Who Are We Talking About?
Handling Person Names in Speech Translation

Anonymous ACL submission

Abstract

Recent work has shown that systems for speech translation (ST) – similarly to automatic speech recognition (ASR) – poorly handle person names. This shortcoming does not only lead to errors that can seriously distort the meaning of the input, but also hinders the adoption of such systems in application scenarios (like computer-assisted interpreting) where the translation of named entities, like person names, is crucial. In this paper, we first analyse the outputs of ASR/ST systems to identify the reasons of failures in person name transcription/translation. Besides the frequency in the training data, we pinpoint the nationality of the referred person as a key factor. We then mitigate the problem by creating multilingual models, and further improve our ST systems by forcing them to jointly generate transcripts and translations, prioritising the former over the latter. Overall, our solutions result in a relative improvement in token-level person name accuracy by 47.8% on average for three language pairs (en→es,fr,it).

1 Introduction

Automatic speech translation (ST) is the task of generating the textual translation of utterances. Research on ST (Anastasopoulos et al., 2021; Bentivogli et al., 2021) has so far focused on comparing the *cascade* (a pipeline of an automatic speech recognition – ASR – and a machine translation – MT – model) and the *direct* paradigms (Bérard et al., 2016; Weiss et al., 2017), or on improving either of them in terms of overall quality. Quality is usually measured with automatic metrics such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006), possibly corroborated by manual analyses. However, the underlying assumption of these efforts is that the generated text is consumed by end users whose goal is understanding the source speech content, disregarding that ST has the potential to be deployed in other application scenarios associated to different user needs.

One possible application is in the context of computer-assisted interpreting (CAI – Fantinuoli 2017a), which supports interpreters during both the preparation phase (Fantinuoli, 2017b; Lim, 2020) and the live interpretation (Prandi, 2018; Desmet et al., 2018). During simultaneous sessions, in fact, interpreters undergo a high cognitive workload in which some elements – namely named entities (NEs) and terminology – are known to play a critical role (Jones, 1998; Gile, 2009). These elements *i)* are hard to remember (Liu et al., 2004), *ii)* can be unknown to interpreters and difficult to recognize (Griffin and Bock, 1998), and *iii)* differently from other types of words, usually have one or few correct translations. As such, interpreters would benefit from automatic systems that reliably recognize and translate these critical elements, without distracting them with wrong suggestions that are even harmful (Stewart et al., 2018). The fluency and intelligibility of the generated translations, instead, plays a marginal role for them, as interpreters are known to be better than machines on these aspects (Fantinuoli and Prandi, 2021).

However, Gaido et al. (2021) recently showed on their newly created benchmark – NEuRoparl-ST – that both ASR models (and thus cascade ST systems) and direct ST systems are currently inadequate to meet these needs. Indeed, they perform poorly on person names, with transcription/translation accuracy of ~40%. Hence, as a first step toward the long-term goal of integrating ST models in assistant tools for live interpreting, this work focuses on *i)* identifying the factors that lead to the wrong transcription and translation of person names, and *ii)* proposing dedicated solutions to mitigate the problem.

To achieve these objectives, our first contribution (§3.1) is the annotation[1] of each person name occurring in NEuRoparl-ST with information about their nationality and the nationality of the speaker (as a

[1] To be released upon paper acceptance.

proxy of the native language) – e.g. if a German person says "*Macron is the French president*", the speaker nationality is German, while the referent nationality is French. Drawing on this additional information, our second contribution (§3.2-3.3) is the analysis of the concurring factors involved in the correct recognition of person names. Besides their frequency, we identify as key discriminating factor the presence in the training data of speech uttered in the referent's native language (e.g. French in the above example). This finding, together with an observed accuracy gap between person name transcription (ASR) and translation (ST), leads to our third contribution (§4): a multilingual ST system that jointly transcribes and translates the input audio, giving higher importance to the transcription task in favour of a more accurate translation of names. Our model shows relative gains in person name translation by 48% on average on three language pairs (en→es,fr,it), producing useful translations for interpreters in 66% of the cases. A manual analysis of the outputs concludes our work (§5), highlighting that most of the errors still produced fall into two categories: omissions and replacements with a different person name. These insights can be the starting point for future work aimed at tackling the identified issues.

