Corpus Creation and Evaluation for Speech-to-Text and Speech Translation

Corey Miller corey.a.miller@nvtc.gov
Evelyne Tzoukermann evelyne.tzoukermann@nvtc.gov
Jennifer Doyon jennifer.doyon@nvtc.gov
Elisabeth Mallard elisabeth.d.mallard@nvtc.gov
National Virtual Translation Center, Washington, DC, USA

Abstract

The National Virtual Translation Center (NVTC) seeks to acquire human language technology (HLT) tools that will facilitate its mission to provide verbatim English translations of foreign language audio and video files. In the text domain, NVTC has been using translation memory (TM) for some time and has reported on the incorporation of machine translation (MT) into that workflow (Miller et al., 2020). While we have explored the use of speech-to-text (STT) and speech translation (ST) in the past (Tzoukermann and Miller, 2018), we have now invested in the creation of a substantial human-made corpus to thoroughly evaluate alternatives. Results from our analysis of this corpus and the performance of HLT tools point the way to the most promising ones to deploy in our workflow.

1. Introduction

Among other offerings, NVTC provides verbatim human translations of both text and audio/video (AV) materials from foreign languages into English. NVTC places a great emphasis on identifying efficient workflows employing the latest HLT tools in the spirit of Augmented Translation (AT), a more encompassing form of Computer-Assisted Translation (CAT) (Miller et al. 2020). This paper focuses on AT in support of translation of AV. Miller and Tzoukermann (2018) showed efficiency advantages through the incorporation of both STT, ST and MT into human audio/video translation workflows. This paper describes the beginning stages of a more comprehensive exploration of that space, focused initially on the creation of a corpus and the running and scoring of several STT and ST engines using it. Subsequent work will focus on an analysis of MT vs. ST and the relative efficiency of such workflows.

2. Corpus

In order to identify relevant tools and processes for its data, NVTC sought to develop a corpus based on data that would be representative of the kinds of AV materials it typically receives for verbatim human translation. Criteria included typical languages, presence of multiple speakers, conversational/colloquial language, and pertinence to domains such as technological/scientific, cultural and political. Table 1 provides a summary of the languages sampled and the quantity of material in hours. All of the material was originally in video format and was converted to audio format so that both could be used as will be described below.

Language	Hours	Number of files
Arabic (Saudis speaking Modern Standard Arabic [MSA])	1	1
French (France)	2	1
Russian	6	4
Persian (Iran)	4	4

Table 1. Languages and quantity of associated data.

Once the source data had been identified, we developed a protocol for what kinds of human-produced output we wished to develop and how to instruct the participants to produce it. While NVTC's human translators (known as "linguists") typically only provide an English verbatim translation of foreign language source material, we sought to also include a foreign language transcription task since the most common speech analytics available today render a transcription in the same language as the AV input.

Accordingly, the first human-produced output we specified was a verbatim source-language transcription. Since verbatim translations (and transcriptions) often require timepoints and indications of who is speaking, we sought to identify a tool to facilitate linguists' annotation of this information. ELAN (2021) was deemed to be the most modern, flexible and well-supported of such tools.

Both the video and audio pertaining to a given file were loaded into an ELAN project. The video was included since it supplies useful information about who is speaking and provides extralinguistic context that facilitates transcription. Audio was provided in the form of a waveform in order to provide an easy way for linguists to demarcate the section being transcribed.

Linguists were asked to put the transcription of each speaker's utterances on a separate tier. They were asked to transcribe a single interpausal unit (IPU, Hosaka et al., 1994) at a time by selecting a portion of the waveform pertaining to the IPU and providing the source language orthographic transcription (to be described in more detail below) on an annotation tier identified with the speaker's name. This method obviated the linguist needing to explicitly annotate the start and end times of each IPU (a process subject to error), since they could be exported automatically from ELAN as will be described below.

Since people often do not speak in well-formed sentences, the IPU represents a convenient segmentation. In addition, its limited size lends itself to STT word error rate (WER) scoring (Jonathan Fiscus, personal communication) and serves as a spoken analogue of the translation unit (TU) (Hosaka et al., 1994), which is a normally a sentence in textual materials. Figure 1 shows the ELAN interface including French video, audio waveform, individual speaker tiers, and source language transcription of two IPUs by two speakers.

