Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 165-176 recibido 03-04-2024 revisado 08-05-2024 aceptado 17-05-2024

Is ASR the right tool for the construction of Spoken Corpus Linguistics in European Spanish?

¿Es el ASR la herramienta adecuada para la construcción de Corpus Lingüísticos Orales en Castellano?

Mirari San Martín, Jónathan Heras, Gadea Mata, Sara Gómez
Universidad de La Rioja
{miren.san-martin, jonathan.heras, gadea.mata, sara.gomezs}@unirioja.es

Abstract: Spoken corpora are a valuable resource to explore naturally occurring discourse. However, large parts of those corpora remain untranscribed due to the high cost of manually transcribing audio files; and, therefore, the access to these resources is limited. This problem could be faced by means of Automatic Speech Recognition (ASR) tools, that have shown their potential to automatically transcribe audio files. In this work, we study two families of ASR models (Whisper and Seamless) for automatically transcribing files from the COSER corpus (that stands for *Corpus Oral y Sonoro del Español Rural*, in English Audible Corpus of Rural Spanish). Our results show that those ASR models can produce accurate transcriptions independently of the dialect of the speakers and their speed-rate; specially with the large v3 version of Whisper that is the model which produces the best results (mean WER of 0.292). However, in some cases the transcriptions do not perfectly align with those produced by humans, since human transcriptors reflect nuances introduced in the speech of speakers that are not captured with the ASR models. This shows that ASR tools can reduce the burden of manually transcribing hours of audios from spoken corpus, but human supervision is still needed.
Keywords: Spoken Corpus, Automatic Speech Recognition, COSER, Dialects, Whisper.

Resumen: Los corpus orales son un recurso muy valioso para explorar el discurso que ocurre de manera natural. Sin embargo, grandes partes de estos corpus permanecen sin transcribir debido al alto coste de transcribir manualmente ficheros de audio; y, por lo tanto, el acceso a estos recursos es limitado. Este problema podría ser abordado mediante herramientas de Reconocimiento Automático del Habla (ASR, por sus siglas en inglés), que han demostrado su potencial para transcribir automáticamente ficheros de audio. En este trabajo, estudiamos dos familias de modelos ASR (Whisper y Seamless) para transcribir automáticamente archivos del corpus COSER (sigla formada a partir de *Corpus Oral y Sonoro del Español Rural*). Nuestros resultados muestran que los modelos de ASR pueden producir transcripciones precisas independientemente del dialecto de los hablantes y su velocidad de habla; especialmente con la versión large v3 de Whisper, que es el modelo que produce los mejores resultados (WER promedio de 0.292). Sin embargo, en algunos casos, las transcripciones no se alinean perfectamente con las producidas por humanos, ya que los transcriptores humanos reflejan matices introducidos por los hablantes que no son capturados con los modelos ASR. Esto muestra que las herramientas ASR pueden reducir la carga de transcribir manualmente horas de audio de los corpus orales, pero aún se necesita supervisión humana.
Palabras clave: Corpus orales, Reconocimiento del habla, COSER, Dialectos, Whisper.

ISSN 1135-5948 DOI 10.26342/2024-73-12

1 Introduction

Corpus linguistics is a discipline devoted to the compilation, annotation and study of written, spoken and multimedia electronic corpora (Kennedy, 2014). This discipline has multiple applications in lexicography, grammar, translation, stylistics, and second language studies (Kennedy, 2014); and, although most well-known corpus are written corpus, there has been a growing interest in spoken corpora since they provide a unique resource to explore naturally occurring discourse (Knight and Adolphs, 2022). Among the existing spoken corpora, we can find many of them in English (Knight et al., 2008) but also in other languages such as Spanish (Fernández-Ordóñez, 2005), Portuguese (Mello, 2014), Korean (Bang et al., 2020), or Arabic (Selouani and Boudraa, 2010). However, large parts of spoken corpus remain untranscribed due to the high cost of manually transcribing audio files (Gorisch, Gref, and Schmidt, 2020); and, therefore, the access to these resources is limited. This drawback can be faced by means of Automatic Speech Recognition (ASR) tools.

In recent years, ASR has undergone a profound transformation fueled by advancements in deep learning techniques (Mehrish et al., 2023). This technology facilitates the conversion of audio signals into text (Yu and Deng, 2016), finding applications in various domains including human-machine interaction such as virtual assistants (Seaborn et al., 2021) and transcription services (Malik et al., 2021), as well as facilitating human-human communication, exemplified by speech-to-speech translation systems for bridging language barriers (Li, Jia, and Chiu, 2023). Additionally, ASR plays a crucial role in addressing accessibility challenges (Pragt et al., 2022). In the context of spoken corpus linguistics, it has been used to curate spoken corpus (Gorisch, Gref, and Schmidt, 2020) and to reduce the burden of manually transcribing hours of recordings (Ramabhadran, Huang, and Picheny, 2003).

As stated in (Orihuela Gracia, 2021), one of the main challenges faced by ASR tools to transcribe spoken corpus is the variability inherent in human speech. This is a well-known drawback of ASR systems (Tatman and Kasten, 2017), particularly when dealing with accents and dialects due to their variations (Forsberg, 2003). Moreover, such a variability is more marked in the case of bilingualism, since people who speak more than one language tend to show more variations in their speech patterns. There are other characteristics that also hinder the work of these systems, such as gender, age, social dialects, speaking styles, and geographic location (O'Shaughnessy, 2008). In addition, ASR tools have been trained with standardized databases, so they will always produce a standard output; and, therefore, they might not capture the nuances and irregularities produced by speakers. While disparities in ASR performance have been extensively studied in English, research in other languages, as Spanish, remains limited (Kantharuban, Vulić, and Korhonen, 2023). Therefore, the aim of this paper is to explore how useful are state-of-the-art ASR systems to build European Spanish spoken corpus that include diverse accents prevalent across various dialects of the European Spanish.

The rest of the paper is organised as follows. In the next section, we present the COSER corpus used in this work — a dataset that comes from recordings of spoken language from rural communities across Spain. Subsequently, we introduce the two families of ASR systems that have been employed to analyse the COSER corpus. In Section 4, we present the results that have been obtained with the ASR systems in the COSER corpus, and discuss the advantages and disadvantages of using ASR tools for the automatic transcription of spoken corpus. We end the paper with some conclusions and further work.