2 Related Work

When the source modality is text, person names can often be "copied", i.e. replicated unchanged, into the output. This task has been shown to be well accomplished by both statistical and neural translation systems (Koehn and Knowles, 2017). On the contrary, when the source modality is speech (as in ASR and ST), systems struggle due to the impossibility to copy the audio source. The recognition of person names from speech is a complex task that has mostly been studied in the context of recognizing a name from a pre-defined list, such as phone contacts (Raghavan and Allan, 2005; Suchato et al., 2011; Bruguier et al., 2016). The scenario of an open or undefined set of possible names is instead under-explored. Few studies (Ghannay et al., 2018; Caubrière et al., 2020) focus on comparing end-to-end and cascade approaches in the transcription and recognition of NEs from speech. They do not directly investigate person names though, as they do not disaggregate their results by NE category. Similarly, Porjazovski et al. (2021) explore NE recognition from speech in low-resource languages, and propose two end-to-end methods: one adds a tag after each word in the generated text to define whether it is a NE or not, and one uses a dedicated decoder. However, they do not provide specific insights on the system ability to correctly generate person names and limit their study to ASR, without investigating ST. Closer to our work, Gaido et al. (2021) highlight the difficulty of ASR/ST neural models to transcribe/translate NEs and terminology. Although they identify person names as the hardest NE category by far, they neither analyse the root causes nor propose mitigating solutions.

3 Factors Influencing Name Recognition

As shown in (Gaido et al., 2021), the translation of person names is difficult both for direct and cascade ST systems, which achieve similar accuracy scores (~40%). The low performance of cascade solutions is largely due to errors made by the ASR component, while the MT model usually achieves nearly perfect scores. For this reason, henceforth we will focus on identifying the main issues related to the transcription and translation of person names, respectively in ASR and *direct* ST.

We hypothesize that three main factors influence the ability of a system to transcribe/translate a person name: *i)* its frequency in the training data, as neural models are known to poorly handle rare words, *ii)* the nationality of the referent, as different languages may involve different phoneme-to-grapheme mappings and may contain different sounds, and *iii)* the nationality of the speaker, as non-native speakers typically have different accents and hence different pronunciations of the same name. To validate these hypotheses, we inspect the outputs of Transformer-based (Vaswani et al., 2017) ASR and ST models trained with the configuration defined in (Wang et al., 2020). For the sake of reproducibility, complete details on our experimental settings are provided in the Appendix.[2]

3.1 Data and Annotation

To enable fine-grained evaluations on the three factors we suppose to be influential, we enrich the NEuRoparl-ST benchmark by adding three (one for each factor) features to each token annotated as *PERSON*. These are: *i)* the token frequency in the target transcripts/translations of the training set, *ii)* the nationality of the referent, and *iii)* the

[2]Upon paper acceptance, we will release both the code and the trained models used in our experiments.

nationality of the speaker. The nationality of the referents was manually collected by the authors through online searches. The nationality of the speakers, instead, was automatically extracted from the personal data listed in LinkedEP (Hollink et al., 2017) using the country they represent in the European Parliament.[3] All our systems are trained on Europarl-ST (Iranzo-Sánchez et al., 2020) and MuST-C (Cattoni et al., 2021), and evaluated on this new extended version of NEuRoparl-ST.

3.2 The Role of Frequency

As a first step in our analysis, we automatically check how the three features added to each *PERSON* token correlate with the correct generation of the token itself. Our aim is to understand the importance of these factors and to identify interpretable reasons behind the correct or wrong handling of person names. To this end, we train a classification decision tree (Breiman et al., 1984). Classification trees recursively divide the dataset into two groups, choosing a feature and a threshold that minimize the entropy of the resulting groups with respect to the target label. Their structure makes them easy to interpret (Wu et al., 2008): the first decision (the root of the tree) is the most important criterion according to the learned model, while less discriminative features are pushed to the bottom.

We feed the classifier with 49 features, corresponding to: *i)* the frequency of the token in the training data, *ii)* the one-hot encoding of the speaker nationality, and *iii)* the one-hot encoding of the referent nationality.[4] We then train it to predict whether our ASR model is able to correctly transcribe the token in the output. To this end, we use the implementation of scikit-learn (Pedregosa et al., 2011), setting to 3 the maximum depth of the tree, and using Gini index as entropy measure.

Unsurprisingly, the root node decision is based on the frequency of the token in the training data, with 2.5 as split value. This means that person names occurring at least 3 times in the training data are likely to be correctly handled by the models. Although this threshold may vary across datasets of different size, it is an indication on the necessary number of occurrences of a person name, eventually useful for data augmentation techniques aimed at exposing the system to relevant instances at training time (e.g. names of famous people in the specific domain of a talk to be translated/interpreted). We validate that this finding also holds for ST systems by reporting in Table 1 the accuracy of person tokens for ASR and the three ST language directions, split according to the mentioned threshold of frequency in the training set. On average, names occurring at least 3 times in the training set are correctly generated in slightly more than 50% of the cases, a much larger value compared to those with less than 3 occurrences.

	All	Freq. >= 3	Freq. < 3
ASR	38.46	55.81	4.55
en-fr	28.69	45.45	0.00
en-es	35.29	53.57	19.05
en-it	29.70	46.77	2.56
Average	33.04	50.40	6.54

Table 1: Token-level accuracy of person names divided into two groups according to their frequency in the training set for ASR and ST (en→es/fr/it) systems.