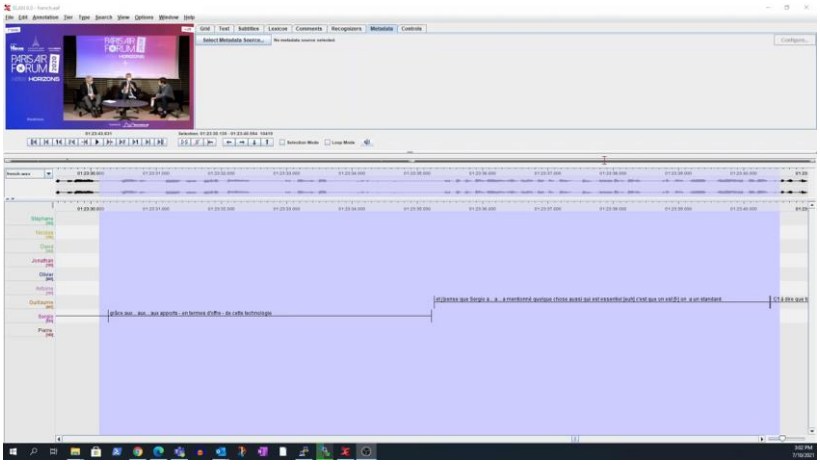

Figure 1. ELAN interface.

Once the transcription of a file was complete, its contents could be exported from ELAN as a tab-delimited text file containing the start time, end time, tier/speaker name and transcription of each IPU. This file could be loaded into an Excel spreadsheet, as shown in Figure 2, and then loaded into a CAT tool to be translated into English. Each transcribed IPU would serve as a source TU that would then be rendered as a target TU and serve toward the construction of a speech-oriented TM. Once the translation was completed, it could be output as an Excel spreadsheet, as shown in Figure 3.

Sergio [3]	4997.66	5000.375	mais il faudra un peu de temps pour l'implémenter
Sergio [3]	5000.375	5005.325	et ilf faudra une progressivité - je pense que on apprendra à utiliser la technologie-
Sergio [3]	5005.335	5010.279	de plus en plus, en fonction de [euh] de cas qui seront découverts
Sergio [3]	5010.279	5015.223	grâce aux... aux... aux apports - en termes d'offre - de cette technologie
Guillaume [2]	5015.266	5020.4	et j'pense que Sergio a... a... a mentionné quelque chose aussi qui est essentiel [euh] c'est que on est [f-] on a un standard.
Guillaume [2]	5020.412	5024.802	C't à dire que tous les aéroports [souffle] [euh] peuvent déployer ce... ce standard [souffle]

Figure 2. Sample Transcription File.

Sergio [3]	4997.66	5000.375	but it will take some time to implement
Sergio [3]	5000.375	5005.325	and it will need to be done gradually - I think we'll learn to use the technology...
Sergio [3]	5005.335	5010.279	more and more as new use cases are discovered
Sergio [3]	5010.279	5015.223	thanks to the benefits brought - in terms of offer - by this technology
Guillaume [2]	5015.266	5020.4	and I think Sergio also mentioned something that is essential, which is the fact that we do have a standard.
Guillaume [2]	5020.412	5024.802	It means that all airports can deploy this ... this standard

Figure 3. Sample Translation File.

In order to facilitate transcription and translation, linguists were instructed to follow their normal style guide. Traditionally, transcription for the purpose of STT evaluation has advised certain normalizations, such as lowercasing, avoiding punctuation and transcription of numbers as words rather than numerals[1]. However, given that we planned to evaluate several STT and ST systems, some of which transcribe numbers and punctuation in sophisticated ways, we felt it best to allow the linguists to transcribe things the way their final products were intended to be presented, e.g., including casing, punctuation and context-dependent representation of numbers as either numerals or words. That would give us an opportunity to evaluate these more sophisticated features should speech analytics attempt them. We also felt that normalization/simplification of standard forms, if necessary, would be easier than trying to infer the more sophisticated forms from simpler ones.

