The WAW Corpus: The First Corpus of Interpreted Speeches and their Translations for English and Arabic

Ahmed Abdelali, Irina Temnikova, Samy Hedaya, Stephan Vogel
Qatar Computing Research Institute & Translation and Interpreting Institute
Hamad Bin Khalifa University, Doha, Qatar
{aabdelali, itemnikova, shedaya, svogel}@hbku.edu.qa

Abstract

This article presents the WAW Corpus, an interpreting corpus for English/Arabic, which can be used for teaching interpreters, studying the characteristics of interpreters' work, as well as to train machine translation systems. The corpus contains recordings of lectures and speeches from international conferences, their interpretations, the transcripts of the original speeches and of their interpretations, as well as human translations of both kinds of transcripts into the opposite language of the language pair. The article presents the corpus curation, statistics, assessment, as well as a case study of the corpus use.

Keywords: corpus, interpreting strategies, annotation, translation, Arabic

1. Introduction

The Arabic Language Technologies research group at the Qatar Computing Research Institute[1] is developing Natural Language Processing (NLP) tools to support Arabic, including an Automatic Speech Recognition (ASR) system (Khurana and Ali, 2016), and a Speech-to-Text (S2T) Machine Translation (MT) system (Dalvi et al., 2017). Despite the advances in automatic MT, such systems still lag behind interpreters in accuracy and fluency. One of the planned applications of our S2T translation system is to support the educational domain and provide translations of lectures. As simultaneous interpretation is different from translation, we wanted to better understand the strategies interpreters apply to hopefully implement some of these strategies to improve our system. A first step for this was to build a corpus of interpreted lectures, which can be used to analyze how interpreters deal with the challenges in simultaneous interpretation. Although a number of Arabic corpora exist (Zaghouani, 2017; Wray and Ali, 2015; Ali et al., 2016a; Ali et al., 2016b; Ali et al., 2017), to the extent of our knowledge there are no publicly available interpreting corpora for Arabic. We therefore collected a corpus of lectures by recording talks given at conferences, together with their interpretations done by professional interpreters. After giving a short survey of the related work, we present the WAW Corpus and details about the collection and curation process. Next, we provide a quantitative and qualitative assessment of the collected data. We also present a pilot/case study of the use of the corpus for extracting interpretation strategies used by interpreters.

2. Related Work

In (Al-Khanji et al., 2000) interpreting strategies in Arabic have been studied, but no re-usable corpus was released. There are also Arabic speech corpora used for MT[2] (Kumar et al., 2014; Zaidan and Callison-Burch, 2014), but they do not include human interpretation of the original speech (only translated speech transcripts are provided). We are not aware of any other publicly available interpreting corpora for Arabic, whereas they exist for Italian, Spanish, English, French, Dutch (Bendazzoli and Sandrelli, 2005; Falbo, 2012), Brazilian Portuguese and German (House et al., 2012), Japanese and Chinese (Tohyama and Matsubara, 2006; Hu and Qing, 2009). Differing from the existing Arabic speech corpora, the WAW corpus contains recordings of the original speakers, the recordings of the interpreters, the transcripts of both recordings, and the translations of all transcripts.

3. WAW Corpus

The WAW corpus comprises recordings from three international conferences, which took place in Qatar: **W**ISE 2013 (World Innovation Summit for Education)[3], **A**RC'14 (Qatar Foundation's Annual Research and Development Conference[4], and **W**ISH 2013 (World Innovation Summit for Health)[5]. The speeches and discussions were mostly in English, some in Arabic.

Professional interpreters were hired by the event organizers to provide simultaneous interpretation from English into Modern Standard Arabic or vice-versa. To the best of our knowledge the interpreters were all native speakers of Arabic, and most talks were interpreted from English into Arabic, so into the native language of the interpreter. As all interpreters were high-level professionals, their level of English was at least at an advanced level. All speakers and interpreters signed a release form to transfer the ownership of their talks to the conferences organizers (Qatar Foundation) and to give permission for their speech to be recorded and used for scientific purposes.