2 The COSER corpus

The COSER corpus (that stands for *Corpus Oral y Sonoro del Español Rural*, in English Audible Corpus of Rural Spanish) (Fernández-Ordóñez, 2005) was created specifically to capture the varieties of rural spoken European Spanish. The corpus includes interviews with elderly rural residents, with an average age of 74.1 years, that were made in the street, not in a lab setting, so some of the interview have background noises. In the recordings, most of the time is the interviewed who is talking (on average an 81.81% of the time with a standard deviation of 7.56); so, we assume that the majority of errors are mainly focused on the

speech of the interviewed. In December 2023, the corpus consisted of 3009 audio recordings, evenly distributed between male and female speakers, totaling 1947 hours of content. Individual recordings vary in duration, ranging from thirty minutes to over two and a half hours, with an average duration of one hour and four minutes per file.

The recordings of the COSER corpus were captured in different Spanish villages, and all Spanish provinces are represented in the corpus. Spain is organized into 17 autonomous communities, with their own institutions and representatives, and certain legislative, executive and administrative powers. Some communities consist of provinces, dating back to the 1833 territorial division of Spain, while others are single-province. We can see this division in Figure 1. Spanish is the official language in the state, but in some communities, there are co-official languages, so bilingualism to varying degrees is common practice. This is the case in Catalonia, Valencia and the Balearic Islands for Catalan; in the Basque Country for Basque, as well as a limited area of Navarre; and in Galicia for Galician. Except for Basque, which is a language isolate, these languages, together with Spanish, are Romance languages. Like in every language, Spanish has its varieties or dialects, which are determined by speaker's geographical background (Shareah, Mudhsh, and AL-Takhayinh, 2015), and the COSER corpus aims to capture them.

Within European Spanish, in monolingual areas, two main dialectal varieties have been traditionally distinguished, Northern-Central or Castilian, and Southern or Andalusian. Some more recent works include a transitional area between these areas. In intonation, there are clear differences between the Spanish dialects, mainly when there is a strong present-day adstrate situation, with bilingualism in another language. Two features create different rhythmic effects: a number of specific contours and durational patterns, specifically the relative duration of pretonic, tonic, and posttonic syllables(Hualde, 2013; Hualde and Prieto, 2015). The rhythm of Basque language and most Romance languages (Catalan, Galician and Spanish) has been reported to be syllable-timed; Portuguese, however, has mixed rhythm (more stress-timed in European varieties and more syllable-timed in Brazilian varieties) (Hualde, 2013; Nazabal, 2021; Frota and Prieto, 2015). Other features separating geographical varieties of Spanish are differences in pronunciation, in lexical items, and in terms of word order, object clitics, and verb tense and mood, in addition to changes due to contact with other languages (Moreno-Fernández and Caravedo, 2022). In this work, we are interested in studying how well state-of-the-art ASR tools deal with this dialectal variation; hence, we have to focus on the part of the COSER corpus that has been transcribed.

Figure 1: Map of Spain with its Autonomous Communities and Provinces.

The COSER corpus not only contains the audio recordings of the interviews; but, for some of them, a manually generated transcription is provided. Namely, as of December 1st, 2023, 227 interviews were available in both formats, but one of them (in particular, a recording from the Lugo province) was not correctly processed; so it was discarded for our study. Without this audio, the transcribed part of the COSER corpus consists of nearly 307 hours of audio content, and the transcriptions contain more than 3 millions words. This dataset of audio-transcription pairs has been used in our work to study how close automatic transcriptions of a given audio are to the manual transcription. Towards that aim, it has been necessary a processing step to be able to compare the manual and the automatically generated transcriptions.

The textual transcripts of the COSER corpus contain, in addition to the spoken content, metadata such as recording location, date, topics covered, and speaker identification for each sentence — an example of one of those transcriptions is provided in Figure 2. Furthermore, the transcriptions are annotated to denote shifts in topics, proper

names, or instances of audio overlap, enhancing their comprehensiveness. While these annotations enrich the transcriptions, they need to be normalised before comparing them with automatically generated transcriptions. Therefore, we have processed the transcriptions as follows.

```
Provincia:      Álava
Enclave:        Menagarai (Ayala), Álava (0109)
Fecha:  26 de marzo de 1993
Duración:       00:51:48
Informantes:    I1: mujer, 51 años
Encuesta:       Alicia Martín, Esperanza Tenorio, Inés Fernández-Ordóñez, Rocío
Transcripción:  María José González Arévalo, Carlota de Benito Moreno
Temas:  1. Matanza del cerdo (chorizo, longaniza, morcilla, jamón, tocino)
2. Alimentación (pan, horno, manteca, huevos, bollos, recetas, pescado)
4. Labores del hogar (agua, fuente, lavadero, jabón, costura, colchones)
7. Ganadería (cabra, oveja, vaca, pastor, leche, queso, lana, abejas, miel)
9. Otros oficios (albañil, modista, etc.)
10. Vida religiosa (cura, iglesia, misa)
11. Vida vecinal (anécdotas, inquietudes, gobierno municipal, guardia civil)
12. Educación (escuela, libros, maestros)
13. Familia (contraste entre generaciones: abuelos, hijos, padres)
14. Bodas y noviazgos (boda, anillos, ajuar)
18. Ejército/Servicio Militar (mili, quintos, guerra)
19. Costumbres (ocio, deporte, baile, juegos infantiles)

E1: |T7| Pues, estábamos impresionados de, de, de lo bonito que era esta comarca
I1: Pues algunos sí tienen, todavía.
E1: [HS:I1 Sí, sí.] Todavía... ¿Qué animales tienen?
I1: Pues aquí se tiene vacas. Vacas, ovejas.
E1: Sí.
I1: Pues, vacas y ovejas, no tienen otra cosa.
```

Figure 2: An example of a transcription of the COSER Corpus.

The first stage of the normalisation involves removing the metadata belonging to the recording, such as "Region", "Enclave" and "Date" from the transcribed text. Subsequently, all annotations within the transcribed text, denoted by markers like "E1:", "[NP]" or "[R-Vhc]", are eliminated. Finally, standard normalisation procedures are applied, including the conversion of all text into lowercase, removal of punctuation marks and abbreviations, and the representation of numbers in their textual form rather than numerical expressions — as an example of this normalisation processs, the normalised version of the text from Figure 2 is presented in Figure 3. Once that the transcriptions have been normalised, they can be compared with the automatic transcriptions generated by ASR systems.

```
pues estábamos impresionados de lo bonito que
era esta comártaca que no conocíamos no
conocíamos vamos de nada que era precioso vamos
a ver nos podrían un poco contar aquí cómo
bueno si la gente sigue teniendo animales pues
algunos sí tienen todavía qué animales tienen
pues aquí se tiene vacas vas o ovejas pues
vacas y ovejas no tienen otra cosa
```

Figure 3: The result of the transcription from Figure 2 after normalisation.

