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.