Table 1 shows the corpus' topics composition with number of files per conference, classified per broad areas (for ARC'14, which is a wide interdisciplinary conference) and topics (for WISE 2013 and WISH 2013, as they are already specialized in the broad areas of Education and Health).

[1] http://www.qcri.org
[2] https://catalog.ldc.upenn.edu/LDC97T19.
[3] http://www.wise-qatar.org
[4] http://www.qf-arc.org
[5] http://www.wish-qatar.org

Conference	Number of files
ARC'14	71
Topics: Energy and Environment(31), Computing and Information Technology(21), Social Sciences, Arts and Humanities(12), Health(7)	
WISE 2013	33
Topics: General(16), Massive Open Online Courses(5), World-Wide Education(4), Web Literacy(4), Education Without Teachers(2), Primary School Teachers(2)	
WISH 2013	22
Topics: Big Data(4), Health and Ethics: End of Life(4), Road Traffic Injury(3), Patient Engagement(3), Innovation of Health Care(3), General(3), Mental health(2)	

Table 1: WAW Corpus number of files per conferences and topics.

3.1. Original Recordings

The original audio files where recorded with two audio channels captured from the audio stream – Arabic and English – using Zoom H4n Handy Recorders hooked into the audio systems of the conference room. The recordings were stored using the WAV format with 24-bit encoding and a sampling rate of 48khz. The stereo recordings were then split per language and per session. Each session presents a speech, a lecture, or a discussion. For consistency, questions and answers segments were separated from the original lecture as an independent session. This resulted in a total of 521 sessions with an average length of 13.56 min, and a total of 119 hours. A subset of 252 sessions (62h52m) among the fully parallel (i.e recordings were completed for both languages with English speaker and Arabic interpretation) have been used for this study. A total of 12 interpreters delivered the live interpretation in these events.

3.2. Audio Transcription

Once the audio collection was completed and separated by session and language, we developed guidelines for transcription to be carried by a professional agency. The guidelines include[6]. :

- Tagging Named Entities (Persons, Locations, Organizations)
- Capturing Non-lexical noises, i.e. Breathing, Laugh, Applause, Music.
- Capturing speech acts such as Repetitions, Hesitations, Interjections, False Starts and Corrections.
- How to limit segment size.

Figure 1 shows excerpts from the transcripts of an English audio and its Arabic interpretation with various annotations for NEs, speech acts and timing information.
The aim of the additional annotation was to capture the maximum features that allow to exploit the verbatim transcriptions for automatic recognition and processing of speech – for example, speaker information can be used for speech diarisation – while the resulting parallel data can be used in building automatic machine translation.

```
514
00:31:56,266 --> 00:31:58,862
well so that is the way I think you know [NE:PER Danny Weitzner]
515
00:31:58,862 --> 00:32:01,986
[REP who is] [HES] [NOISE] is was [REP in] [FALSE DC]
[REP in] in [NE:LOC Washington DC] for a while
516
00:32:01,986 --> 00:32:05,072
[HES] [FALSE for the] working [FALSE with the] [REP for the]
for the government.
517
00:32:05,072 --> 00:32:06,605
and he is back at [NE:ORG CSAIL] now.
518
00:32:06,605 --> 00:32:09,785
he is a lawyer [NOISE] [REP and] and a principal investigator
at [NE:ORG CSAIL].
------------------------------
160
00:31:59,489 --> 00:32:12,985
[تردد] [علم: شخص داني وايتتر] [تردد] موجود في [علم: مكان وشنطن دي سي] بالنسبة
[تردد] والآن عاد إلى [علم: مكان سيئل] أو [علم: مكان سياتل] [تردد] وهو محامي.
```

Figure 1: Speaker's (English) and corresponding interpreter's (Arabic) original transcripts with various annotations and time information.

3.3. Translation of Transcripts

The transcripts of both the original audio and the interpretations were translated into Arabic/English in accordance to which language the original speech was given in.
The goal for transcripts translation was to provide ground for comparison between the original text and the output of the interpreter. Because of this, the transcripts were translated to be able to perform experiments comparing translation features with interpretation features. Another reason was that we wanted to be able to run comparison of the original speaker's transcript and the interpreter in the same language. This could be achieved cross-lingually but the ideal scenario would be to have the material in the same language.
The translations were done by a professional translation agency[7]. The translation was guided by rules to ensure the quality and the style. Some of the translation rules include:

- The translation should be faithful to the original text in terms of meaning, cultural assumptions, and style, while preserving grammaticality, fluency, and naturalness.
- The translator is expected to maintain the same speaking style or register as the source. For example, if the source is polite, the translation should maintain the politeness. If the source is excited or angry, the translation should convey the same tone.