3 ASR models

Nowadays, most ASR models are usually trained on either English-only data or multilingual data; therefore, it is necessary the usage of multilingual models to obtain transcriptions in Spanish. Among the open-source alternatives, we can find models such as Conformer-CTC (Gulati et al., 2020), PocketSphinx (Huggins-Daines et al., 2006) or Wav2Vec (Baevski et al., 2020); however, the state-of-the-art models are based on the Whisper and SeamlessM4T architectures — this claim is based on the ASR leaderboard [1] on December 2023. For our work, we have used different versions of these models provided by the HuggingFace library on an Nvidia GPU Geforce RTX 3080.

Whisper (Radford et al., 2023) is a transformer based model trained on 680,000 hours of multilingual and multitask supervised data collected from the web. The Whisper model process audio by splitting it into 30 second chunks that are converted into log-Mel spectrograms. Whisper is provided in 5 sizes: tiny, base, small, medium, and large; and there are three different versions of the large size model, called large, large-v2, and large-v3. The difference between the large and large-v2 versions is that the latter was trained longer than the former; and, the difference between the large-v2 and large-v3 versions is that the latter was trained with additional data (part of it generated automatically with the large-v2 version). The 8 versions of Whisper can be directly applied for transcribing audios in multiple languages, including Spanish, of any duration. For our study, we have used all versions of Whisper but the large version since there are almost no difference between such a version and the large-v2 version.

Moreover, we have also considered the SeamlessM4T architecture (Barrault et al., 2023). This architecture not only supports speech-to-text translation, but also speech-to-speech translation, text-to-speech translation and text-to-text translation for up to 100 languages, including Spanish. SeamlessM4T was trained with 1 million hours of open speech audio data. In our case, we are using SeamlessM4T in terms of speech-to-text translation in Spanish, and we use the SeamlessM4T v2 large model — the only model of

[1] https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

this family that is available at HuggingFace. On the contrary of Whisper, it is common to have out of memory errors with the SeamlessM4T v2 large model when the length of the audios to be transcribed increases — we detected this problem with audios that last more than a minute. Therefore, we have to split the audio files into one-minute segments, with a 4-second overlap between consecutive segments. In this way, once we obtain the transcription of each segment, we can join them taking into account the overlap of the segments.

Finally, in order to evaluate the transcriptions of the ASR models from the COSER corpus, we have used the *Word Error Rate (WER)* (Woodard and Nelson, 1982), a common metric to measure the performance of ASR systems. WER is defined from the Levenshtein distance (Levenshtein and others, 1966) and works at the word level. Given a reference sentence and an automatically generated sentence, WER is computed using the following formula:

$$WER = \frac{S + D + I}{N}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the reference; therefore, the lower the WER value, the better. It is worth mentioning that to compute the WER value, the reference sentence and an automatically generated sentence are normalised by removing punctuation marks and lower casing the sentences. This should not be necessary in the case of Whisper models that include capital letters and punctuation in their transcription, but that is not the case for the SeamlessM4T models. Moreover, punctuation is an orthographic feature of written text and not from oral transcription; so, it makes sense to work with a normalised version of the text.

4 Result & Discussion

In this section, we analyse the performance of the aforementioned ASR models in the COSER corpus. We have considered three perspectives for our study, two quantitative and the other qualitative. First of all, we consider the overall differences among the ASR models; subsequently, we inspect the performance of the best overall model in the different Spanish provinces; and finally, we conduct a qualitative study to determine what kinds of errors are produced. In the paper, we only include the main statistics of our study, but the interested reader can consult all the conducted experiments in the supplementary materials available at `https://github.com/joheras/SEPLN2024/`.

4.1 Differences among ASR systems

We start by analysing the performance of the studied ASR models on the COSER corpus — we show the mean and standard deviation of each model in Figure 4. As we can see in those results, the mean performance of the models range from 0.81 in the case of the Whisper tiny model, to 0.292 in the case of the Whisper large v3 model. Moreover, we can notice that, as expected, increasing the size of the Whisper model reduces the errors produced by the model. However, there is not a significant difference between the large v2 and large v3; and large v2 and medium versions of Whisper; hence, in this context, the version trained with more data does not provide a significant benefit. Finally, the performance of the Seamless model is only better than the tiny version of Whisper; therefore, this model does not seem a suitable alternative to the family of Whisper models.

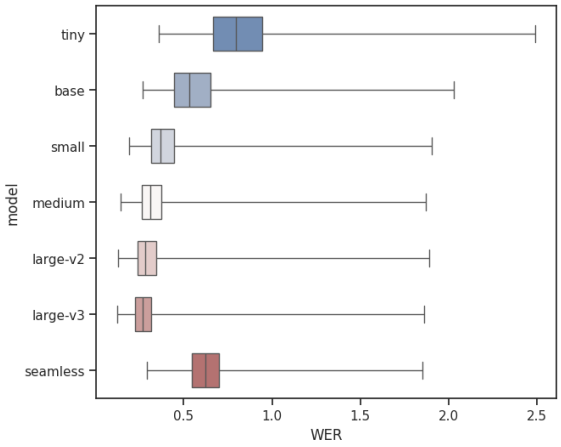


Figure 4: Box and whisker graph that represents how each model behaves for the analyzed audios.

We have also studied how much time does it take to each ASR model to process a 1 minute audio using a GPU NVIDIA GeForce RTX 3080, see Table 1 — note that the two large versions of Whisper take the same time. In the case of the Whisper models, the bigger

Model	Time (secs)
Whisper tiny	4.75
Whisper base	5.32
Whisper small	8.69
Whisper medium	15.25
Whisper large-v2 and -v3	23.89
Seamless	4.36

Table 1: Inference times of the ASR models.

the model, the slower; namely, the tiny version took approximately 4.75 seconds to process the audio, but the large models took almost 24 seconds. It is worth noticing that the Seamless model is the fastest of the analysed ASR systems, even faster than the Whisper tiny model, but as we previously mentioned, its performance is not on par to the bigger models of the Whisper family.

From these results, we can conclude that the model that produces the most accurate transcriptions is the large v3 version of Whisper; however, it is considerably slower than its smaller counterparts. In the context of building the transcriptions of a spoken corpus, processing time is not usually an issue, since the automatic transcription process can be run in the background, and after it finishes, a manual evaluator can fix the errors — therefore, the fewer errors, the better. However, large models might require special hardware to run in a reasonable time; in such cases, the medium version of Whisper provides a good trade-off between accuracy in the transcription and inference speed.