The other nodes of the classification tree contain less interpretable criteria, which can be considered as spurious cues. For instance, at the second level of the tree, a splitting criterion is "*is the speaker from Denmark?*" because the only talk by a Danish speaker contains a mention to *Kolarska-Bobinska* that systems were not able to correctly generate.

We hence decided to perform further dedicated experiments to better understand the role of the the other two factors: referent and speaker nationality.

3.3 The Role of Referent Nationality

Humans often struggle to understand names belonging to languages that are different from their native one or from those they know. Moreover, upon manual inspection of the system outputs, we observed that some names were Englishized (e.g. *Youngsen* instead of *Jensen*). In light of this, we posit that a system trained to recognize English sounds and to learn English phoneme-to-grapheme mappings might be inadequate to handle non-English names.

We first validate this idea by computing the accuracy for names of people from the United Kingdom[5] ('UK" henceforth) and for names of people from the rest of the World ("non-UK"). Looking

[3] For each speech in Europarl-ST, the speaker is referenced by link to LinkedEP.

[4] Speakers and referents respectively belong to 17 and 31 different nations.

[5] We are aware that our annotation is potentially subject to noise, due to the possible presence of UK citizens with non-anglophone names. A thorough study on the best strategies to maximise the accuracy of UK/non-UK label assignment is a task *per se*, out of the scope of this work. By now, as a manual inspection of the names revealed no such cases in our data, we believe that the few possible wrong assignments do not undermine our experiments, nor the reported findings.

Referent	ASR	en-fr	en-es	en-it	Freq.
UK	52.38	59.09	63.16	41.18	46.21
non-UK	35.78	22.00	30.00	27.38	21.96
All	38.46	28.69	35.29	29.70	25.65

Table 2: Token-level accuracy of ASR and ST (en-fr, en-es, en-it) systems for UK/non-UK *referents*.

at Table 2, we notice that our assumption seems to hold for both ASR and ST. However, the scores correlate with the frequency (Freq.) of names in the training set[6] as, on average, UK referents have more than twice the occurrences (46.21) of non-UK referents (21.96). The higher scores for UK referents may hence depend on this second factor.

To disentangle the two factors and isolate the impact of referents' nationality, we create a training set with balanced average frequency for UK and non-UK people by filtering out a subset of the instances containing UK names from the original training set.[3] To ensure that our results are not due to a particular filtering method, we randomly choose the instances to remove and run the experiments on three different filtered training sets. The results for the three ST language pairs and ASR (see Table 3) confirm the presence of a large accuracy gap between UK and non-UK names (9.22 on average), meaning that the accuracy on non-UK names (23.62) is on average ~30% lower than the accuracy on UK names (32.84). As in this case we can rule out any bias in the results due to the frequency in the training set, we can state that the nationality of the referent is an important factor.

	ASR	en-fr	en-es	en-it	Avg.
UK	42.86	25.76	33.33	29.41	32.84
non-UK	29.05	22.67	23.33	19.44	23.62
ΔAccuracy	13.81	3.09	10.00	9.97	9.22

Table 3: Token-level accuracy of UK/non-UK *referents* averaged over three runs with balanced training sets.

3.4 The Role of Speaker Nationality

Another factor likely to influence the correct understanding of person names from speech is the speaker accent. To verify its impact, we follow a similar procedure to that of the previous section. First, we check whether the overall accuracy is higher for names uttered by UK speakers than for those uttered by non-UK speakers. Then, to ascertain whether the results depend on the proportion of UK/non-UK speakers, we randomly create three training sets featuring a balanced average frequency of speakers from the two groups.

[6]Notice that the ASR and the ST training sets mostly contain the same data, so frequencies are similar in the four cases.

Speaker	ASR	en-fr	en-es	en-it	Freq.
UK	41.03	32.43	36.84	29.41	34.55
non-UK	37.36	27.06	34.57	29.85	21.76
All	38.46	28.69	35.29	29.70	25.65

Table 4: Token-level accuracy of ASR and ST systems for names uttered by UK/non-UK *speakers*.

Table 4 shows the overall results split according to the two groups of speaker nationalities. In this case, the accuracy gap is minimal (the maximum gap is 5.37 for en-fr, while it is even negative for en-it), suggesting that the speaker accent has marginal influence, if any, on how ASR and ST systems handle person names.

The experiments on balanced training sets (see Table 5) confirm the above results, with an average accuracy difference of 2.78 for ASR and the three ST language directions. In light of this, we can conclude that, differently from the other two factors, speakers' nationality has negligible effects on ASR/ST performance on person names.

Speaker	ASR	en-fr	en-es	en-it	Avg.
UK	29.91	29.73	28.95	23.53	28.03
non-UK	33.33	22.75	25.51	19.40	25.25
ΔAccuracy	-3.42	6.98	3.43	4.13	2.78

Table 5: Token-level accuracy of person names uttered by UK/non-UK *speakers* averaged over three runs with balanced training sets.