[6]Complete transcription guidelines are available at http://alt.qcri.org/resources/wawcorpus

[7]Transcription and translation were done by different professional agencies.

- The translation is expected to contain the exact meaning conveyed in the source text, and it should neither add, nor delete information. For instance, if the original text uses "Bush" to refer to the former US President, the translation should not be rendered as "President Bush, George W. Bush", etc.

The structure of the resulting corpus is shown in Figure 2, where for each speech and its corresponding interpretation, there are transcriptions and translations of the transcriptions into the other language.
The corpus contains additional meta data, including the language of the original speech and its interpretation, information regarding which interpreter interpreted which speech, the length (in seconds and in words) of the speech and of its interpretation[8].

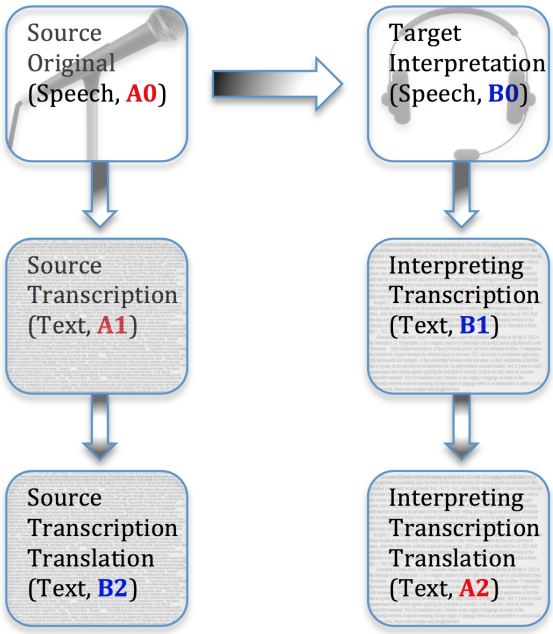

Figure 2: WAW Corpus structure. For each audio and its corresponding interpretation (A0, B0) there are transcripts (A1, B1) and translations (B2, A2).

4. Corpus Content Assessment

The transcriptions of the original audios and their respective interpretations were carried out independently. This resulted in *differences in time segmentation* and different number and length of segments in the original speakers' transcripts versus the interpreters' ones. In Figure 1 we see an example where the English part has been split by the transcriber into 5 short segments whereas the transcriber of the Arabic recording combined it into 1 segment.
We can also see a *difference in the number of speaker's words versus the ones generated by the interpreter*. This difference is due to the *language pair* and to the *strategies applied by the interpreter* (see Sections 4.1. and 5.1.). A detailed summary of corpus statistics is provided in Table 2.

[8]http://alt.qcri.org/resources/wawcorpus/releases/WAW-Readme.txt.

	English	Arabic
N. of files	126	126
Total Time	31:26:24	31:26:24
N. of Segments	26,572	9,413
N. of Words	286,024	156,814
N. of Words Translation	222,025	194,927

Table 2: WAW Corpus size in total number of files, recorded time, number of lines/segments, and number of words.

The simultaneous interpreter is *usually lagging a few seconds behind the speaker*. In the example in Figure 1 the interpreter started with 3 seconds delay and also ended with a delay of approximately 3 seconds. This décalage has been investigated in a number of studies, e.g. (Kroll and De Groot, 2009).

