Translating Dislocations or Parentheticals: Investigating the Role of Prosodic Boundaries for Spoken Language Translation from French into English

Nicolas Ballier nicolas.ballier@u-paris.fr
Laboratoire de linguistique formelle/ CLILLAC-ARP, Université Paris Cité, Paris, F-75013, France

Behnoosh Namdarzadeh behnoosh.namdar@gmail.com
Maria Zimina-Poirot maria.zimina-poirot@u-paris.fr
CLILLAC-ARP, Université Paris Cité, Paris, F-75013, France

Jean-Baptiste Yunès jean-baptiste.yunes@u-paris.fr
IRIT, Department of Computational Science, Université Paris Cité, Paris, F-75013, France

Abstract

This paper examines some of the effects of prosodic boundaries on ASR outputs and Spoken Language Translations into English for two competing French structures (*"c'est"* dislocation vs. *"c'est"* parentheticals). One native speaker of French read 104 test sentences that were then submitted to two systems. We compared the outputs of two toolkits, SYSTRAN Pure Neural Server (SPNS9) (Crego et al., 2016) and Whisper. For SPNS9, we compared the translation of the text file used for the reading with the translation of the transcription generated through Vocapia ASR. We also tested the transcription engine for speech recognition uploading an MP3 file and used the same procedure for AI Whisper's Web-scale Supervised Pretraining for Speech Recognition system (Radford et al., 2022).

We reported WER for the transcription tasks and the BLEU scores for the different models. We evidenced the variability of the punctuation in the ASR outputs and discussed it in relation to the duration of the utterance. We discussed the effects of the prosodic boundaries. We described the status of the boundary in the speech-to-text systems, discussing the consequence for the neural machine translation of the rendering of the prosodic boundary by a comma, a full stop, or any other punctuation symbol. We used the reference transcript of the reading phase to compute the edit distance between the reference transcript and the ASR output. We also used textometric analyses with iTrameur (Fleury and Zimina, 2014) for insights into the errors that can be attributed to ASR or to Neural Machine translation.

Keywords: MT with speech recognition, quality estimation, toolkit comparison, prosodic boundaries, parentheticals, dislocations

Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, pages 119–131
September 4–8, 2023, Macau SAR, China.

1 Introduction

In French (Fagyal, 2002), as in different languages (Dehé and Kavalova, 2007), parentheticals have specific prosodic patterns. Several papers have shown the crucial role of prosodic boundaries for dislocations in French (Ashby, 1994; Buthke et al., 2010; Avanzi, 2012). We aimed to investigate the effect of prosodic boundaries and analyse whether the second prosodic boundary of the parenthetical was accurately translated and distinguished from the final rise of the left periphery dislocations. Our small-scale analysis compares a successive pipeline including VOCAPIA and SYSTRAN automated speech translation generated by SYSTRAN Pure Neural Server (SPNS9)[1] and a multitask multilingual pipeline using Whisper, an Automatic Speech Recognition (ASR) system trained on audio data for transcription and translation.

Speech technologies such as Automatic Speech Recognition were already coupled with automatic translation within Phrase-Based statistical Machine Translation (PBMT) (Reddy et al., 2007). As part of a more general project on error evaluation of ASR systems, ERA project (adVanced ERrors Analysis for speech recognition), ASR systems have been analysed in (Santiago et al., 2015). In a more pragmatically oriented paper, eight ASR platforms were assessed for accuracy and time-saving purposes on five documents from different fields of research in humanities (Tancoigne et al., 2022).

Preliminary investigations of chatGPT-3 for translation suggest a better performance (Hendy et al. (2023) and Jiao et al. (2023); other audio Large Language models have been built for speech recognition such as LXSR-53 large model Grosman (2021)). The Whisper paper describes its performance in relation to other systems such as mSLAM (Bapna et al. (2022)), a multilingual Speech and Language Model that learns cross-lingual cross-modal representations of speech, trained on LibriSpeech and other resources like a thousand hours of speech from Babel.

The rest of the paper is organised as follows: Section 2 details our experiments; Section 3 presents the results; Section 4 discusses them and outlines further research.