4 Improving Person Name Translation

The previous section has uncovered that only two of the three considered factors actually have a tangible impact: the frequency in the training set, and the referent nationality. The first issue can be tackled either by collecting more data, or by generating synthetic instances (Alves et al., 2020; Zheng et al., 2021). Fine-tuning the model on additional material is usually a viable solution in the use case of assisting interpreters since, during their preparation phase, they have access to various sources of information (Díaz-Galaz et al., 2015), including recordings of previous related sessions. Focusing on the second issue, we hereby explore *i)* the creation of models that are more robust to a wider range of phonetic features and hence to names of different nationalities (§4.1), and *ii)* the design of solutions to close the gap between ASR and ST sys-

	Monolingual				Multilingual				
	ASR	en-fr	en-es	en-it	ASR	en-fr	en-es	en-it	
	WER (↓)	BLEU (↑)			WER (↓)	BLEU (↑)			
Europarl-ST	13.65	32.42	34.11	25.72	13.29	33.92	35.59	26.55	
MuST-C	11.17	32.81	27.18	22.81	11.86	33.34	27.72	23.02	
	Token-level Person Name Accuracy (↑)								Avg. Δ
Overall	38.46	28.69	35.29	29.70	46.15	38.52	44.54	36.63	+8.43
UK	52.38	59.09	63.16	41.18	66.67	59.09	63.16	52.94	+6.51
non-UK	35.78	22.00	30.00	27.38	42.20	34.00	41.00	33.33	+8.84

Table 6: Transcription/translation quality measured respectively with WER and SacreBLEU[7] (Post, 2018) and token-level person name accuracy, both overall and divided into UK/non-UK referents. *Avg.* Δ indicates the difference between multilingual and monolingual systems averaged over the ASR and the three ST directions.

tems attested by previous work (Gaido et al., 2021) and confirmed by our preliminary results shown in Table 1 (§4.2).

4.1 Increasing Robustness to non-UK Referents

As illustrated in §3.3, one cause of failure of our ASR/ST models trained on English audio is the tendency to force every sound to an English-like word, distorting person names from other languages. Consequently, we posit that a multilingual system, trained to recognize and translate speech in different languages, might be more robust and, in turn, achieve better performance on non-English names.

We test this hypothesis by training multilingual ASR and ST models that are fed with audio in different languages, and respectively produce transcripts and translations (into French, Italian, or Spanish in our case). The ST training data (*→es/fr/it) consists of the en→es/fr/it sections of MuST-C and the {nl, de, en, es, fr, it, pl, pt, ro}→es/fr/it sections of Europarl-ST. Notice that, in this scenario, the English source audio constitutes more than 80% of the total training data, as MuST-C is considerably bigger than Europarl-ST and the English speeches in Europarl-ST are about 4 times those in the other languages.[8] For ASR, we use the audio-transcript pairs of the *-it training set defined above. Complete details on our experimental settings are provided in the Appendix.[2]

We analyze the effect of including additional languages both in terms of general quality (measured as WER for ASR, and BLEU for ST) and in terms of person name transcription/translation accuracy. Looking at the first two rows of Table 6, we notice that the improvements in terms of generic translation quality (BLEU) are higher on the Europarl-ST than on the MuST-C test set – most likely because the additional data belongs to the Europarl domain – while in terms of speech recognition (WER) there is a small improvement for Europarl-ST and a small loss for MuST-C. Turning to person names (third line of the table), the gains of the multilingual models (+8.43 accuracy on average) are higher and consistent between ASR and the ST language pairs.

By dividing the person names into the two categories discussed in §3.3 – UK and non-UK referents – we see that results become less consistent across language pairs. On ST into French and Spanish, the accuracy of UK names remains constant, while there are significant gains (respectively +12 and +11) for non-UK names. These results seem to support the intuition that models trained on more languages are able to recognize a wider range of phonetic content and, having learned phoneme-to-grapheme mappings also for other languages, they better handle non-English names. However, the results for ASR and for ST into Italian seemingly contradict our hypothesis, as they show higher improvements for UK names (~11-14) than for non-UK names (~6-7).

We investigate this behavior by further dividing the non-UK group into two sub-categories: the names of referents whose native language is included in the training set ("in-train" henceforth), and those of referents whose native language is not included in the training set ("out-of-train"). For in-train non-UK names, the monolingual ASR accuracy is 33.33 and is outperformed by the multilingual counterpart by 16.66, i.e. by a margin higher than that for UK names (14.29). For the out-of-train names, instead, the gap between the monolingual ASR accuracy (36.71) and the multilingual ASR accuracy (39.24) is marginal (2.5). Similarly, for ST into Italian the in-train group accuracy improves by 8.70 (from 34.78 to 43.48), while the

[7] `BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.5.0`

[8] For instance, in *-fr the training set amounts to 671 hours of audio, 573 (i.e. 83%) having English audio.