4.2 Differences among Spanish regions

For the second experiment of our study, we are interested in analysing the performance of the ASR models across the regions of Spain (both for provinces and Autonomous Communities). Towards that aim, we have grouped the audio files by province and by Autonomous Community and analysed the performance of the ASR models, see Table 2 for the results per province, and Table 3 for the results per Autonomous Community. From those tables, the first conclusion that we can draw is that the large v3 version of Whisper not only provides the overall best result, but also the best result independently of the Spanish region (both for provinces and Autonomous Communities). Therefore, in the rest of the section, we only present the results for this model, but the interested speaker can see the results of our experiments for the other models in the supplementary materials.

We focus now on the results of the models per province. The intonation of Spanish speakers are very different depending on the Spanish province (Hualde, 2013), so it is natural to wonder whether the models are affected by the speaker's accent. In our analysis, we can observe in Figure 5a that there is not much difference between the WER values obtained across provinces; namely, we observed only significant differences between the provinces of Barcelona and Ourense — p-value of 0.021 in the Large model obtained from the Kruskal-Wallis statistical test, non-parametric method. The two provinces, one in the east and the other in the west of Spain have distinctly different accents and have a different second official language, Catalan in the case of Barcelona and Galician in the case of Ourense. Moreover, in the case of Ourense, its proximity to Portugal could reduce the vowel inventory in final contexts. These may be some of the reasons for this significant difference between these two provinces, which, in the future, we will study in more detail from a linguistic point of view. Another province to consider is Girona, whose WER metric is the second highest. Therefore, paying attention to this province from a linguistic point of view — remember that here Spanish is in contact with Catalan — is advisable to improve our transcription results.

If we focus now on the WER achieved by the large v3 model per Autonomous Community, we can see in Figure 5b that there is a certain degradation from north to south, except in Galicia and Catalonia. One interpretation of these results is the possible influence of the co-official languages of these areas (Galician and Catalan, respectively) when expressing oneself in Spanish. This may be because Galician and Catalan are Romance languages like Spanish. In the case of the Basque Country, this is not the case even though there is a second official language, since it should be noted that Basque is not a Romance language. From the results of the Autonomous Communities, a statistical study has also been conducted, and no significant differences have emerged among them.