The style guide advises linguists to use standard orthography. We anticipated this might be a problem in Persian where there is typically a wide "diglossic" divergence between the written (standard) and spoken (colloquial) registers (Miller and Saeli, 2016; Saeli and Miller, 2018). However, we were surprised to see that French transcribers introduced a number of colloquial spellings as well, to be described below.

Finally, the style guide permits linguists to provide "exegetical remarks" in square brackets. In our case, these provided a useful way to isolate fillers/disfluencies such as *um* and *uh*, non-speech (e.g., music, coughs) and cut-off words (such as *hel-* or *-lo* for *hello*).

3. Speech Analytics and Scoring

Since our linguists most often translate foreign language source AV into English, our earlier work (Tzoukermann and Miller, 2018) led us to believe that ST would ultimately provide the best accuracy and efficiency outcomes with respect to enhancing translation workflows with HLT. Since ST goes from source language audio directly to target language text, it has access to rich audio information, such as stress/focus and emotion that would be lost in typical text STT output that the alternative of an STT + MT pipeline would provide. Salesky et al. (2021) offer a promising methodology for comparing STT+MT pipelines vs. ST that we hope to follow in our next stage of research.

Until then, we sought to obtain a baseline assessment of STT performance. The traditional metric is WER, but it should be noted there are several additional metrics we would like to explore as we proceed, including diarization error rate (DER), punctuation error rate (PER), and other advanced features considered in NIST's Rich Transcription Evaluation series[2].

WER calculations require an evaluation tool, reference transcriptions and hypothesis transcriptions for a given set of files. We used two evaluation tools, NIST's *sclite*[3] and a government off the shelf (GOTS) tool called *compute-wer*. Both tools take reference transcriptions in stm format and hypothesis transcriptions in ctm format. Figure 4 provides an example portion of an stm file corresponding to the transcription file shown above; they are both segmented at the IPU level. Note that it has been lowercased and most punctuation has been removed. In addition, square brackets have been converted to parentheses, so that this material can be ignored for the purposes of WER calculation (per *sclite's* -D or *compute-wer's* --sclite-parse options). Note also the presence of speaker names which allows speaker-specific WER calculation. This was helpful in identifying issues such as codeswitching as will be described below.

[1] Examples include https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-human-labeled-transcriptions and https://aws.amazon.com/blogs/machine-learning/evaluating-an-automatic-speech-recognition-service/.

[2] https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

[3] https://github.com/usnistgov/SCTK

Figure 5 provides an example of a hypothesis ctm file from one of the STT systems we evaluated. Note that it is segmented at the word level. Most of the STT engines we evaluated provide their output in json format. We are surprised that there does not seem to be any W3C guidance or standard for the presentation of STT output. Nevertheless, we were able to straightforwardly convert the various output formats to ctm via Python script.

```
french A Sergio 5000.37 5005.32 et ilf faudra une progressivité - je pense que on apprendra à utiliser la technologie-
french A Sergio 5005.33 5010.27 de plus en plus  en fonction de (euh) de cas qui seront découverts
french A Sergio 5010.27 5015.22 grâce aux  aux  aux apports - en termes d'offre - de cette technologie
french A Guillaume 5015.26 5020.40 et j'pense que Sergio a  a  a mentionné quelque chose aussi qui est essentiel (euh) c'est que on est (f-) on  a un standard
french A Guillaume 5020.41 5024.80  C't à dire que tous les aéroports (souffle) (euh) peuvent déployer ce  ce standard (souffle)
```

Figure 4. Sample portion of a reference stm file.

```
french A 5010.27 0.7599999999993088 grâce 0.993
french A 5011.04 0.6400000000003274 aux 0.983
french A 5011.69 0.2000000000007276 aux 0.974
french A 5011.89 0.7799999999997453 apports 0.96
french A 5012.82 0.2600000000002183 en 1.0
french A 5013.08 0.3599999999996726 termes 0.519
french A 5013.44 0.5300000000006548 d'offre 0.661
french A 5013.98 0.17000000000007276 de 1.0
french A 5014.15 0.18000000000029104 cette 0.993
french A 5014.33 0.5699999999999709 technologie 0.992
french A 5015.12 0.32999999999992724 Et 1.0
french A 5015.45 0.07999999999992724 je 1.0
french A 5015.53 0.17000000000007276 pense 1.0
french A 5015.7 0.11999999999989086 que 1.0
french A 5015.82 0.11000000000058208 c'est 1.0
french A 5015.93 0.0999999999994543 un 1.0
french A 5016.03 0.13000000000010914 jeu 0.998
```

Figure 5. Sample portion of hypothesis ctm file.