2 Materials and Methods

2.1 Challenge Set Recording

We adopted a challenge set approach (Isabelle et al., 2017), by recording challenging examples compiled and adapted from attested data. We used adapted data from the CFPP corpus, ie *le Corpus de Français Parlé Parisien* (Branca-Rosoff and Lefeuvre, 2016). Our dataset also includes examples from (Tellier and Valois, 2006) and (Blasco-Dulbecco, 1999) for reported examples of dislocations in spoken French. Our challenge test is aimed at evaluating whether the systems correctly process the dislocation or the parenthetical structures and the punctuation symbols used in their transcripts. Our challenge set is much more modest than previous work in the field, such as (Besacier et al., 2014). We have centered our analysis on the potential ambiguity between parentheticals and dislocations, having noticed that dislocation is a troublesome construction for neural machine translation systems Namdarzadeh and Ballier (2022) and that when the dislocation was properly translated in the DeepL outputs, it nevertheless could entail potential ambiguities with parentheticals. We also wanted to analyze the ability of the models to translate right and left dislocations, so that we replicated textbook examples using the same constituent either in right or left periphery. The overall assumption is in spoken data constructions that structure is even more used and may be consequently troublesome for neural machine translation, given its rarity in the training data. We did not resort to an anechoic chamber for our recordings but used a standard headset when recording over Zoom in a quiet office. We

[1] The service is available via the platform *Pure Neural Server – CLILLAC-ARP:* `https://plateformes.u-paris.fr`

Size	Parameters	Required VRAM	Relative speed
tiny	39 M	1 GB	32x
base	74 M	1 GB	16x
small	244 M	2 GB	6x
medium	769 M	5 GB	2x
large	1550 M	10 GB	1x
large-v2	1550 M	10? GB	1?x

Table 1: Whisper models tested for this experiment

voluntarily used a Zoom recording facility for a more ecologically valid acoustic environment. We included several types of ambiguities to gauge the impact of the detection of silent pauses.

Our dataset also includes examples from Tellier and Valois (2006) and (Blasco-Dulbecco, 1999) for reported examples of dislocations in spoken French. We did not test the sound file with the best sampling rates. We converted the mp4 file generated by Zoom into an mp3 file that was compatible with the Vocapia SPNS9 system. It should be noted that Whisper down-converts to lower sampling rates.

2.2 Parameters

We used the Huggging Face distribution of the models trained on multilingual data.[2] Table 1 sums up the number of parameters for each model size from `tiny` to `large` models. As indicated in Radford et al. (2022), the distinction between the `large` model and the largest model (`large-v2`) is not based on a difference in the number of parameters but rather on a fine-tuning of the large model. As reported in the appendix of the Whisper paper, French is the fifth language for hours of speech in the training data for speech recognition with 9,752 hours and the eighth for translation (4,481 hours of audio).

The data was processed on a server using an NVIDIA A100 GPU.[3] We measured our carbon foootprint using the codecarbon **?**library (version 2.4.4), we used the 8 CPUs of an A100. 79s were required for the processing of our experiments and we reckon that it corresponds to an estimated total emission of 0.0002048757071268 g of CO_2.

2.3 Evaluation Metrics

We resorted to quantitative and qualitative analyses. With the Natural Language Toolkit (NLTK) library (Bird, 2006), we used BLEU score (Papineni et al., 2002) for the comparison based on our in-house translation dataset and Word Error Rate (WER) for the analysis of the discrepancies between the original script and the ASR transcriptions. As is well-known, WER is computed by adding substitutions (S), insertions (I), and deletions (D), divided by the N total words in the reference transcription, and multiplied by 100 as expressed in the formula (1).

$$WER = \frac{I + D + S}{N} X100(1)$$

We did not normalise the outputs in terms of capitals and punctuation, whereas Whisper has been tested using a normalisation procedure described in the appendix of the Whisper paper (Radford et al., 2022). To distinguish "innocuous differences in wording and genuine mistranscriptions", they used text normalisation to minimise the difference between strings like "ten thousand dollars" and "$10000".

[2] `https://huggingface.co/models?search=openai/whisper`

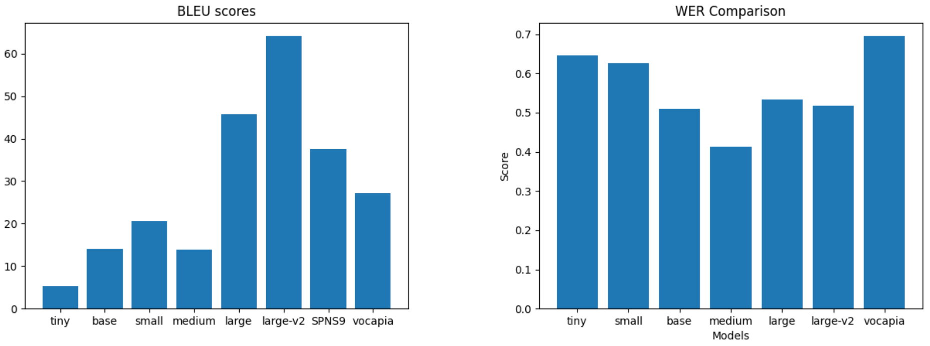
[3] `https://u-paris.fr/plateforme-paptan/`