We finish this part of the study by wonder-

Province	Num	Tiny	Base	Small	Medium	Large-v2	Large-v3	Seamless
Álava	4	0.72 (0.07)	0.47 (0.05)	0.33 (0.04)	0.27 (0.04)	0.24 (0.04)	0.22 (0.04)	0.67 (0.03)
Albacete	4	0.98 (0.11)	0.73 (0.14)	0.53 (0.09)	0.43 (0.07)	0.41 (0.08)	0.37 (0.06)	0.7 (0.08)
Alicante	4	0.69 (0.05)	0.47 (0.07)	0.33 (0.03)	0.28 (0.05)	0.25 (0.05)	0.24 (0.04)	0.56 (0.04)
Almería	4	0.88 (0.19)	0.64 (0.17)	0.47 (0.15)	0.39 (0.12)	0.36 (0.1)	0.35 (0.11)	0.67 (0.17)
Asturias	5	0.74 (0.18)	0.49 (0.1)	0.35 (0.06)	0.3 (0.05)	0.27 (0.05)	0.26 (0.05)	0.5 (0.06)
Ávila	6	0.82 (0.14)	0.53 (0.1)	0.37 (0.07)	0.31 (0.06)	0.29 (0.05)	0.27 (0.05)	0.69 (0.04)
Badajoz	5	0.68 (0.08)	0.46 (0.06)	0.32 (0.03)	0.27 (0.02)	0.25 (0.02)	0.23 (0.02)	0.58 (0.1)
Baleares	3	0.67 (0.19)	0.48 (0.15)	0.34 (0.09)	0.3 (0.07)	0.28 (0.07)	0.27 (0.06)	0.52 (0.1)
Barcelona	4	0.47 (0.09)	0.33 (0.04)	0.23 (0.02)	0.21 (0.03)	0.19 (0.03)	0.19 (0.02)	0.49 (0.07)
Burgos	5	0.93 (0.08)	0.6 (0.04)	0.42 (0.02)	0.34 (0.02)	0.31 (0.03)	0.29 (0.03)	0.66 (0.07)
Cáceres	5	0.95 (0.1)	0.68 (0.11)	0.48 (0.09)	0.38 (0.08)	0.36 (0.08)	0.33 (0.07)	0.68 (0.08)
Cádiz	4	0.78 (0.11)	0.56 (0.06)	0.41 (0.08)	0.34 (0.07)	0.32 (0.05)	0.3 (0.06)	0.64 (0.1)
Cantabria	5	0.93 (0.11)	0.6 (0.11)	0.38 (0.07)	0.3 (0.06)	0.27 (0.05)	0.25 (0.05)	0.63 (0.11)
Castellón	4	0.74 (0.19)	0.49 (0.11)	0.33 (0.04)	0.28 (0.05)	0.26 (0.04)	0.24 (0.04)	0.5 (0.1)
Ciudad Real	5	0.75 (0.08)	0.5 (0.07)	0.35 (0.04)	0.28 (0.02)	0.26 (0.02)	0.24 (0.02)	0.66 (0.05)
Córdoba	4	0.76 (0.21)	0.57 (0.15)	0.42 (0.12)	0.36 (0.09)	0.33 (0.09)	0.31 (0.08)	0.66 (0.11)
Cuenca	5	0.87 (0.14)	0.56 (0.09)	0.38 (0.07)	0.3 (0.06)	0.27 (0.06)	0.25 (0.05)	0.62 (0.09)
Gerona	4	1.22 (0.85)	0.93 (0.73)	0.78 (0.72)	0.73 (0.72)	0.72 (0.74)	0.7 (0.73)	0.86 (0.53)
Granada	4	0.91 (0.15)	0.63 (0.14)	0.44 (0.08)	0.37 (0.06)	0.34 (0.07)	0.32 (0.05)	0.63 (0.12)
Guadalajara	4	0.81 (0.21)	0.55 (0.16)	0.39 (0.15)	0.31 (0.1)	0.29 (0.1)	0.26 (0.09)	0.62 (0.18)
Guipúzcoa	5	0.85 (0.22)	0.57 (0.16)	0.42 (0.13)	0.36 (0.1)	0.32 (0.1)	0.3 (0.09)	0.65 (0.16)
Huelva	4	0.85 (0.2)	0.59 (0.12)	0.42 (0.07)	0.35 (0.05)	0.32 (0.04)	0.3 (0.04)	0.67 (0.08)
Huesca	4	0.93 (0.07)	0.72 (0.14)	0.53 (0.17)	0.42 (0.12)	0.41 (0.12)	0.37 (0.11)	0.69 (0.05)
Jaén	4	0.82 (0.12)	0.53 (0.08)	0.37 (0.06)	0.3 (0.05)	0.28 (0.04)	0.25 (0.04)	0.6 (0.14)
La Coruña	4	0.98 (0.26)	0.62 (0.16)	0.42 (0.13)	0.36 (0.1)	0.34 (0.09)	0.3 (0.09)	0.55 (0.16)
La Rioja	5	0.77 (0.13)	0.5 (0.08)	0.34 (0.06)	0.28 (0.05)	0.25 (0.05)	0.23 (0.05)	0.62 (0.04)
Las Palmas	5	0.71 (0.14)	0.51 (0.09)	0.4 (0.11)	0.36 (0.12)	0.33 (0.12)	0.33 (0.12)	0.55 (0.09)
León	6	0.85 (0.14)	0.57 (0.12)	0.4 (0.1)	0.31 (0.08)	0.29 (0.09)	0.27 (0.09)	0.73 (0.11)
Lérida	4	0.75 (0.21)	0.5 (0.15)	0.33 (0.07)	0.28 (0.06)	0.26 (0.05)	0.23 (0.05)	0.5 (0.08)
Lugo	3	0.77 (0.27)	0.52 (0.16)	0.37 (0.12)	0.31 (0.11)	0.3 (0.11)	0.28 (0.09)	0.49 (0.17)
Madrid	4	0.75 (0.13)	0.53 (0.14)	0.39 (0.14)	0.34 (0.12)	0.31 (0.12)	0.31 (0.16)	0.65 (0.1)
Málaga	4	0.82 (0.11)	0.59 (0.08)	0.42 (0.06)	0.36 (0.05)	0.34 (0.04)	0.32 (0.04)	0.61 (0.07)
Murcia	5	0.83 (0.17)	0.64 (0.21)	0.52 (0.26)	0.47 (0.28)	0.45 (0.3)	0.44 (0.3)	0.7 (0.14)
Navarra	6	0.83 (0.16)	0.52 (0.11)	0.34 (0.05)	0.28 (0.04)	0.24 (0.03)	0.22 (0.03)	0.58 (0.09)
Orense	4	1.31 (0.63)	1.09 (0.64)	0.94 (0.65)	0.87 (0.67)	0.86 (0.69)	0.84 (0.69)	1.04 (0.55)
Palencia	6	0.7 (0.16)	0.47 (0.12)	0.33 (0.08)	0.29 (0.07)	0.26 (0.05)	0.24 (0.05)	0.63 (0.06)
Pontevedra	1	0.99 (0)	0.72 (0)	0.43 (0)	0.38 (0)	0.37 (0)	0.32 (0)	0.61 (0)
Salamanca	5	0.88 (0.18)	0.56 (0.15)	0.36 (0.09)	0.27 (0.06)	0.25 (0.07)	0.23 (0.06)	0.61 (0.09)
Santa Cruz de Tenerife	5	0.69 (0.06)	0.47 (0.05)	0.32 (0.04)	0.28 (0.03)	0.26 (0.04)	0.25 (0.03)	0.49 (0.1)
Segovia	4	0.52 (0.16)	0.36 (0.08)	0.27 (0.06)	0.24 (0.04)	0.22 (0.04)	0.2 (0.04)	0.57 (0.11)
Sevilla	4	0.92 (0.13)	0.67 (0.09)	0.46 (0.03)	0.39 (0.03)	0.36 (0.03)	0.34 (0.04)	0.7 (0.07)
Soria	4	1.05 (0.13)	0.82 (0.23)	0.53 (0.15)	0.41 (0.11)	0.36 (0.09)	0.33 (0.08)	0.65 (0.16)
Tarragona	4	0.83 (0.08)	0.55 (0.07)	0.39 (0.05)	0.33 (0.05)	0.3 (0.05)	0.29 (0.04)	0.62 (0.09)
Teruel	5	0.68 (0.26)	0.46 (0.17)	0.31 (0.11)	0.26 (0.09)	0.23 (0.1)	0.21 (0.09)	0.57 (0.12)
Toledo	6	0.83 (0.16)	0.58 (0.14)	0.43 (0.12)	0.36 (0.11)	0.33 (0.11)	0.31 (0.1)	0.71 (0.07)
Valencia	6	0.68 (0.17)	0.46 (0.11)	0.33 (0.06)	0.28 (0.04)	0.25 (0.05)	0.24 (0.04)	0.57 (0.04)
Valladolid	6	0.62 (0.14)	0.41 (0.1)	0.29 (0.07)	0.25 (0.05)	0.23 (0.04)	0.21 (0.04)	0.63 (0.08)
Vizcaya	5	0.68 (0.13)	0.46 (0.12)	0.3 (0.06)	0.25 (0.06)	0.24 (0.05)	0.22 (0.05)	0.51 (0.1)
Zamora	5	0.85 (0.17)	0.63 (0.18)	0.42 (0.12)	0.34 (0.09)	0.32 (0.08)	0.29 (0.07)	0.65 (0.15)
Zaragoza	5	0.77 (0.09)	0.53 (0.04)	0.39 (0.04)	0.34 (0.01)	0.31 (0.02)	0.29 (0.02)	0.65 (0.09)
Total	226	0.81 (0.23)	0.56 (0.2)	0.4 (0.18)	0.34 (0.17)	0.31 (0.17)	0.29 (0.17)	0.63 (0.15)

Table 2: WER mean (std) results obtained per each model in the different provinces.

ing whether the speaking rate has an impact on the performance of the ASR models. The study was conducted for all models and splitting the audios by province and Autonomous Community, but we only include here the results for the Whisper large v3 model and splitting the audios per Autonomous Community, see Table 4; the interested reader can consult all the results on the supplementary materials. From those results, we can see that some of the Autonomous Communities with the lowest speed-rates, such as Galicia, has worse error rates than other Automomous Communities with faster speakers, like La Rioja. However, it is not possible to claim that there is a correlation between the speed-rate and the performance of the Whisper large v3 model — Pearson Correlation Coefficient of -0.1891. Therefore, speed rate does not seem to be a factor that influences the performance of ASR models.