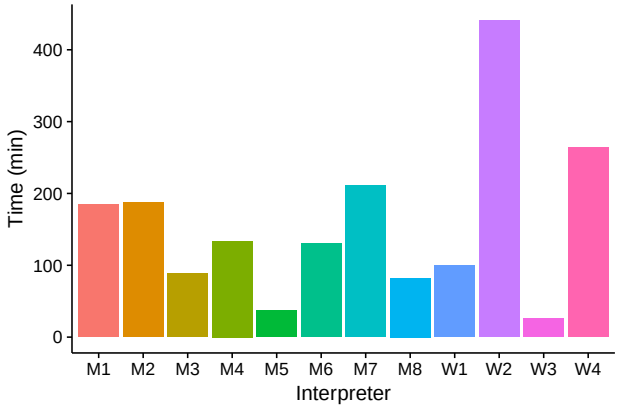

Figure 3: Interpreters' total time in the corpus (Mx: Male, Wx: Female)

In the following, we present an overall inspection of the contents of the corpus, in order to get an idea of the quality and the nature of its data. To achieve that, we used two measurements: a *lexical measurement of token ratios* and a *semantically oriented measurement of the distribution of "Named Entity" (NE) tags* used in the corpus and compared them cross-linguistically between the speaker and the interpreter.

4.1. Token Ratios

Arabic is an agglutinative language, in which several morphemes get fused together to compose a single Arabic word. As noted in corpora studies, the ratio between Arabic to English in typical written text is around 0.7 (Salameh et al., 2011). This is what we do also observe in the WAW corpus when we compare the number of tokens of the Arabic translations (B2) compared with the transcript of the original speaker's speech (A1, in English). However, when we compare the number of word tokens in the transcripts of the interpreters (B1) with the number of tokens in the translations of these transcripts (A2), we see a much smaller ratio. Figure 4 shows these ratios across documents grouped by interpreters. On average less text is produced by the interpreters compared to the original speakers and also com-

pared to the translators. In addition, we observe a wider distribution of ratios for the interpretations than for the translations. This reflects the challenging aspect of the interpretation work and the cognitive load the interpreter has to deal with while carrying the task of interpretation.

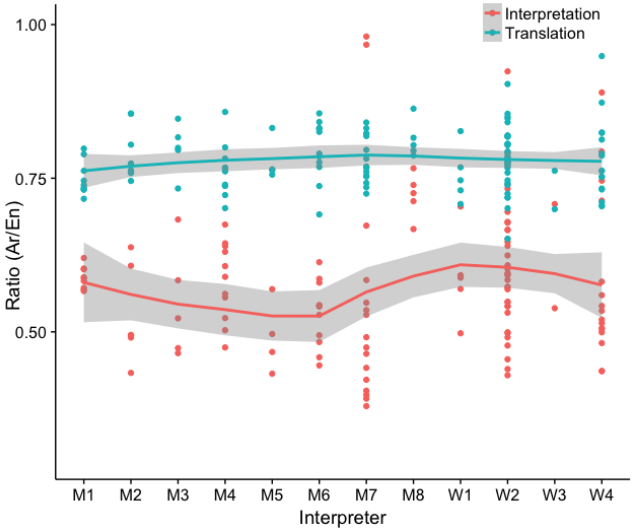

Figure 4: Comparison of the interpretation/original speaker and translation/original speaker ratios.

4.2. Distribution of Annotated Named Entities

Measuring the interpretation accuracy is not a straightforward task (Kahane, 2000; Zwischenberger and Pöchhacker, 2010). To measure it, we selected the "Named Entities" tag as an indicator or a feature to compare them cross transcripts. The expectation is that in the ideal scenario the interpretation should contain the same (or most of the) named entities that are mentioned in the original transcript. We realize that this is not a precise evaluation as the interpreter or the speaker could use pronominal or other types of references, which might not be captured while annotating the data. This approach nonetheless could still be helpful in the initial assessment. Figure 5 shows the distribution of the NE tag ratios between the original and interpreter transcripts. Most of the medians are around 0.8, which indicates that the interpreters were able to reproduce 80% of the NEs mentioned by the speaker. Interpreter M2 is a clear outlier, generating on average only 30% of the named entities. It will require more in-depth analysis to see if this is the result of a specific strategy, e.g. substituting names by pronouns, or just the indication of a sub-par interpretation.