Once the stm and ctm files were prepared, we were able to calculate WER for each file, language and speaker for each speech engine that featured the language. Table 2 shows the engines that we evaluated, in anonymized form. We considered four commercial off the shelf (COTS) and three GOTS engines. Each engine has a different set of languages available, and some engines provide more than one locale per language. We used the most relevant locales when available. Even though our French file was from France, COTS 2 only had Canadian French (CA), so we also tested Canadian French in addition to European French (FR) with COTS 1, which had both. Only one engine provided ST output; however, that engine also provided STT output, so that is what was used in the evaluation described here.

STT	ST	Languages
COTS 1		Arabic (SA[4], AE[5]), French (FR, CA), Persian, Russian
COTS 2		Arabic (EG[6]), French (CA), Persian, Russian
COTS 3		French (FR), Russian
COTS 4	✓	Arabic (SA, AE), French (FR), Russian
GOTS 1		Arabic, Russian
GOTS 2		Arabic, Russian
GOTS 3		Russian, Persian

Table 2. Speech Engines Evaluated.

4. STT Results

We present WER results per language, distinguishing between files when there is more than one. For French, we additionally provide per-speaker results. Since WER is an error rate, lower is better, so we order the engines in increasing order, with the better performing ones on top.

4.1. French

French STT results are shown in Table 3 where five STT engines were available. As discussed above, where possible, both Canadian and European French were tested, and when only Canadian French was available, that was used. As shown in Table 3, European French and Canadian French STT were very close in results for COTS 1, which had both locales.

Engine	WER
COTS 4	18.4
COTS 1-European French	20.3
COTS 1-Canadian French	20.8
COTS 3	24.4
COTS 2-Canadian French	49.1

Table 3. French STT Results.

Table 4 below breaks the results down by speaker; number of words are provided in order to indicate the relative quantity of speech per speaker. Note that the speaker who uttered the largest number of words, Guillaume, was generally better recognized than Stéphane who uttered less than half as many words. This shows that the WER is not a function of the amount of uttered speech, but rather a function of the quality of the uttered speech. Indeed, Guillaume was the facilitator of the debate, and he may well have been trained to speak very clearly. Sergio, who spoke the second-highest number of words, was the best recognized of all speakers across all the engines. His speech rate was slightly slower than the other speakers which we speculate accounts for the better performance on his speech.

[4] Saudi Arabia
[5] United Arab Emirates
[6] Egypt

		COTS 1 CA	COTS 1 FR	COTS 4	COTS 2	COTS 3
Speaker	**# Words**	**WER**	**WER**	**WER**	**WER**	**WER**
Antoine	1904	20.4	22	17.6	56.6	27.5
David	2008	24.2	24.4	22.3	53.9	24.4
Guillaume	5127	20.9	20.4	19.2	46.1	25.7
Jonathan	1681	24.2	21.1	18.2	57.4	24
Nicolas	2296	18.3	18.1	16.9	55.4	20.7
Olivier	1965	23.3	22	21.8	50.1	26.9
Pierre	1785	19.6	19	16	47.5	25.8
Sergio	3964	15.7	15.1	14.3	36.8	18.7
Stéphane	1216	30.2	29.7	24.2	60.4	33.6
Sum/Avg	**21946**	**20.8**	**20.3**	**18.4**	**49.1**	**24.4**

Table 4. French STT Results by Speaker.

The error analysis showed discrepancies between colloquial French and more formal French. Colloquial examples supplied in the reference include *y'a* for *il y a* 'there is', *p'tit* for *petit* 'small', *c'qui* for *ce qui* 'which'. These appear to be efforts by the transcribers (in contrast to the instructions in their style guide) to reflect the conversational nature of the speech by trying to capture a fast speech pronunciation rule, schwa deletion (Barnes and Kavitskaya, 2002), in a colloquial orthography. This would be akin to representing a word such as English *running* as *runnin'* to indicate the speaker had not articulated the standard /ŋ/. While it is possible such colloquial spellings might be welcome in some contexts, they are a source of errors unless an STT engine happens to use these at the same time as a transcriber. This introduces interesting questions about how register should be accommodated and controlled in STT, a topic we discussed earlier with respect to MT and CAT (Miller et al., 2018).