Model	WER (↓)	BLEU (↑)			Person Accuracy					
	ASR	en-es	en-fr	en-it	ASR	en-es	en-fr	en-it	ST Avg.	ASR-ST
Base	13.29	35.86	33.99	26.80	46.15	44.54	38.52	36.63	39.90	6.25
Triangle	14.25	37.42	35.44	28.20	42.31	43.70	41.80	41.58	42.36	-0.05
λ_{ASR}=0.8, λ_{ST}=0.2	13.75	36.48	34.85	27.30	47.69	44.54	43.44	50.50	46.16	1.53

Table 7: WER (for ASR), SacreBLEU (for ST), and token-level person name accuracy computed on the NEuRoparl-ST test set. For triangle models, ASR scores are computed on the transcript output of the *-it model, as throughout the paper we evaluate ASR on the English transcript of the en-it section. *ST Avg.* is the the average accuracy on the 3 language pairs (en→es,fr,it) and *ASR-ST* is the difference between the ASR and the average ST accuracy.

out-of-train group accuracy has a smaller gain of 4.92 (from 24.59 to 29.51). These results indicate that adding a language to the training data helps the correct handling of person names belonging to that language, even when translating/transcribing from another language. Further evidence is exposed in §5, where we analyse the errors made by our systems and how their distribution changes between a monolingual and a multilingual one.

4.2 Closing the Gap Between ASR and ST

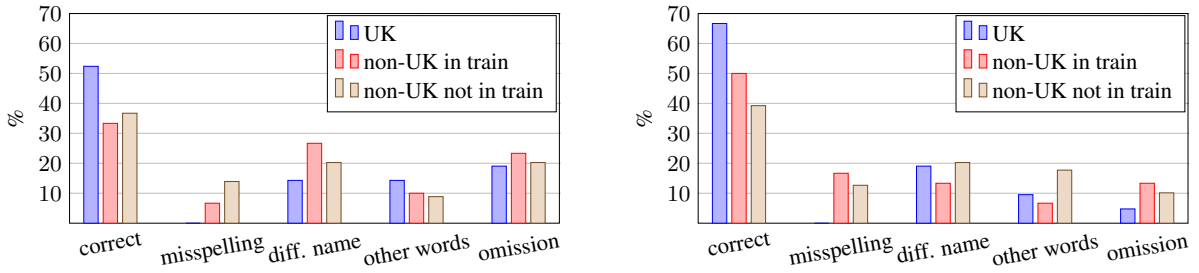
The previous results – in line with those of Gaido et al. (2021) – reveal a gap between ASR and ST systems, although their task is similar when it comes to person names. Indeed, both ASR and ST have to recognize the names from the speech, and produce them as-is in the output. Contextually, Gaido et al. (2021) showed that neural MT models are good at "copying" from the source or, in other words, at estimating $p(Y|T)$ – where Y is the target sentence and T is the textual source sentence – when Y and T are the same string. Hence, we hypothesize that an ST model can close the performance gap with the ASR by conditioning the target prediction not only on the input audio, but also on the generated transcript. Formally, this means estimating $p(Y|X, T')$, where T' denotes a representation of the generated transcript, such as the embeddings used to predict them; and this estimation is what the triangle architecture (Anastasopoulos and Chiang, 2018) actually does.

The triangle model is composed of a single encoder, whose output is attended by two decoders that respectively generate the transcript (ASR decoder) and the translation (ST decoder). The ST decoder also attends to the output embeddings (i.e. the internal representation before the final linear layer mapping to the output vocabulary dimension and softmax) of the ASR decoder in all its layers. In particular, the output of the cross-attention on the encoder output and the cross-attention on the ASR decoder output are concatenated and fed to a linear layer. The model is optimized with a multi-loss objective function, defined as follows:

$$L(X) = -\sum_{x \in X} \Big(\lambda_{ASR} * \sum_{t \in T_x} log(p_\theta(t_i|x, t_{i-1,\dots,0})) + \lambda_{ST} * \sum_{y \in Y_x} log(p_\theta(y_i|x, T, y_{i-1,\dots,0})) \Big)$$

where T is the target transcript, Y is the target translation, and x is the input utterance. λ_{ASR} and λ_{ST} are two hyperparameters aimed at controlling the relative importance of the two tasks. Previous works commonly set them to 0.5, giving equal importance to the two tasks (Anastasopoulos and Chiang, 2018; Sperber et al., 2020). To the best of our knowledge, ours is the first attempt to inspect performance variations in the setting of these two parameters, calibrating them towards the specific needs arising from our application scenario.