As a conclusion of this part of the study, we can claim that ASR models, and in particular the Whisper large v3 model, work

Aut. Community	Tiny	Base	Small	Medium	Large-v2	Large-v3	Seamless
Andalusia	0.84 (0.15)	0.6 (0.11)	0.43 (0.08)	0.36 (0.07)	0.33 (0.06)	0.31 (0.06)	0.64 (0.1)
Aragon	0.78 (0.19)	0.56 (0.16)	0.4 (0.14)	0.33 (0.1)	0.31 (0.11)	0.28 (0.1)	0.63 (0.1)
Asturias	0.74 (0.18)	0.49 (0.1)	0.35 (0.06)	0.3 (0.05)	0.27 (0.05)	0.26 (0.05)	0.5 (0.06)
Balearic Islands	0.67 (0.19)	0.48 (0.15)	0.34 (0.09)	0.3 (0.07)	0.28 (0.07)	0.27 (0.06)	0.52 (0.1)
Basque Country	0.75 (0.16)	0.5 (0.13)	0.35 (0.1)	0.3 (0.09)	0.27 (0.08)	0.25 (0.07)	0.61 (0.13)
Canary Islands	0.7 (0.1)	0.49 (0.07)	0.36 (0.08)	0.32 (0.1)	0.29 (0.09)	0.29 (0.1)	0.52 (0.1)
Cantabria	0.93 (0.11)	0.6 (0.11)	0.38 (0.07)	0.3 (0.06)	0.27 (0.05)	0.25 (0.05)	0.63 (0.11)
Castilla La Mancha	0.84 (0.15)	0.58 (0.13)	0.41 (0.11)	0.33 (0.09)	0.31 (0.09)	0.28 (0.08)	0.66 (0.1)
Castille and Leon	0.8 (0.2)	0.54 (0.17)	0.37 (0.11)	0.31 (0.08)	0.28 (0.07)	0.26 (0.06)	0.65 (0.1)
Catalonia	0.82 (0.48)	0.58 (0.4)	0.43 (0.39)	0.38 (0.38)	0.37 (0.39)	0.35 (0.39)	0.62 (0.29)
Extremadura	0.81 (0.17)	0.57 (0.14)	0.4 (0.1)	0.33 (0.08)	0.3 (0.08)	0.28 (0.07)	0.63 (0.1)
Galicia	1.04 (0.43)	0.76 (0.43)	0.58 (0.44)	0.52 (0.44)	0.51 (0.45)	0.48 (0.45)	0.7 (0.39)
La Rioja	0.77 (0.13)	0.5 (0.08)	0.34 (0.06)	0.28 (0.05)	0.25 (0.05)	0.23 (0.05)	0.62 (0.04)
Madrid	0.75 (0.13)	0.53 (0.14)	0.39 (0.14)	0.34 (0.12)	0.31 (0.12)	0.31 (0.16)	0.65 (0.1)
Murcia	0.83 (0.17)	0.64 (0.21)	0.52 (0.26)	0.47 (0.28)	0.45 (0.3)	0.44 (0.3)	0.7 (0.14)
Navarre	0.83 (0.16)	0.52 (0.11)	0.34 (0.05)	0.28 (0.04)	0.24 (0.03)	0.22 (0.03)	0.58 (0.09)
Valencian Community	0.7 (0.14)	0.47 (0.09)	0.33 (0.04)	0.28 (0.04)	0.25 (0.04)	0.24 (0.04)	0.55 (0.07)
Total	0.81 (0.23)	0.56 (0.2)	0.4 (0.18)	0.34 (0.17)	0.31 (0.17)	0.29 (0.17)	0.63 (0.15)

Table 3: WER mean (std) results obtained per each model in the different Autonomous Communities.

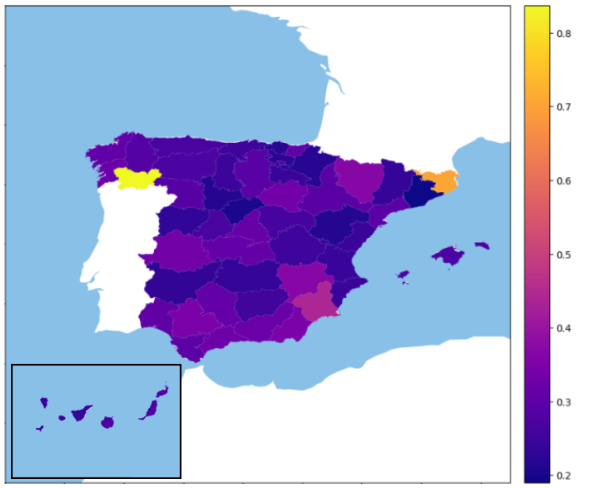
Autonomous Community	# audios	avg (std) words	avg (std) time (sec)	avg (std) speed	avg (std) WER large v3
Castilla La Mancha	24	14171.0 (4742.7)	4702.7 (1431.4)	3.012 (0.456)	0.283 (0.080)
Madrid	4	17048.2 (5325.2)	5813.9 (2391.5)	3.011 (0.283)	0.313 (0.156)
La Rioja	5	14497.0 (10504.7)	4963.0 (3786.9)	2.951 (0.174)	0.233 (0.051)
Castille and Leon	47	12148.0 (4501.4)	4211.3 (1636.0)	2.930 (0.390)	0.258 (0.064)
Aragon	14	12540.5 (5044.5)	4383.8 (1778.7)	2.923 (0.370)	0.285 (0.097)
Murcia	5	18619.6 (6826.1)	6506.4 (2514.0)	2.882 (0.196)	0.438 (0.303)
Extremadura	10	15876.3 (8531.9)	5415.2 (2579.6)	2.876 (0.367)	0.282 (0.071)
Andalusia	32	15295.1 (5023.2)	5383.6 (1905.5)	2.874 (0.378)	0.311 (0.062)
Basque Country	14	11320.2 (4153.2)	4124.0 (1703.7)	2.821 (0.489)	0.249 (0.070)
Cantabria	5	9607.6 (3341.8)	3559.0 (1059.8)	2.695 (0.297)	0.247 (0.051)
Valencian Community	14	15119.2 (6635.9)	5657.25 (2132.0)	2.626 (0.335)	0.241 (0.038)
Navarre	6	8581.3 (3070.5)	3340.2 (1211.2)	2.620 (0.418)	0.219 (0.027)
Canary Islands	10	16044.9 (3897.1)	6320.4 (1725.9)	2.578 (0.310)	0.287 (0.096)
Balearic Islands	3	17390.3 (9076.6)	6819.4 (2506.1)	2.466 (0.364)	0.266 (0.061)
Catalonia	16	12882.9 (5581.5)	5183.5 (1735.6)	2.459 (0.529)	0.354 (0.392)
Galicia	12	11351.6 (5369.5)	4821.0 (1815.0)	2.336 (0.643)	0.477 (0.451)
Asturias	5	11822.2 (4388.7)	5354.5 (2132.3)	2.231 (0.238)	0.264 (0.046)

Table 4: Summary of speed-rate per Autonomous Community.