5. Case Study: Interpreting Strategies Annotation

As a use case of the WAW interpretation corpus, we present a study of manually annotating the interpreting strategies in the corpus. The aim of this study was to reveal which strategies interpreters from English into Arabic use, and how often. The hope is that this might eventually provide some indications if our speech-to-text automatic translation system (Dalvi et al., 2017) could benefit from implementing some of these human interpreting strategies.

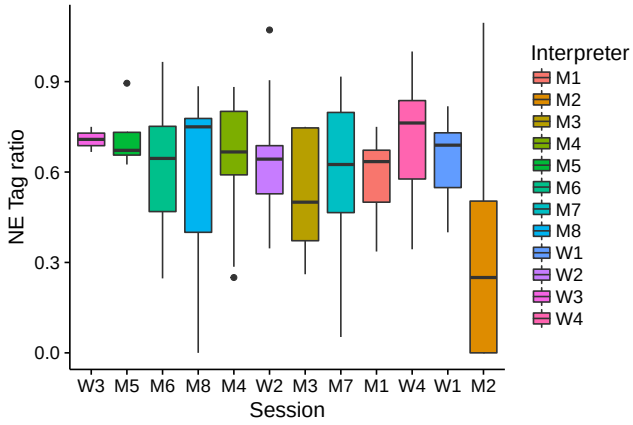

Figure 5: Ratios of NEs between the interpretation transcripts and the original speaker's transcripts (Mx: Male, Wx: Female)

The study was also motivated by our previous observation on the differences in the word ratios between interpretations and translations and the discrepancy in number of named entities tags.

This study is a follow-up of (Temnikova et al., 2017), where we conducted a preliminary annotation of the WAW corpus for interpreting strategies by analyzing a small sample of 7500 words (English+Arabic) from the transcripts of 4 sessions, with 2 female interpreters, including W2 (the most productive interpreter in our corpus), and 2 talks for each of these interpreters. For the study reported in the current paper we expanded the sample to 8 sessions, done by 4 interpreters, 2 women (W2, W4) and 2 men (M1, M7), adding up to 16,955 words in English and 9,477 words in Arabic. We selected these specific interpreters based on the transcript ratios against the original speakers, picking one with a high ratio and one with a low ratio for both male and female interpreters. As something new, added to our previous preliminary study, we also combined female and male interpreters, in order to further investigate whether there is any gender difference.

Finally, we also revised our annotation guidelines by adding more categories. A professional translator – native speaker of Arabic and fluent in English – with expertise in translation strategies annotation, was recruited to carry these annotation tasks.

5.1. Interpreting Strategies Annotation Guidelines

Our annotation guidelines are based on the state-of-the-art work on interpreting strategies (Roderick, 2002; Shlesinger, 1998; Bartłomiejczyk, 2006; Kalina, 1998; Al-Khanji et al., 2000), and on practical observations of the WAW corpus (Temnikova et al., 2017). Our updated annotation guidelines are available online[9].

Our annotation categories can be grouped into five major groups (as defined by our annotator):

1. **Summarizing**: The interpreter combines two clauses

[9] http://alt.qcri.org/resources/wawcorpus/releases/WAWCorpusSegmentationandAlignmentGuidelines.docx

into one or summarizes a single but longer clause.

2. **Self-correction**: The interpreter usually uses "أو" (or) or repetition to alter a lexical choice or correct a mispronounced word. Also, the interpreter may correct or modify the part of speech.

3. **Omissions**: The interpreter omits words which were in the source language text. We have a further subdivision into 3 categories of omissions (further information can be consulted in our guidelines).

4. **Additions**: The interpreter adds additional information, which is not found in the source text. We have 5 varieties of additions, which can be seen in our guidelines.

5. **Transliterations**: The interpreter uses the source language word in the target language. E.g. using "الموكس" for MOOCS.

For these categories, the expert annotator was asked to both provide a classification category of an interpreting strategy, and to provide an evaluation whether this "change" was *acceptable* or *not acceptable* (in order to further identify interpreters' errors versus successful strategies). (For more details and examples, please see our annotation guidelines.)