3 Experiments and Results

Our analysis compares two translation pipelines for spoken data: ASR (Vocapia) then translation (SPNS9 MT engine) and audio LLM translation output from the speech signal with Whisper. In the Whisper output, the numbers of segments produced by the different models do not match for translation and transcription tasks, so that we can reasonably assume that the translation is not based on the transcription. We both resort to quantitative and to qualitative analyses. We report WER and BLEU scores for the different systems, analyse vocabulary growth curves for the data sets produced by different models and then discuss some characteristic phenomena, such as the use of punctuation marks (“.” and “,”) more qualitatively.

3.1 Quantitative Analysis of Translations and Transcriptions

Figure 1a sums up the results for the BLEU score and the effect of ASR errors that can be revealed when comparing the Vocapia ASR output to the SPNS9 translation which is based on the original transcript. The performance on translations for Whisper outputs needs to be related to the performance on the transcription task. It should be noted that not all the utterances we read were actually transcribed. This is why we realigned the translation outputs and, in the case of repetitions, we assumed the first occurrence was actually translated.



(a) BLEU scores (realigned output)

(b) WER scores (raw outputs)

Figure 1: Performance on translation (BLEU) and on transcription (WER)

We used the Python JIWER library to compute WER as shown in Figure 1b. The striking difference is the `medium` model does much better for transcription but not for translation.

3.2 Textometric Analysis

We used textometric analyses with iTrameur (Fleury and Zimina, 2014)[4] to compare raw translations of test models. Repeated segments computation (Lebart et al., 1997) might shed light on the automatic chunking produced by the machine to recognise text patterns and insert punctuation marks in translated output, as in the following lines produced by `small` model (segments with 10 or more repetitions in the test set output are underlined): *Comedian, he will always remain. Comedian, he will always remain. He was still a comedian. He was still a comedian. He was still a comedian.* One way to give an account of the textual production is to surmise that the machine tries to generate text chunks that are compatible with training data. It may as well be that the presence of these repeated segments is only the artefact of the somewhat artificial character of the textbook examples that are used in our data.

[4] `https://itrameur.clillac-arp.univ-paris-diderot.fr`

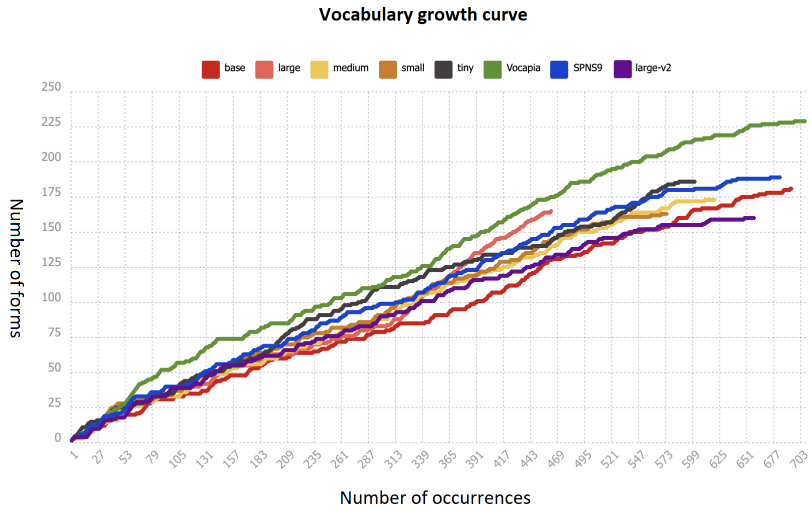


Figure 2: Vocabulary Growth Curves of the different models

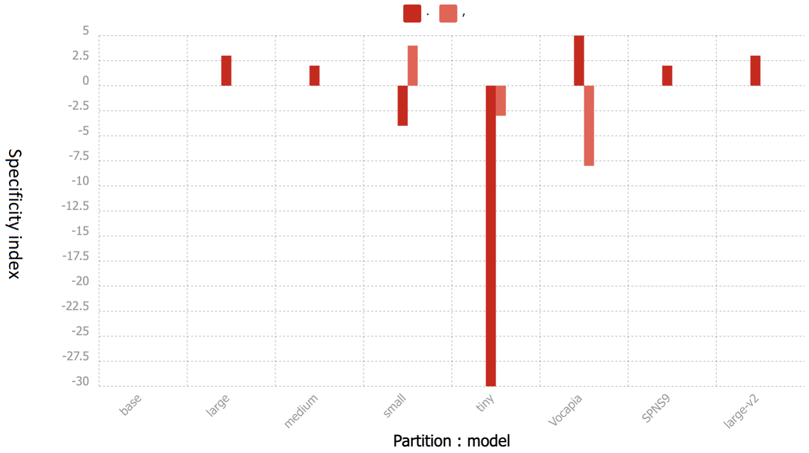


Figure 3: Characteristic elements: specificity of a full stop and a comma in raw translations