Additionally, word boundaries were the cause of multiple errors, particularly for French hyphenated words, where reference hyphenated multiword units such as *est-ce* 'is this', *c'est-à-dire* 'that is to say', *peut-être* 'perhaps', and *quand-même* 'still', were rendered differently by some STT engines, resulting in errors. One of the complexities of a multi-engine evaluation such as ours is that transcription normalization for the purpose of achieving "comparable" WERs would need to be engine-specific. Our philosophy at this stage is to get a general idea of performance without substantial investment in normalization, under the assumption that different engines will both benefit and suffer from the reference transcriptions as they are, and intensive normalization would not be likely to cause the engines to stratify particularly differently in terms of performance. Another consideration is that if we take the reference transcriptions as indeed what the target should look like, then altering them to achieve a "more realistic" WER would be counter-productive since any edit distance between the reference and the STT would have to be "corrected" by a linguist.

4.2. Russian

The Russian data consisted of four separate files and seven STT engines were available to test. Results for each system are provided in Table 5. Russian 2 and Russian 3 had some speakers speaking English, which appears to have worsened results compared to Russian 1. At present, we have run only Russian STT on these files, but we hope to experiment with language diarization so that English STT can be run when English segments are detected.

	Russian 1	Russian 2	Russian 3	Russian4
Engine	**Word Error Rate**			
GOTS 2	19.9	27.4	30.3	32.8
GOTS 1	28.4	35.7	36.8	35.3
COTS 4	27.5	36.1	43.2	35.4
COTS 1	34.8	44.8	45.6	44.4
COTS 2	37.8	46.8	50.2	49.9
GOTS 3	40.1	46.4	49.6	48.4
COTS 3	53.2	53.7	56.5	58.8

Table 5. Russian STT Results by Engine and File.

We focused on content words, rather than function words since content words are more semantically meaningful. When possible, we sought to determine which words in the reference transcriptions did not appear in the STT engine's lexicon: the out-of-vocabulary (OOV) words. We also examined the reference words that did not appear in an engine's hypotheses; these consisted of both OOV and in-vocabulary (IV) words. For the IV words, we suppose that an engine's failure to recognize them had to do either with the engine's pronunciation or language models or with the pronunciation or audio conditions of words as uttered.

Another class of errors consists of words that are not recognized for multiple reasons including text normalization, realization of numbers, word segmentation, and morphology. One example of text normalization is letter ё 'yo', which is often realized by transcribers and STT engines as e 'ye'. The interesting part is that all these classes overlap, thus the number of OOV words combined with morphology largely increases the number of problematic tokens. For example, the single adjective аддитивный meaning '3-d', as in '3-d printing', generates 186 morphologically inflected tokens covering a dozen inflected types.

For Russian, we particularly studied the results of GOTS 1, where 30% of the reference words did not appear in the hypotheses. Among these, 35% were OOVs and 65% were IVs but were presumably not recognized due to accent, position of the word in the sentence, ambient noise, etc. The following list samples recognition errors of various types of words:

- **OOV:** technical words and compounds, such as аддитивный '3-d', физическо-химических 'physico-chemical', экосистемы 'eco-systems'.
- **Mixed Russian and English Borrowings:** бизнес-задача 'business task', бизнес-модели 'business models', бизнес-секции 'business sections', интернет-площадке 'internet site'.
- **Borrowings:** слайд 'slide', принт 'print', лидер 'leader'
- **Morphology:** Russian has three genders (feminine, masculine and neuter) and 6 inflectional cases; this means that when one word is not recognized, all its inflected and derived forms will also likely be unrecognized. Morphological errors of IV items also occur such as технологий → технологии 'technology', которые → который 'which', развиваются → развивается 'are/is developing'.
- **Word segmentation:** какой-то / то 'some', вице-президент / президент 'vice-president / president', пост-обработка / постобработка 'post-processing / postprocessing'.