5.2. Analysis of the Results

The analysis of the results is based on these five aforementioned categories (see Figure 6). As it can be seen in the figure, corroborating the results of our previous study, the omissions are the highest number of interpretation strategies used, followed by additions. The ratio of omissions correlates with the Arabic to English word ratios with a coefficient of 0.89%. This is typically a strong correlation, but the other strategies show loose correlation with the Arabic to English word ratios (the coefficient varies between -0.12% and 0.38%).

On the other hand, we know that the different strategies affect the text of the transcript in a different way. "Additions" expand the text by adding newer content that was not in the original version of the text. This impacts negatively the ratio. In our scenario for interpreting from English into Arabic, the ratio "English/Arabic" tokens gets smaller when the denominator gets bigger (when more text is added to the Arabic interpretation). For this reason, we thought of looking at the strategies in combination. The different categories of strategies impact on the resulting transcript as a whole rather than individual. The new formalism we obtained was:

$$Combined = Addition - Omission + Self\ Correction - Summarizing \quad (1)$$

The new combined metric in Eq. 1 modeled better the impact on text that the various strategies employed by the interpreters, had. The correlation ratio between *Combined* and *Ratio (En/Ar)* is *0.92%*. This better explains the drop in the text ratio and how closely it is related to the number of strategies employed.

From another point of view, we did not notice any gender-related differences (at least for this small sample).

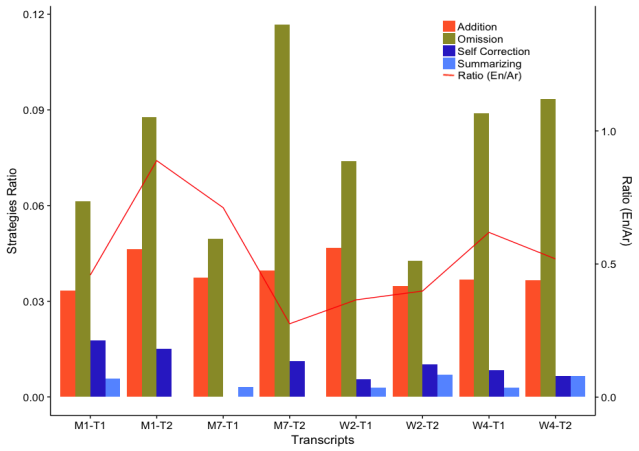

Figure 6: Distribution of Strategies normalized by the transcript length for each interpreter versus the ratio of Arabic tokens over English.

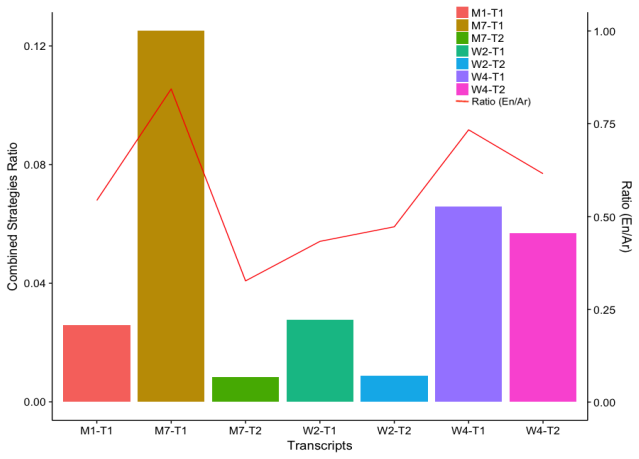

Figure 7: Combined Strategies normalized by the transcript length versus the ratio of Arabic over English.

6. Conclusions

This article presents our (a first) corpus of conference speeches, interpreted, transcribed, and translated, for the English-Arabic language pair. We provide some statistics and assessment of the corpus, as well as a case study involving interpreting strategies annotation by an expert. The findings from the annotations explain the differences between translated and interpreted material. The amount of omitted and summarized material skew the normal ratio between original documents and their translation. We would like to further exploit these findings and employ them for Machine Translation and S2T system where time is more critical. As future work, we plan to continue collecting more annotations at this level of granularity (omissions, additions, etc.), as well as to deepen the analysis into more detailed strategies. Further, we plan to automatize the annotation at the omissions and additions level, as well as to train our in-house speech-to-text machine translation system (Dalvi et al., 2017) on the basis of our findings.