Exploring Vocabulary growth curves in Figure 2, one can notice that both output length (number of occurrences) and lexical diversity (number of different forms) are greater in the case of more "mature" models, such as `Vocapia`. The discrepancy shows that some tokens were misrecognised by "weaker" ASR models, such as `tiny`, hence the variability that can be observed here. From that point of view, the vocabulary growth analysis is a potential reflection of the errors that can be attributed to a specific ASR system: the difference between the Vocapia curve and the SPNS9 one corresponds to ASR-generated artefacts.

Figure 3 shows the results of characteristic elements computation of two punctuation marks (a comma "," and a full stop "."). For each model, the specificity index (Lebart et al. (1997)) reveals characteristic presence or absence of the two punctuation marks in the raw translations. For example, one can easily notice the absence of ``.") in the output of the `tiny` model (specificity index: -30), suggesting that the next token prediction on the basis of the transcribed data does not allow the model to identify the final stop. Conversely, the results presented on Figure 3 show that sentence boundaries are better transcribed by the `Vocapia`, `large` and `large-v2` models. At the same time, while the full stop is over-represented in the translation generated on the basis of the `Vocapia` transcription (specificity index: +5), clause boundaries isolated by commas are under-represented in this output (specificity idex: -8) revealing the presence of over-segmentation, for example: *I stayed there for a long time. In there. Do you know what he's doing.*

Thus, by studying the presence and absence of punctuation marks, it is possible to see the trough of the models, as for instance in the case of `tiny` and `small`, which often fail to produce translations in chunks of sentences.

3.3 Qualitative Insights

This sub-section attempts to characterise the observed behaviour of transcriptions/translations in particular the number of segments produced by each Whisper model. [5]

For the `large` model, it looks like some of the sentences that are repeated are not perceived as such, and a certain number of spoken utterances are reduced to only one sentence in the output. It remains to be seen whether a certain form of threshold for the duration between the different utterances can be observed. It may be the case that the distinction is not so much about pauses but about models. For example, the `small` model produces *"[...], free I never follow it at all[...]"* for *"libre, je ne le suis à peu près jamais"*, where the homonymous *"suis"* has been translated by *"follow it"* (in French: *"je le suis"*), which means that the sequence *"I am not"* was not related to the dislocated item *"libre"*. The absence of translation of the dislocation and the re-analysis of the sequence *"suis"* as being *"followed"* seems to prove that the dislocation was not perceived, perhaps due to the duration of the pause between the dislocated item *"libre"* and the corresponding predicate *"je ne le suis"*.

We then had to manually realign the different utterances to the corresponding sentences. We tried to keep the original punctuation of the model output so that many sentences ended with a comma where the original signal would have a full stop and a pause, a major boundary pause. The realignment process was not easy and guided with the original text that was used for the realisation of the sentences. For the `medium` model, the discrepancy between the transcription model output and the translation output is the most striking. The translation has 48 segments and the transcriptions have 102 segments. From the point of view of AI faithfulness, the `medium` model is pretty accurate in the translation of *libre, je ne le suis pratiquement jamais, free I'm almost never*, but the erasure or absence of reproduction of repetitions is also very striking. 56 sentences were omitted and the BLEU score would have been more degraded if the sentences were longer. We need to investigate whether the transcription output can duplicate the copied or repeated segments from the sound file, but do not include them in the translation output. It may be the case that the `medium` model might be the most efficient to suppress disfluencies, with the very unfortunate consequence that repeated segments get to be omitted in the transcription or at least in the translation. The `base` model gives examples of some absurd translations: with the use of Chinese character 465 and the translation "I asked him who's that pomm, he asked his poms you and asked the poms". Using a detached structure in the left periphery with a pause may trigger a phonological reanalysis in the left periphery of the dislocated constituent, this could account for the transcription of the sequence *danser,* as *dans ses* in the translation *In (=dans?) them (=ses??), she will do all her life*.