In Table 7, we compare the multilingual models introduced in §4.1 with triangle ST multilingual models trained on the same data (second row). Although the transcripts are less accurate (about +1 WER), the translations have higher quality (+1.4-1.6 BLEU on the three language pairs). Person names follow a similar trend: in the transcript the accuracy is lower (-3.84), while in ST it increases (on average +2.46). Interestingly, the accuracy gap between ASR and ST is closed by the triangle model (see the ASR-ST column), confirming our assumption that neural models are good at copying. However, since the accuracy in the transcript is lower (42.31), the ST accuracy (42.36) does not reach that of the base ASR model (46.15). The reason of this drop can be found in the different kind of information required by the ASR and ST tasks. Chuang et al. (2020) showed that the semantic content of the utterance is more important for ST, and that joint ASR/ST training leads the model to focus more on the semantic content of the utterance, yielding BLEU gains at the expense of higher WER. As person names are usually close in the semantic space (Das et al., 2017), the higher



(a) Base ASR errors.

(b) Multilingual ASR errors.

Figure 1: Correct person names and the categories of errors of the baseline and multilingual ASR systems.

focus on semantic content may be detrimental to their correct handling and hence explain the lower person name accuracy.

In light of this observation, we experimented with changing the weights of the losses in the triangle training, assigning higher importance to the ASR loss (third row of Table 7). In this configuration, as expected, transcription quality increases (-0.5 WER) at the expense of translation quality, which decreases (-0.8 BLEU on average) but remains higher than that of the base model. The accuracy of person names follows the trend of transcription quality: the average accuracy on ST (46.16) increases by 3.8 points over the base triangle model (42.36), becoming almost identical to that of the base ASR model (46.15). All in all, our solution achieves the same person name accuracy of an ASR base model without sacrificing translation quality compared to a base ST system.

5 Error Analysis

While the goal is the correct rendering of person names, not all the errors have the same weight. For interpreters, for instance, minor misspellings of a name may not be problematic, an omission can be seen as a lack of help, but the generation of a wrong name is harmful, as potentially distracting and/or confusing. To delve into these aspects, we first carried out a manual analysis on the ASR outputs (§5.1) and then compared the findings with the same analysis on ST outputs (§5.2).

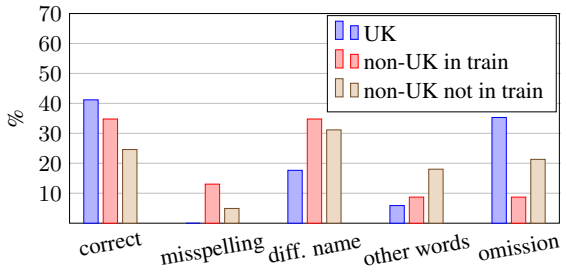
5.1 ASR Analysis

Two authors with at least C1 English knowledge and linguistic background annotated each error assigning it to a category.[9] The categories, chosen by analysing the system outputs, are: **misspelling** – when a person name contains minor errors leading to similar pronunciation (e.g. *Kozulin* instead of *Kazulin*); **replacement with a different name** – when a person name is replaced with a completely different one in terms of spelling and/or pronunciation (e.g. *Mr Muhammadi* instead of *Mr Allister*); **replacement with other words** – when a proper person name is replaced by a common noun, other parts of speech, and/or proper nouns that do not refer to people, such as geographical names (e.g. *English Tibetan core* instead of *Ingrid Betancourt*) **omission** – when a person name, or part of a sentence containing it, is ignored by the system.

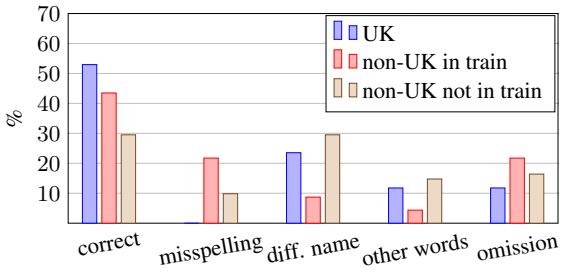
The results of the annotations are summarized in the graphs in Figure 1. Looking at the baseline system (Figure 1a), we notice that omissions and replacements with a different name are the most common errors, closely followed by replacements with other words, although for non-UK names the number of misspellings is also significant. The multilingual system (Figure 1b) does not only show a higher percentage of correct names, but also a different distribution of errors, in particular for the names belonging to the languages added to the training set (non-UK in train). Indeed, the misspellings increase to the detriment of omissions and replacements with a different name and other words. Omissions also decrease for UK names and for names in languages not included in the training set (non-UK not in train). For UK names, the previously-missing names fall either into the correct names or into the replacements with a different name; for the non-UK not in train, instead, they are replaced by different names or other words.

Considering multilingual outputs, we observe that for the languages in the training set (including

[9]The inter-annotator agreement on label assignments was calculated using the *kappa coefficient* in Scott's π formulation (Scott, 1955; Artstein and Poesio, 2008), and resulted in 87.5%, which means "almost perfect" agreement in the standard interpretation (Landis and Koch, 1977).



(a) Base en-it ST errors.



(b) Multilingual ST *-it errors.