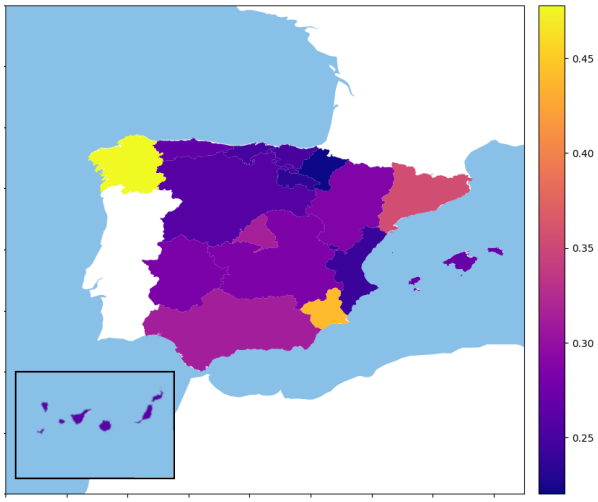
equally well independently of the region of Spain where the audio was captured and the speed-rate of speakers. These results might serve to demystify some bias related to the accents and dialects of some regions of Spain; for instance, that is more difficult to understand Andalusian people since they speak faster than in the rest of Spain; or that people from regions with a co-official language introduce more errors in their discourse. However, more research in that direction is needed since the sample size of our study is relatively small.

4.3 Qualitative results

We finish our study with a more fine-grained analysis of the errors introduced by the ASR tools. In particular, we focus on the errors introduced by the Whisper large v3 model since it is the ASR model that produced transcriptions with less errors. As we have explained in Section 3, the WER metric takes into account three kinds of errors: substitutions, insertions and deletions; and the three of them appear in the generated transcriptions, see Table 5. The main kind of error introduced by the Whisper large v3 model are deletions (that is, some words of the original transcription are not included in the automatically generated text), followed by substitutions (that is, some words of the original transcription are replaced by others), and the least common errors are insertions (that is, the model introduces some words that are not in the original manuscript). These results are also true for most Autonomous Communities



(a) Mean WER per Spanish province.



(b) Mean WER per Spanish Aut. Community.

Figure 5: Distribution of WER for the results of the Whisper Large v3 model

Autonomous Community	% del/error	% ins/error	% subs/error
Andalusia	0.380	0.118	0.502
Aragon	0.465	0.139	0.395
Asturias	0.373	0.181	0.446
Balearic Islands	0.518	0.143	0.339
Basque Country	0.505	0.133	0.362
Canary Islands	0.477	0.125	0.398
Cantabria	0.367	0.207	0.426
Castilla La Mancha	0.398	0.149	0.453
Castille and Leon	0.475	0.140	0.386
Catalonia	0.447	0.238	0.315
Extremadura	0.386	0.128	0.487
Galicia	0.365	0.347	0.288
La Rioja	0.516	0.122	0.362
Madrid	0.589	0.072	0.339
Murcia	0.580	0.056	0.364
Navarre	0.480	0.104	0.415
Valencian Community	0.483	0.155	0.362
Total	0.459	0.150	0.390

Table 5: Percentages of the kinds of errors produced b the Whisper large v3 model grouped by Autonomous Community.

but Andalusia, Asturias, Cantabria, Castilla la Mancha, and Extremadura where substitution errors are more common than deletions. Hence, the first conclusion that we can draw is that the Whisper model does not usually hallucinate by introducing new words in the transcriptions, but has other kinds of errors.

To understand the errors produced by the Whisper large v3 model, we consider an extract of a transcription where all errors are represented, see Table 6. A first kind of deletion error that has been detected occurs when the Whisper model removes words that are repeated in a row; for instance, "*pero, pero*" (but, but) is transcribed just as "*pero*" (but)). Another deletion error is related to taglines (such as "*ummm*" or "*ehh*") since Whisper ignores them. The last kind of deletion error that has been detected comes from words that speakers start to say but are not finished (for example "*co-*"). This shows that the Whisper model is not perfectly aligned with the aim of the COSER corpus, since the model tries to construct sentences that are grammatically correct, but the human transcriptors of the COSER corpus try to reflect how people speak with all their nuances.

A similar situation arises with substitution errors since the speaker might present different mismatches when speaking, and the model corrects them. Examples of those substitutions are contractions ("*pa*" and "*to*" are transcribed as "*para*" and "*todo*" respectively), words where the speaker omitted a sound ("*matao*" is transcribed as "*matado*"), words with the wrong gender ("*primer*" is transcribed as "*primera*" since the speaker was taking about a feminine term), or verbs with the correct agreement ("*estaba*" is transcribed as "*estaban*" when the speaker is talking in plural). Hence, the model is not only transcribing the dialogues of speakers but also correcting their non-normative pronunciations and constructions. We can say that this is a posh transcription that might not reflect how people actually speak, and that is the actual aim of a corpus like COSER.

Finally, the insertion errors that have been detected are either voices in the background (that are transcribed by Whisper but are not included in the manual transcription), and long words that might not appear in the dictionary since they refer, for instance, to locations (an example is the word "*Calamocha*"

Human transcription
Pero siempre he **matao**, siempre. Hasta cinco o seis he **matao** y me iba yo a **Calamocha** a vender los lomos y los costillares y los magros y todo eso y me quedaba lo demás. Y con eso me defendía **pa** los pagos de casa. Porque mi marido de dos añicos se quedó sin el padre y de seis, sin la madre. Y tuvieron que ir a parar a tíos carnales y, claro, se casó joven él porque yo le llevo casi dos años, un año y ocho meses le llevo. Y, y desde **primer** hora pedazar mucho y todo eso porque entonces no había los **haberes** de agora. Y nos gobernábamos a lo mejor con vender lo bueno del tocino y **que la luz del** tocino **pa** comerlo **en**, en casa. Que por eso estaba bueno **to** que hacíamos, **pero**, pero mi vida sacrificada siempre. E1: ¿Cómo se mata un cerdo? I1: Pues **co-, se**, se cuida todo el año. E1: ¿Un año entero? I1: El año entero. Ya antes el año entero, **agora** a lo mejor les echan harinas compuestas y los crían antes, pero como lo natural no hay nada.
Transcription of the model Large-v3 (Whisper)
pero siempre he **matado** siempre hasta cinco o seis he **matado** y me iba yo **acá a la noche** a vender los lomos y los costillares y los magros y todo eso y me quedaba lo demás y con eso me defendía **para** los pagos de casa porque mi marido de dos añicos se quedó sin el padre y de seis sin la madre y tuvieron que ir a parar a tíos carnales y claro se casó joven él porque yo le llevo casi dos años un año y ocho meses le llevo y desde **primera** hora a pedazar mucho y todo eso porque entonces no había los **abuelos** de agora y nos gobernábamos a lo mejor con vender lo bueno del tocino y **quedarnos el** tocino **para** comerlo en casa que por eso estaba bueno **todo lo** que hacíamos pero mi vida sacrificada siempre cómo se mata un cerdo pues se cuida todo el año un año entero el año entero ya antes el año entero **ahora** a lo mejor les echan harinas compuestas y los crían antes pero como lo natural no hay nada