As reported in the Whisper paper, Named entity recognition (NER) still remains an issue: In French, the initial consonant for 'Chomsky' is realised as a voiceless fricative and not as an affricate, so that the closest transcription 'Jomski' (`large` model) fails to recognise the

[5] Data to be found on `https://github.com/nballier/NMT/tree/master/MTS2023`.

named entity. Interestingly enough, models have different transcriptions for this named entity: *je me skie* (`tiny`) *j'aime ce qui* (`base`), *James Key* (`medium`), *Jomski* (`large`) and *Jamsky* (`small`/`largev-2`). The tokens predicted for the translation of this named entity by the smaller models seem to correspond to a grammatical sequence, and models beyond the `small` one correspond to plausible proper nouns.

3.4 Punctuation and Prosodic Boundaries

Reference transcripts for the evaluation of Automatic Speech Recognition (ASR) usually imply removing punctuation (Matassoni et al., 2013) except apostrophes when normalising data before computing Word Error Rate (WER), sometimes reported as case-insensitive word error rates (Despres et al., 2013). For neural machine translation, a change in punctuation may entail on-the-fly modifications of the translation outputs on available on-line systems. Properly assigning punctuation symbols proves crucial for ASR systems and spoken language translation. Errors may entail linguistic ambiguity when prosodic boundaries help to recognise sentence structures such as dislocation (source: *"La traduction automatique neuronale, c'est impressionnant"* target: *"neural machine translation is impressive"*) and parentheticals (source: *"la traduction automatique neuronale, c'est vrai, est impressionnante"*, target: *"True, neural machine translation is impressive"*). In this context, when the autonomous parenthetical accent phrase is not perceived as parenthetical, the translation is *"neural machine translation is true"*.

4 Discussion

This experiment is really intended as a pilot study, we did not control for the effect of speech rate nor did we rely on inter-annotator agreement for the reference transcription TextGrid of the time stamps represented Figure 5.

4.1 Contextualisation

For a strict parentheticals versus dislocations comparison, as one of the reviewers suggested, we would need to report more information about the frequency of dislocations and parentheticals in the source language to judge impact more reliably. Taking reference treebanks as a proxy for integral corpus queries, we found that dislocations are more frequent than parentheticals : if we take the example of the CFPP Treebank, only 14 occurrences of parataxis (a more general label than just parentheticals) to be compared to 264 dislocations.

4.2 Time-Stamps in the Transcription Task

As described in the corresponding reference paper (Radford et al., 2022) Whisper uses text normalisation but little is known about the punctuation restoration task and how it fares on test datasets (Lerner et al., 2022): semi-columns are absent in the translation output dataset.

It may be an effect of the training data, and another feature of the training probably takes its toll, the segmentation into 30s windows. As described in the methodology of the Whisper systems "when a final transcript segment is only partially included in the current 30-second audio chunk, we predict only its start time token for the segment when in timestamp mode, to indicate that the subsequent decoding should be performed on an audio window aligned with that time, otherwise we truncate the audio to not include the segment." (Radford et al., 2022) That decision may explain why some of the timestamp boundaries in the SLT format often correspond to speech and not to pauses. In the figure describing the overview of the approach, the multitask training format does mention the timestamp tokens and their operationalizations as time-aligned transcription. Nevertheless, the variability across models of this time-aligned transcription is not reported in the Whisper paper. Admittedly, the emphasis of these generative pre-trained models is on generating texts but the generation of time stamps does not seem to have not been monitored so closely. Speech alignment is acknowledged to be potentially problematical above the 30 second window the models were trained with. The result is also a variability in the segmentation of speech. The same sound file produces different segmentations (and corresponding time stamps) across models for the translation and transcription task. Figure 4 recaps the effect of the size of the model on the number of segments for our data. Above the medium model, the intervals get bigger and segments are arbitrarily cut off as 2,3 or 5 second intervals, sometimes in the middle of the speech signal.