- **Numerals**
  - Normalization: 30 / тридцать '30 / thirty'
  - Normalization and morphology: 30-му / тридцатом '30 (dative) / thirty (prepositional)'

4.3. Persian

Our Persian data consisted of seven files, four of which have been analyzed so far. Results are presented in Table 6.

	Persian 1	**Persian 3**	**Persian 4**	**Persian 6**
Engine	**WER**			
COTS 1	45.8	32.9	52.2	38.3
GOTS 3	62	48	86.6	60.7
COTS 2	89.2	84.6	92.5	83.6

Table 6. Persian STT Results by File.

Typical errors were similar to those noted above under the colloquial rubric for French but often in reverse. For example, transcribers often used standard representations such as می کنند 'they do' and ندارد 'doesn't have' in cases where the best performing STT output colloquial forms such as می کنن and نداره. As in French and Russian, word segmentation issues also arose; for example, a transcriber might write میتونه where STT output می تونه 'is able'. Finally, we did make a concession to normalization by accounting for encoding issues, as different engines (and transcribers) sometimes used different Unicode codepoints for the letters ک 'kāf' and ی 'ye'.

4.4. Arabic

Arabic results are shown in Table 7. It turns out that Arabic, despite the perception that it is a complex language to recognize, demonstrates the best STT results. Top confusions evinced similar normalization issues to those discussed above, such as variable placement of *hamza* in reference and hypothesis.

Engine	**WER**
GOTS 2	12
COTS 2	19.8
GOTS 1	22
COTS 4 SA	22.2
COTS 4 AE	22.3
COTS 1 AE	27.8
COTS 1 SA	33.4

Table 7. Arabic STT results.

5. Conclusions

Since our main goal is to identify worthwhile insertions of HLT into the AV translation workflow, the work described here is really just the beginning. We are collecting additional details from linguists, such as time on task, which we are hoping to factor into our analysis. In addition, since completed translations also contain indications of who is speaking, we hope to incorporate an analysis of speaker diarization and potentially, speaker recognition. As has been made evident in the WER analyses of all the languages discussed here, getting to the bottom of how exactly certain classes of words should be represented in final transcriptions and translations, including register issues, will be important in order to assess to what extent speech analytics are contributing toward those objectives. We hope to look more carefully at the representation of numerals and punctuation, since if these are required in the end product, speech analytics that accurately represent them will be potentially more useful than those that omit or misrepresent them. Finally, we are keen to determine whether ST offers promise over STT and MT pipelines; if so, perhaps many of the source language transcription issues we have been discussing will cease to be important, since the focus will be on the translated English output.

References

Barnes, J. and Kavitskaya, D. (2002). Phonetic Analogy and Schwa Deletion in French. In *Berkeley Linguistics Society*, pages 39-50.

ELAN (Version 6.2) [Computer software]. (2021). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan.

Hosaka, J., Seligman, M., and Singer, H. (1994). Pause as a Phrase Demarcator for Speech and Language Processing. In Proceedings of the 15th Conference on Computational Linguistics, vol. 2, pages 987-991, Kyoto.

Miller, C., Higgins, C., Havens, P., Van Guilder, S., Morris, R., and Silverman, D. (2020). Plugging into Trados: Augmenting Translation in the Enclave. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas*, pages 469-477.

Miller, C. and Saeli, H. (2016). Second-level pluricentricity in the Persian of Tehran. In *Pluricentric Languages and Non-Dominant Varieties Worldwide*, R. Muhr (ed.), pages 191-204. Frankfurt.

Miller, C., Silverman, D., Jurica, V., Richerson, E., Morris, R. and Mallard, E. (2018). Embedding Register-Aware MT into the CAT Workflow. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, vol. 2, pages 275-282.

Saeli, H. and Miller, C. (2018). Some linguistic indicators of sociocultural formality in Persian. In *Trends in Iranian and Persian Linguistics*, A. Korangy and C. Miller (eds.), pages 163-182. Berlin.

Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D. and Post, M. (2021). The Multilingual TEDx Corpus for Speech Recognition and Translation. In *Proceedings of Interspeech 2021*. Brno.

Tzoukermann, E. and Miller, C. (2018). Evaluating Automatic Speech Recognition in Translation. In *Proceedings of AMTA 2018*, vol. 2, pages 294-302, Boston.