Figure 2: Correct person names and the categories of errors of the baseline and multilingual ST-into-Italian systems.

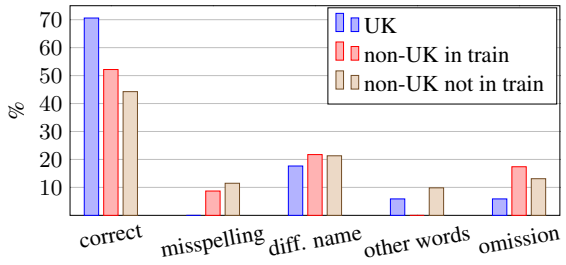


Figure 3: Correct person names and the different categories of errors of the ST-into-Italian triangle system with λ_{ASR}=0.8, λ_{ST}=0.2 expressed in percentages.

English), in 66% of the cases the system generates a name that could be helpful for an interpreter (either correct or with minor misspellings). Confusing/distracting outputs (i.e. replacements with a different person name) occur in about 15% of the cases. Future work should precisely assess whether these scores are sufficient to help interpreters in their job, or which level of accuracy is needed.

Moreover, we notice that the system is able to discern when a person name should be generated (either correct, misspelled, or replaced by a different name) in more than 80% of the cases. This indicates their overall good capability to recognize patterns and/or appropriate contexts in which a person name should occur.

5.2 ST Analysis

The same analysis was carried out for ST systems translating into Italian (see Figure 2) by two native speakers, co-authors of this paper. Although results are lower in general, when moving from the monolingual (Figure 2a) to the multilingual (Figure 2b) system we can see similar trends to ASR, with the number of omissions and replacements with a different name that decreases in favor of a higher number of correct names and misspellings. Looking at the analysis of the triangle model with λ_{ASR}=0.8, λ_{ST}=0.2 presented in §4.2 (Figure 3), we observe that misspellings, omissions, and replacements with other words diminish, while correct names increase. Moreover, both the accuracy (i.e. *correct* in the graphs) and the error distributions of this system are similar to those of the ASR multilingual model (Figure 1b). On one side, this brings to similar conclusions, i.e. ST models can support interpreters in ∼66% of the cases, and can discern when a person name is required in the translation in ∼80% of the cases. On the other, it confirms that the gap with the ASR system is closed, as observed in §4.2.

6 Conclusions

Humans and machines have different strengths and weaknesses. Nonetheless, we have shown that when it comes to person names in speech, they both struggle in handling names in languages they do not know and names that they are not used to hear. This finding seems to insinuate that humans cannot expect help from machines in this regard, but we demonstrated that there is hope, moving the first steps toward ST systems that can better handle person names. Indeed, since machines are faster learners than humans, we can train them on more data and more languages. Moreover, we can design dedicated architectural solutions aimed to add an inductive bias and to improve the ability to handle specific elements. Along this line of research, we have shown that a multilingual ST model, which jointly predicts the transcript and conditions the translation on it, has relative improvements in person name accuracy by 48% on average. We also acknowledge that much work is still needed in this area, with large margin of improvements available, especially to avoid the two most common type of errors pointed out by our analysis: omissions and replacements with different person names.

References

Diego Alves, Askars Salimbajevs, and Mārcis Pinnis. 2020. Data augmentation for pipeline-based speech translation. In *9th International Conference on Human Language Technologies - the Baltic Perspective (Baltic HLT 2020)*, Kaunas, Lithuania.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Antonios Anastasopoulos and David Chiang. 2018. Tied Multitask Learning for Neural Speech Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and regression trees*. Routledge.

Antoine Bruguier, Fuchun Peng, and Françoise Beaufays. 2016. Learning Personalized Pronunciations for Contact Name Recognition. In *Interspeech 2016*, pages 3096–3100.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2020. Where are we in Named Entity Recognition from Speech? In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4514–4520, Marseille, France. European Language Resources Association.

Shun-Po Chuang, Tzu-Wei Sung, Alexander H. Liu, and Hung-yi Lee. 2020. Worse WER, but Better BLEU? Leveraging Word Embedding as Intermediate in Multitask End-to-End Speech Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5998–6003, Online. Association for Computational Linguistics.

Arjun Das, Debasis Ganguly, and Utpal Garain. 2017. Named Entity Recognition with Word Embeddings and Wikipedia Categories for a Low-Resource Language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 16(3).

Bart Desmet, Mieke Vandierendonck, and Bart Defrancq. 2018. Simultaneous interpretation of numbers and the impact of technological support. In Claudio Fantinuoli, editor, *Interpreting and technology*, Translation and Multilingual Natural Language Processing, pages 13–27. Language Science Press.

Stephanie Díaz-Galaz, Presentacion Padilla, and María Teresa Bajo. 2015. The role of advance preparation in simultaneous interpreting: A comparison of professional interpreters and interpreting students. *Interpreting*, 17(1):1–25.