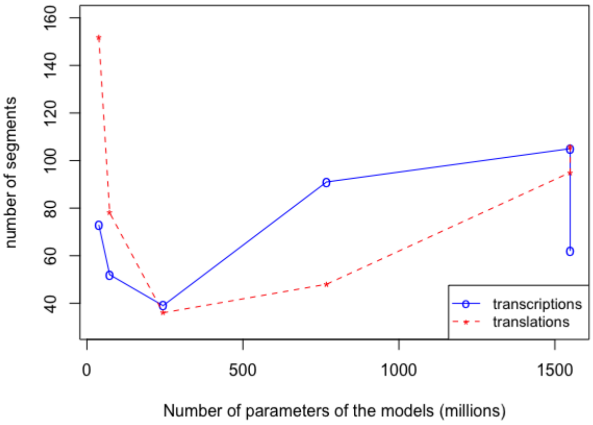
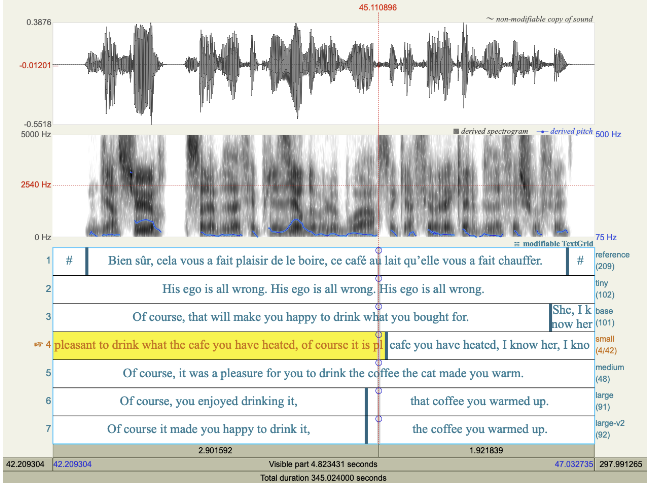


Figure 4: Variability of whisper segmentation across model size



1 # Bien sûr, cela vous a fait plaisir de le boire, ce café au lait qu’elle vous a fait chauffer. # reference (209)

2 His ego is all wrong. His ego is all wrong. His ego is all wrong. tiny (102)

3 Of course, that will make you happy to drink what you bought for. She, I know her base (101)

4 pleasant to drink what the cafe you have heated, of course it is pl cafe you have heated, I know her, I kno small (4/42)

5 Of course, it was a pleasure for you to drink the coffee the cat made you warm. medium (48)

6 Of course, you enjoyed drinking it, that coffee you warmed up. large (91)

7 Of course it made you happy to drink it, the coffee you warmed up. large-v2 (92)

2.901592 1.921839

42.209304 42.209304 Visible part 4.823431 seconds 47.032735 297.991265

Total duration 345.024000 seconds

Figure 5: Variability of whisper segmentation of time intervals across models

Silence portions are sometimes used as left initial boundary signals. In other terms, the chunk speech is represented as beginning with a pause. We have used the SRT output of the whisper models to convert them into a Praat (Boersma and Weenink, 2023) TextGrid (the standard tool for phonetic analysis) and we have realigned the original sound file onto a Praat transcription. Figure 4 displays the waveform, the spectrogram of the speech signal and the corresponding read speech (reference) and the tiers below it represent the different segmentations of the same speech utterance contained in the time stamps in the .SRT file outputs of the different models. The reference tier shows the presence of boundaries corresponding to the speech pauses represented by hashtags and the absolute absence of them in the different time representation of the translations. One can see that a single French sentence might potentially correspond to several segments in the base transcription. The second interval in the `small` tier includes an overlap of two utterances, while the `large` tier splits the reference into two intervals/segments. The vertical lines correspond to the segments/intervals and, interestingly enough, one can see that for the `large` and `large-v2` models the dislocated item almost corresponds to the phonetic boundary in the original sound file (represented by the vertical lines that crosses tiers). The beginning of the phrase is actually beginning on the vertical line following the acoustic cues. The interval boundary proposed for the `small` model somewhat reflects the beginning of the signal whereas the `base` model has almost the initial silence as a cue for the beginning of the following intervals. This Praat representation is very representative of the mismatches between the different intervals produced with the different models for the translation task (but this is also true for the Whisper transcriptions). For each tier, under each name of the model is the number of intervals. For the upper reference tier, we have 104 utterances and 105 silent intervals. One can see the variability of the different numbers of segments that are produced to supposedly align with the signal. It is striking that a unique sound file should have so many different time representations of the corresponding speech, especially for the transcription task.

4.3 Further Research and Generalisability

More experiments are needed to address the same phenomenon and perform testing with variable phonological environments in order to determine for example whether the liaison as a cue is really taken by the model. It is also important to note that the sentences of the test set were read successively but were realised in isolation, no co-referential cues were available in the data, contrary to what would be found in continuous speech.

It may be the case that some Whisper models eliminate disfluencies, hence the absence of repeated segments. Again, disfluencies and repairs are much less frequent in the training data based on read books. We segmented the sound file into smaller sound files using `ffmpeg`. We investigated whether we obtained more consistent intervals on these re-cut files. We did not systematically investigate the role of the variability of the speech rate on the cut-off points for full stops or commas but we represent the variability of the speech rate (number of syllables duration) in our data as calculated by de Jong and Wempe algorithm (De Jong and Wempe, 2009). Figure 6 shows the variability of the speed of the different segments.