References

Al-Khanji, R., El-Shiyab, S., and Hussein, R. (2000). On the use of compensatory strategies in simultaneous interpretation. *Meta: Journal des traducteurs/Meta: Translators' Journal*, 45(3):548–557.

Ali, A., Bell, P., Glass, J., Messaoui, Y., Mubarak, H., Renals, S., and Zhang, Y. (2016a). The MGB-2 Challenge: Arabic multi-dialect broadcast media recognition. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*.

Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., Bell, P., and Renals, S. (2016b). Automatic dialect detection in Arabic broadcast speech. In *Interspeech, San Francisco, USA*, pages 2934–2938.

Ali, A., Vogel, S., and Renals, S. (2017). Speech recognition challenge in the wild: Arabic MGB-3. In *Automatic Speech Recognition and Understanding (ASRU), 2017 IEEE Workshop on*. IEEE.

Bartłomiejczyk, M. (2006). Strategies of simultaneous interpreting and directionality. *Interpreting*, 8(2):149–174.

Bendazzoli, C. and Sandrelli, A. (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.

Dalvi, F., Zhang, Y., Khurana, S., Durrani, N., Sajjad, H., Abdelali, A., Mubarak, H., Ali, A., and Vogel, S. (2017). QCRI live speech translation system. *EACL 2017*, page 61.

Falbo, C. (2012). CorIT (Italian Television Interpreting Corpus): Classification criteria. *Breaking Ground in Corpus-based Interpreting Studies, Bern et al., Peter Lang*, pages 157–185.

House, J., Meyer, B., and Schmidt, T. (2012). CoSi-A corpus of consecutive and simultaneous interpreting. *Multilingual corpora and multilingual corpus analysis*, 14:295.

Hu, K. and Qing, T. (2009). Hanying huiyi kouyi zhong yupian yiyi xianhjua jiqi dongyin yanjiu–yixiang jiyu pingxing yuliaoku de yanjiu [A corpus-based study of explicitation of textual meaning in chinese-english conference interpreting]. *PLA Foreign Studies University Journal*, 4:67–73.

Kahane, E. (2000). Thoughts on the quality of interpretation. *aiic.net May 13, 2000. Accessed September 30, 2017. http://aiic.net/p/197*.

Kalina, S. (1998). *Strategische Prozesse beim Dolmetschen: Theoretische Grundlagen, empirische Fallstudien, didaktische Konsequenzen*, volume 18. G. Narr.

Khurana, S. and Ali, A. (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*.

Kroll, J. F. and De Groot, A. M. (2009). *Handbook of bilingualism: Psycholinguistic approaches*. Oxford University Press.

Kumar, G., Cao, Y., Cotterell, R., Callison-Burch, C., Povey, D., and Khudanpur, S. (2014). Translations of the CALLHOME Egyptian Arabic corpus for conversational speech translation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*.

Roderick, J. (2002). *Conference Interpreting Explained (Translation Practices Explained)*. St. Jerome Publishing.

Salameh, M., Zantout, R., and Mansour, N. (2011). Improving the accuracy of English-Arabic statistical sentence alignment. In *Int. Arab J. Inf. Technol.*, volume 8, pages 171–177.

Shlesinger, M. (1998). Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta: Journal des traducteurs/Meta: Translators' Journal*, 43(4):486–493.

Temnikova, I., Abdelali, A., Hedaya, S., Vogel, S., and Al Daher, A. (2017). Interpreting strategies annotation in the WAW corpus. In *The First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT), co-located with RANLP 2017. September 7 2017, Varna, Bulgaria.*

Tohyama, H. and Matsubara, S. (2006). Collection of simultaneous interpreting patterns by using bilingual spoken monologue corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Wray, S. and Ali, A. (2015). Crowdsource a little to label a lot: Labeling a Speech Corpus of Dialectal Arabic. In *Interspeech*. (in press).

Zaghouani, W. (2017). Critical survey of the freely available Arabic corpora. *arXiv preprint arXiv:1702.07835*.

Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Zwischenberger, C. and Pöchhacker, F. (2010). Survey on quality and role: Conference interpreters expectations and self-perceptions. *Communicate! AIIC Webzine*.