignment was done by one expert, native speaker of Arabic with advanced level of English, who was a specialist in video subtitles segmentation. The process took in total 7h 49min 16sec. Besides the obvious learning curve for this highly specific task, the average speed was 0.17 English w/sec (words/second) and 0.10 Arabic w/sec.

Another difficulty was finding an appropriate tool for both re-segmentation, alignment, and annotation. We split this task into 1. Re-segmentation and alignment and 2. Annotation. For (1), our expert used Excel spreadsheet and then our in-house web-based editing tool which better visualizes the parallel segments and outputs an unique merged file. The procedure was to 1. Segment the original speaker’s transcript, 2. Align and segment the interpreter’s transcript, according to (1).

File	Interpreter	time (sec)	En Segments	En Words	Ar Segments	Ar Words	Ratio
1	I10	231	43	559	19	331	1.69
2	I09	1296	286	3315	94	1963	1.69
3	I10	2050	444	5557	163	2386	2.33
4	I09	1101	274	3147	89	1728	1.82

Table 2: Expert Evaluation Data.

Initially, we followed video subtitles segmentation rules, but this resulted in too short segments, which created problems for aligning, as often the interpreters were changing the whole structure and the order of clauses and phrases.

Next, we have set as main rule to have an aligned segment in Arabic, while keeping the length of the original English sentence as short as possible.

The manually re-segmented and aligned version of Figure 2 is shown in Figure 4. The empty lines are left when there is no correspondence in the other language.

The final WAW re-segmentation and alignment guidelines are available online⁴.

great importance we had a 2008 end-of-life care strategy	لدينا 2008 لدينا استراتيجية رعاية في مرحلة الاختصار في العام 2008
which we now are in the process of refreshing	
and we are needing to ensure that we attack each of those three areas.	ونحن نريد أن نحرص على أن نتطرق إلى كل المجالات.
Dr. Ghali referred to other issues about the cost of end-of-life care	لقد تحدث الدكتور غالي عن كلفة مرحلة الاختصار
and that is also really critical in the choice as to the home death	و هناك الموت في البيت
or the death in an acute setting with expensive intervention occurring.	أو الموت في قطاع معين في وضعية معينة.
too often we make the inappropriate choice.	في أغلب الأحيان نحن نتخذ الخيارات الخاطئة
to make the right choice is not one to be dictated by money;	ربما نتخذ القرارات أو الخيارات الجيدة لا تعتمد على المال.
it is to make a choice that reflects the real wishes as quotably expressed by the patient.	بل الخيارات هي التي تعود إلى رغبة المريض فعلاً والتي عبر عنها
thank you.	شكراً
thank you, thank you very much.	شكراً شكراً جزيلاً

Figure 4: Manually Re-Segmented and Re-Aligned Transcript.

6 Annotation

Our **annotator** was a professional translator with expertise in annotating translation strategies, native speaker of Arabic, and fluent in English. The annotator passed a training on annotating around 1500 words. Training annotation was done using Word. Four strategies have been explored: *Summarizing*, *Omission affecting the meaning*, *Omission not affecting the meaning* and *Correction*.

The **annotation categories** for main annotation were selected: 1) out of the interpreting strategies listed in the Section 2, 2) filtered during an-

notator’s training, 3) coarse-grained. We have also asked the expert to evaluate whether some of these strategies were **needed** (“*tolerant*”) or **unnecessary** (“*intolerant*”). Our final annotation categories are: *Summarizing*, *Omissions (tolerant, out-of-delay, and intolerant)*, *intolerant Additions*, and *Self-correction*.

Finding an annotation tool was also one of our difficulties, as we needed to find a way for both aligned segments to be annotated in parallel. After asking in corpora mailing list⁵, we found our own solution. After producing a merged file (Section 5) with both parallel segments one after the other, we used GATE⁶. Our annotation guidelines are available online⁷.

Four files with a total of 4941 words in English and respectively 2767 words in Arabic were annotated. The source language in all files was English and the target – Arabic. Here are our annotation categories with their (expert’s) definitions and an example for each category.

Summarizing (Table 3): The interpreter combines two clauses into one clause capturing the main idea and conforming to the structure of Arabic. A single longer clause may also be summarized by the interpreter.

Original Speaker’s Transcript	Interpreter’s Transcript
So for instance we now from my group have spin out, they do mental health assessment	في مجموعتي مثلاً قمنا بتقييم للصحة النفسية Translation: In my group, for instance, we assessed the psychological health.

Table 3: Summarizing Strategy Example

Self-correction (Table 4): The interpreter usually uses “أو” (or) or repetition to alter a lexical choice or correct a mispronounced word.

⁵<https://mailman.uib.no/public/corpora/2017-May/026526.html>.

⁶<https://gate.ac.uk/>.

⁷Link: <http://goo.gl/hjyhAz>.

⁴Link: <http://goo.gl/hjyhAz>.

Original Speaker's Transcript	Interpreter's Transcript
We know where to focus on our clinical interventions.	علينا أن نكون مركزين أكثر في مداخل أو في تدخلاتنا السريرية. <i>Translation: We have to be more concentrated on interventions or on our clinical interventions.</i>

Table 4: Self-correction Strategy Example

Omission-tolerant (Table 5): This strategy is used when the information introduced by the speaker seems to not have essential effect on the entire meaning of the context. May also result from the speakers frequent repetitions of the same idea or clause.

Original Speaker's Transcript	Interpreter's Transcript
Thank you very much.	

Table 5: Omission-tolerant Strategy Example

Omission-intolerant (Table 6): These omissions affect the overall meaning of the context. They stem from interpreter's delay, miscomprehension, lack of anticipation, or/and the speaker's speed.