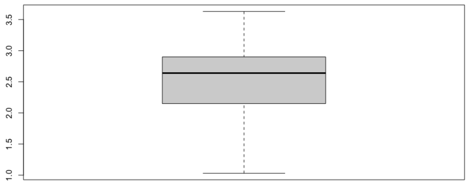


Figure 6: Variability of the speech rate (number of syllables per second) across utterances

As one of the reviewers wondered how portable the findings are to other language combinations, we replicated the experiments with a prototype dataset we designed for Persian (Namdarzadeh et al., 2022). We read it in the same recording conditions, but it should be noted that our test for Persian did not include repeated utterances with shorter pauses between the dislocated items and the predicate. We also read aloud the number of the utterance. From the point of view of transcriptions, we observed inconsistencies among different models of Whisper. In the case of `tiny`, `small`, and `medium` models, transcriptions are incomplete and they encountered difficulties in accurately identifying the correct graphic representation of a given phoneme in Persian, considering the possibility of several letters for the same sound. Furthermore, it should be noted that in the `tiny` model of Whisper, we observe a few Chinese and Spanish words. However, when it comes to `large` and `largev2` models, there are significant improvements. Although there may still be some instances of incorrect alphabet detection, these models outperform the previously mentioned models in terms of word detection. Regarding translations, the `tiny` and `small` models do not produce meaningful outputs as most of the lines are empty or lack proper translation. The `medium` model, while showing some progress, stops translating after a few translations that are not very accurate and then only focuses on translating the numbers recorded by the Persian speaker at the beginning of each sentence. In the case of the `large` and `largev2` models, there are notable improvements. However, there are still instances where the translated outputs failed to properly incorporate the dislocated item from the Persian source text. This suggests that there are some limitations in capturing the specific linguistic phenomenon of dislocation or understanding the intended meaning behind it.

Another generalisable aspect is the discrepancy between the time stamps reported in the .SRT files and the sound file. The use of linebreaks and commas can probably be generalised across languages but we will need to recode the end of line with or without punctuation symbols to analyze the frequency of the carrier return in transcriptions to investigate how it could be analysed to better understand how prosodic chunking is repressented in the whisper outputs.

5 Conclusion

In this paper, we compared the performance of different Whisper models for the translation task from French into English. We compared these multilingual models trained on multimodal data with SYSTRAN Pure Neural Server translations, generated from the transcribed text and from the Vocapia ASR output, and analysed the different translation outputs. Whisper large models and Vocapia fared better, but for Whisper some translations generated by smaller size models were more accurate for some sentences, including a better containment of the gender bias effect. For translations, the main finding is that the medium model does much better on the transcription task than for the translation task, probably because in our data the translation segment often corresponds to two utterances on the sound files. For the transcription task, the key finding is the apparent anarchic distribution of time stamps across models for the same speech signal.

More research is needed to better understand the time interval (mis)management of the Whisper transcriptions and translations encoded in the SRT file outputs. Should Whisper be used for the translations of subtitles, one may wonder about the absence of pauses in the time stamps. More research is needed to evaluate the potential architectural effect of the training of Whisper on 30s windows of speech.

Author contributions

Nicolas Ballier designed the study, developed the validation procedures with the speech signal and wrote the first draft of the manuscript. Behnoosh Namdarzadeh and Nicolas Ballier designed the test set and Nicolas Ballier recorded it. Maria Zimina-Poirot managed the test settings for Vocapia-SPNS9 translations and conducted textometric analysis of raw translation outputs for all test models. Jean-Baptiste Yunès implemented the JupyterHub, supervised some of the experiments and conducted impact measurements. All authors contributed to the analysis of the outputs.

Acknowledgements

We thank the four anonymous reviewers for their input on a preliminary version of this paper.

This publication has emanated from research supported in part by a 2021 research equipment grant (PAPTAN project)[6] from the Scientific Platforms and Equipment Committee, under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université de Paris).

Nicolas Ballier benefited from a CNRS research leave at LLF (Laboratoire de Linguistique Formelle), which is gratefully acknowledged.

References

Ashby, W. J. (1994). An acoustic profile of right-dislocations in French. *Journal of French Language Studies*, 4(2):127–145.

Avanzi, M. (2012). *L'Interface Prosodie Syntaxe en français : Dislocations, Incises et Asyndètes*. Peter Lang, Bruxelles.

Bapna, A., Cherry, C., Zhang, Y., Jia, Y., Johnson, M., Cheng, Y., Khanuja, S., Riesa, J., and Conneau, A. (2022). mSLAM: Massively multilingual joint pre-training for speech and text.