Table 6: **Above.** Example of text written by a human transcriber from audio COSER_4117-01. **Below.** text written by the Whisper model for the same model. The errors are marked with colors according to their type: substitutions in bold, deletions in red, and insertions in blue.

that is transcribed as "*acá a la noche*"). For the former kind of insertion error, Whisper is transcribing the whole audio, but the manual transcription might focus only on the main speakers and ignore background sounds and noises. For the latter, it seems that Whisper split long unknown words into sensible tokens that might form words.

5 Conclusions and further work

In this paper, we have analysed how ASR models can be used to facilitate the task of transcribing audios from European Spanish spoken corpus. Our results show that those models can produce mostly accurate transcriptions independently of the dialect of the speakers and their speed-rate; specially with the large v3 version of Whisper that is the model which produces the best results. However, in some cases the transcriptions do not perfectly align with those produced by humans, since human transcriptors reflect nuances introduced in the speech of speakers that are not captured with the ASR models. This shows that ASR tools can reduce the burden of manually transcribing hours of audios, but human supervision is still needed.

In future work, we are interested in providing transcriptions that align better with the actual speech of speakers, and this might require to fine-tune models like Whisper or Seamless with corpus like COSER. Additionally, up to now, we have not distinguished the different speakers that talk in the audios, and this might be solved by diarisation techniques — this will provide a more fine-grained analysis of the errors. Moreover, the number of hours per province is quite unbalanced and, in some cases, relatively small; so, it would be necessary to collect more data to draw conclusions at the province level. Finally, we have only focused on the COSER corpus but there are other spoken corpus in Spanish that investigate issues such as contact languages (see the COREC corpus) or the Spanish that is talked around the world (see the PRESEEA corpus); and, therefore, it is worth studying how ASR systems behave in those contexts.

Acknowledgments

This work was supported by Ministerio de Ciencia e Innovación [PID2020-115225RB-I00 / AEI / 10.13039/501100011033], and by the Government of La Rioja through Proyecto Inicia 2023/01.

References

Baevski, A., Y. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Bang, J.-U., S. Yun, S.-H. Kim, M.-Y. Choi, M.-K. Lee, Y.-J. Kim, D.-H. Kim, J. Park, Y.-J. Lee, and S.-H. Kim. 2020. Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*, 10(19):6936.

Barrault, L., Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Fernández-Ordóñez, I. 2005. Coser. corpus oral y sonoro del español rural.

Forsberg, M. 2003. Why is speech recognition difficult. *Chalmers University of Technology*.

Frota, S. and P. Prieto. 2015. *Intonation in Romance: Systemic similarities and differences*. Oxford University Press.

Gorisch, J., M. Gref, and T. Schmidt. 2020. Using automatic speech recognition in spoken corpus curation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6423–6428.

Gulati, A., J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Hualde, J. I. 2013. *Los sonidos del español: Spanish Language edition*. Cambridge University Press.

Hualde, J. I. and P. Prieto. 2015. Intonational variation in spanish: European and american varieties. In *Intonation in romance*. Oxford University Press.

Huggins-Daines, D., M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *2006 IEEE international conference on acoustics speech and signal processing proceedings*, volume 1, pages I–I. IEEE.

Kantharuban, A., I. Vulić, and A. Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore, December. Association for Computational Linguistics.

Kennedy, G. 2014. *An introduction to corpus linguistics*. Routledge.

Knight, D. and S. Adolphs. 2022. Building a spoken corpus: What are the basics? In *The Routledge Handbook of Corpus Linguistics*. Routledge, pages 21–34.

Knight, D., S. Adolphs, P. Tennent, and R. Carter. 2008. The nottingham multimodal corpus: A demonstration. In *Programme of the Workshop on Multimodal Corpora*, page 64.

Levenshtein, V. I. et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10(8), pages 707–710. Soviet Union.

Li, X., Y. Jia, and C.-C. Chiu. 2023. Textless direct speech-to-speech translation with discrete speech representation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Malik, M., M. K. Malik, K. Mehmood, and I. Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457.

Mehrish, A., N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria. 2023. A review of deep learning techniques for speech processing. *Information Fusion*, page 101869.

Mello, H. 2014. What corpus linguistics can offer contact linguistics: the c-oral-brasil corpus experience. *PAPIA: Revista Brasileira de Estudos do Contato Linguístico*, pages 407–427.

Moreno-Fernández, F. and R. Caravedo. 2022. Dialectología hispánica the routledge handbook of spanish dialectology.

Nazabal, O. J. 2021. Euskararen erritmoa neurtzen. *Fontes linguae vasconum: Studia et documenta*, 53(132):257–278.

Orihuela Gracia, S. 2021. *Del lenguaje oral al lenguaje escrito: la transcripción como documento de archivo*. Ph.D. thesis, Universitat Autònoma de Barcelona.

O'Shaughnessy, D. 2008. Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10):2965–2979.

Pragt, L., P. van Hengel, D. Grob, and J.-W. A. Wasmann. 2022. Preliminary evaluation of automated speech recognition apps for the hearing impaired and deaf. *Frontiers in Digital Health*, 4:806076.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Ramabhadran, B., J. Huang, and M. Picheny. 2003. Towards automatic transcription of large spoken archives-english asr for the malach project. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Seaborn, K., N. P. Miyake, P. Pennefather, and M. Otake-Matsuura. 2021. Voice in human–agent interaction: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–43.

Selouani, S. A. and M. Boudraa. 2010. Algerian arabic speech database (algasd): corpus design and automatic speech recognition application. *Arabian Journal for Science and Engineering*, 35(2):157–166.

Shareah, M., B. Mudhsh, and A. H. AL-Takhayinh. 2015. An overview on dialectal variation. *International Journal of Scientific and Research Publications*, 5(6):1–5.

Tatman, R. and C. Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Interspeech*, pages 934–938.

Woodard, J. and J. Nelson. 1982. An information theoretic measure of speech recognition performance. In *Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA*.

Yu, D. and L. Deng. 2016. *Automatic speech recognition*, volume 1. Springer.