Claudio Fantinuoli. 2017a. *Chapter 7: Computer-assisted Interpreting: Challenges and Future Perspectives*, pages 153–174. Brill, Leiden, The Netherlands.

Claudio Fantinuoli. 2017b. Computer-assisted preparation in conference interpreting. *Translation & Interpreting*, 9:24–37.

Claudio Fantinuoli and Bianca Prandi. 2021. Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 245–254, Bangkok, Thailand (online). Association for Computational Linguistics.

Marco Gaido, Susana Rodríguez, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2021. Is "moby dick" a Whale or a Bird? Named Entities and Terminology in Speech Translation.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity extraction from speech.

Daniel Gile. 2009. *Basic Concepts and Models for Interpreter and Translator Training: Revised edition*. John Benjamins.

Zenzi M. Griffin and Kathryn Bock. 1998. Constraint, Word Frequency, and the Relationship between Lexical Processing Levels in Spoken Word Production. *Journal of Memory and Language*, 38(3):313–338.

Laura Hollink, Astrid van Aggelen, Henri Beunders, Martijn Kleppe, Max Kemman, and Jacco van Ossenbruggen. 2017. Talk of Europe - The debates of the European Parliament as Linked Open Data.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Roderick Jones. 1998. Conference interpreting explained. *Interpreting*, 3(2):201–203.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Lily Lim. 2020. *Interpreting Training in China: Past, Present, and Future*, pages 143–159. Springer Singapore, Singapore.

Minhua Liu, Diane L. Schallert, and Patrick J. Carroll. 2004. Working memory and expertise in simultaneous interpreting. *Interpreting*, 6(1):19–42.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of Interspeech 2019*, pages 2613–2617, Graz, Austria.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Dejan Porjazovski, Juho Leinonen, and Mikko Kurimo. 2021. Attention-Based End-to-End Named Entity Recognition from Speech. In *Text, Speech, and Dialogue*, pages 469–480, Cham. Springer International Publishing.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Bianca Prandi. 2018. An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation.

Hema Raghavan and James Allan. 2005. Matching Inconsistently Spelled Names in Automatic Speech Recognizer Output for Information Retrieval. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 451–458, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

William A. Scott. 1955. Reliability of Content Analysis:The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3):321–325.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,*, pages 223–231, Cambridge. Association for Machine Translation in the Americas.

Matthias Sperber, Hendra Setiawan, Christian Gollan, Udhyakumar Nallasamy, and Matthias Paulik. 2020. Consistent Transcription and Translation of Speech. *Transactions of the Association for Computational Linguistics*, 8:695–709.

Craig Stewart, Nikolai Vogler, Junjie Hu, Jordan Boyd-Graber, and Graham Neubig. 2018. Automatic Estimation of Simultaneous Interpreter Performance. In

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 662–666, Melbourne, Australia. Association for Computational Linguistics.

Atiwong Suchato, Proadpran Punyabukkana, Patanan Ariyakornwijit, and Teerat Namchaisawatwong. 2011. Automatic speech recognition of Thai person names from dynamic name lists. In *The 8th Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand - Conference 2011*, pages 962–966.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, Nevada, United States.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proc. of Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5998–6008, Long Beach, California.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast Speech-to-Text Modeling with Fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proceedings of Interspeech 2017*, pages 2625–2629, Stockholm, Sweden.

Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.

Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. 2021. Using Synthetic Audio to Improve the Recognition of Out-of-Vocabulary Words in End-to-End Asr Systems. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678.

A Experimental Settings

Our ASR and ST models share the same architecture. Two 1D convolutional layers with a Gated Linear Unit non-linearity between them shrink the input sequence over the temporal dimension, having 2 as stride. Then, after adding sinusoidal positional embeddings, the sequence is encoded by 12 Transformer encoder layers, whose output is attended by 6 Transformer decoder layers. We use 512 as Transformer embedding size, 2048 as intermediate dimension of the feed forward networks, and 8 heads. In the case of the triangle model, we keep the same settings and the configurations are the same for the two decoders. The number of parameters is ∼74M for the base system and ∼117M for the triangle model.

We filter out samples whose audio segment lasts more than 30s, extract 80 features from audio segments, normalize them at utterance level, and apply SpecAugment (Park et al., 2019). The target text is segmented into subwords using 8,000 BPE (Sennrich et al., 2016) merge rules with SentencePiece (Kudo and Richardson, 2018).

Models are optimized with Adam (Kingma and Ba, 2015) to minimize the label smoothed cross entropy (Szegedy et al., 2016). The learning rate increases up to 1e-3 for 10,000 warm-up updates, then decreases with an inverse square-root scheduler. We train on 4 K80 GPUs with 12GB of RAM, using mini-batches containing 5,000 tokens, and accumulating the gradient for 16 mini-batches. We average 5 checkpoints around the best on the validation loss. All trainings last ∼4 days for the multilingual systems, and ∼3 days for the base system.