Original Speaker's Transcript	Interpreter's Transcript
So I deal with individuals who have traditional, you can say traditional cultural values	وأنا أتعاطى بالمسائل الثقافية الإسلامية <i>Translation: And I deal with Islamic cultural questions/issues.</i>

Table 6: Omission-intolerant Strategy Example

Omission-out of delay-intolerant (Table 7): This omission usually results from a long period of delay. The interpreter loses information because he/she may be unable to comprehend what is being said or because of the speaker's speed.

Addition-unnecessary (Table 8): The interpreter adds information that seems out of context (usually happens out of delay). However, some interpreters use this strategy to provide more explanations to the audience.

7 Annotation Results

182 instances out of 1047 segments were annotated (around 17%). Out of these, 135

Original Speaker's Transcript	Interpreter's Transcript
And they examined it, and they came out	
They were extremely concerned about the Dutch system	
And also the system in Oregon and some of the states in the United States.	

Table 7: Omission-out-of-delay-intolerant Strategy Example

Original Speaker's Transcript	Interpreter's Transcript
I did some work for our Royal College of Physicians on professionalism	أنا أتحدث دائماً عن المهنية والاحتراف <i>Translation: I am always taking about being professional and professionalism</i>
And we thought very deep and hard about what is professionalism.	ونحن نفكر بعمق عما هي المهنية، <i>Translation: And we think deeply about what is professionalism</i>

Table 8: Addition unnecessary Strategy Example

were “Omission”; 21 were “Addition”; 16 “Self-correction” and 10 “Summarizing”. Figure 5 shows the tags distribution per file. The annotation provided some insights about our initial observations. In the cases when the num. of words ratio was high (File 3 in Table 2), the annotations showed a high amount of “Omissions” in the Arabic interpretation vs. the original English speech. “Summarizing” contributes to this too, as shown in Figure 5. Omissions are the major cause of information loss “Addition” and “Self-correction” could balance this loss, but they are a too low number to compensate. File 1 was an exception. The length of the file (231 sec only, vs. 1000-2000 sec the other files, each) could potentially be the reason why we did not observe much.

8 Conclusions & Future Work

The WAW corpus is a collection of parallel lectures translated and interpreted from English into Arabic (mostly) and vice-versa. In the process of exploiting this resource for teaching a S2T automatic translation system, we investigated the characteristics of professional interpretation. An expert translator annotated the professional interpreters' strategies in a sample of the corpus by following our guidelines to segment, align and an-

Figure 5: Distribution of Annotation Tags per Transcript.

notate the transcripts. The findings from this pilot experiment confirmed and explained the previously observed anomalies. The discovered strategies will be tested within our in-house S2T translation system. The WAW corpus can be used by student interpreters to learn real, quality interpretation, by researchers in related fields, as well as to improve MT. We aim to expand these tasks further, as well as to automatize some of the previous steps. For future work, we are in the process of involving more annotators, better defining the annotation guidelines, and processing more texts.

References

- Raja Al-Khanji, Said El-Shiyab, and Riyadh Hussein. 2000. On the use of compensatory strategies in simultaneous interpretation. *Meta: Journal des traducteurs/Meta: Translators' Journal* 45(3):548–557.
- Magdalena Bartłomiejczyk. 2006. Strategies of simultaneous interpreting and directionality. *Interpreting* 8(2):149–174.
- 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* (7).
- Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pages 81–89.
- 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.
- Christa Hauenschild and Susanne Heizmann. 1997. *Machine translation and translation theory*, volume 1. Walter de Gruyter.
- 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.
- Kaibao Hu. 2016. Corpus-based interpreting studies. In *Introducing Corpus-based Translation Studies*, Springer, pages 193–221.
- William John Hutchins and Harold L Somers. 1992. *An introduction to machine translation*, volume 362. Academic Press, London.
- Sylvia Kalina. 1998. *Strategische Prozesse beim Dolmetschen: Theoretische Grundlagen, empirische Fallstudien, didaktische Konsequenzen*, volume 18. G. Narr.
- Judith F Kroll and Annette MB De Groot. 2009. *Handbook of bilingualism: Psycholinguistic approaches*. Oxford University Press.
- Marianne Lederer. 1978. Simultaneous interpretation-units of meaning and other features. In *Language interpretation and communication*, Springer, pages 323–332.
- Konstantina Liantou. 1996. Strategies in German-to-Greek simultaneous interpreting: A corpus-based approach. *Gamma: Journal of Theory & Criticism* 19:37–56.
- Konstantina Liantou. 2012. *Anticipation in German to Greek simultaneous interpreting*. Ph.D. thesis, Uni-wien.
- Xiaoyi Ma. 2010. Champollion: A robust parallel text sentence aligner. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation*. pages 489–492.
- 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.
- Jones Roderick. 2002. *Conference Interpreting Explained (Translation Practices Explained)*. St. Jerome Publishing.
- Mohammad Salameh, Rached Zantout, and Nashat Mansour. 2011. Improving the accuracy of english-arabic statistical sentence alignment. In *Int. Arab J. Inf. Technol.*, volume 8, pages 171–177.
- 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)*.
- Miriam Shlesinger. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta: journal des traducteurs/Meta: Translators' Journal* 43(4):486–493.
- Vivek Kumar Rangarajan Sridhar, John Chen, and Srinivas Bangalore. 2013. Corpus analysis of simultaneous interpretation data for improving real time

- speech translation. In *INTERSPEECH*. pages 3468–3472.
- Hitomi Tohyama and Shigeki Matsubara. 2006. Collection of simultaneous interpreting patterns by using bilingual spoken monologue corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.
- Dániel Varga, László Németh, P. Halcsy, András Kornai, Viktor Trón, and Viktor Nagy. 2006. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*. pages 590–596.
- Binhua Wang. 2012. Interpreting strategies in real-life interpreting .