Besacier, L., Lecouteux, B., Luong, N. Q., Hour, K., and Hadjsalah, M. (2014). Word confidence estimation for speech translation. In *Proceedings of The International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA.

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Blasco-Dulbecco, M. (1999). *Les dislocations en français contemporain. Etude syntaxique*. Honoré Champion, Paris.

Boersma, P. and Weenink, D. (2023). Praat: doing phonetics by computer [computer program]. version 6.3.10. *Retrieved May*, 3:2023.

Branca-Rosoff, S. and Lefeuvre, F. (2016). Le corpus de français parlé parisien des années 2000: Constitution, outils et analyses. le cas des interrogatives indirectes. *Corpus*, 15:265–284.

Buthke, C., Sichel-Bazin, R., and Meisenburg, T. (2010). Sujets disloqués vs. sujets doublés: À la recherche de la frontière prosodique. In *Journées PFC (Phonologie du Français Contemporain) Paris (Des normes à la périphérie). Journées PFC Paris, décembre 2010 :des normes à la périphérie*.

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., et al. (2016). Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

[6] Plateforme pour l'apprentissage profond pour la traduction automatique neuronale, in English: Deep Learning for Machine Translation at Universite Paris Cité. See the description of the platform on the project website: `https://u-paris.fr/plateforme-paptan`

De Jong, N. H. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390.

Dehé, N. and Kavalova, Y. (2007). *Parentheticals*, volume 106. John Benjamins Publishing.

Despres, J., Lamel, L., Gauvain, J.-L., Vieru, B., Woehrling, C., Le, V. B., and Oparin, I. (2013). The vocapia research asr systems for evalita 2011. In *Evaluation of Natural Language and Speech Tools for Italian: International Workshop, EVALITA 2011, Rome, January 24-25, 2012, Revised Selected Papers*, pages 286–294. Springer.

Fagyal, Z. (2002). Prosodic boundaries in the vicinity of utterance-medial parentheticals in French. *Probus*, 14(1):93–111.

Fleury, S. and Zimina, M. (2014). Trameur: A framework for annotated text corpora exploration. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 57–61.

Grosman, J. (2021). Fine-tuned XLSR-53 large model for speech recognition in English. `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english`.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Isabelle, P., Cherry, C., and Foster, G. (2017). A challenge set approach to evaluating machine translation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486—-2496.

Jiao, W., Wang, W., Huang, J.-t., Wang, X., and Tu, Z. (2023). Is ChatGPT a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Lebart, L., Salem, A., and Berry, L. (1997). *Exploring textual data*, volume 4. Springer Science & Business Media.

Lerner, P., Bergoënd, J., Guinaudeau, C., Bredin, H., Maurice, B., Lefevre, S., Bouteiller, M., Berhe, A., Galmant, L., Yin, R., and Barras, C. (2022). Bazinga! A Dataset for Multi-Party Dialogues Structuring. In *13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3434–3441, Marseille, France. European Language Resources Association (ELRA).

Matassoni, M., Brugnara, F., and Gretter, R. (2013). Evalita 2011: Automatic speech recognition large vocabulary transcription. In *Evaluation of Natural Language and Speech Tools for Italian: International Workshop, EVALITA 2011, Rome, January 24-25, 2012, Revised Selected Papers*, pages 274–285. Springer.

Namdarzadeh, B. and Ballier, N. (2022). The neural machine translation of dislocations. *ExLing 2022*, 28:127–131.

Namdarzadeh, B., Ballier, N., Wisniewski, G., Zhu, L., and Yunès, J.-B. (2022). Toward a test set of dislocations in persian for neural machine translation. In *The Third International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2022)*, pages 14–21.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Reddy, A. M., Rose, R. C., and Désilets, A. (2007). Integration of ASR and machine translation models in a document translation task. In *INTERSPEECH*, pages 2457–2460.

Santiago, F., Dutrey, C., and Adda-Decker, M. (2015). Towards a typology of ASR errors via syntax-prosody mapping. In Adda, G., Mititelu, V. B., Mariani, J., and Vasilescu, D. T. . I., editors, *Errors by Humans and Machines in Multimedia, Multimodal and Multilingual Data Processing. Proceedings of ERRARE 2015*, pages 175–192. Editura Academiei Române.

Tancoigne, E., Corbellini, J. P., Deletraz, G., Gayraud, L., Ollinger, S., and Valero, D. (2022). Un mot pour un autre ? Analyse et comparaison de huit plateformes de transcription automatique. *Bulletin de Méthodologie Sociologique / Bulletin of Sociological Methodology*, 155(1):45 – 81.

Tellier, C. and Valois, D. (2006). *Constructions méconnues du français*. Presses Universitaires de Montréal.