Challenges in Automatic Speech Recognition in the Research on Multilingualism

Edyta Jurkiewicz-Rohrbacher[1,2], Thomas Asselborn[3]

[1] *University of Hamburg, Institut für Slavistik, Von-Melle-Park 6, 20146 Hamburg, Germany*

[2] *Universität Regensburg, Institut für Slavistik, Universitätsstraße 31, 93053 Regensburg, Germany*

[3] *University of Hamburg, Institute of Humanities-Centered Artificial Intelligence (CHAI), Warburgstraße 28, 20354 Hamburg, Germany*

Abstract

This paper explores the potential of using Large Language Models in multilingualism research to accelerate the management and processing of spoken data. The speech-to-text processing of utterances by multilingual speakers are in the focus. Qualitative discussion of the main issues relating to the non-standard language use of bilingual individuals is provided, using Polish-German recordings from the LangGener corpus as an example.

Keywords

Multilingualism, ASR, Transcription, Polish, German,

1. Introduction

The main challenge in multilingualism research is collecting sufficient data from a homogeneous sample of multilingual speakers to achieve robust results. Perfectly balanced bilingualism in speech and writing is extremely rare. Most speakers have significant discrepancies in their writing skills across different languages, often because they were not educated in one of them. Consequently, multilingualism research, which is necessarily usage-based, relies mostly on spoken data. Managing spoken data is considerably more challenging than managing written data, which is more consistent, even in less formal varieties, and therefore easier to parse and annotate. To perform any kind of quantitative analysis, spoken data must first be transcribed into digital text for further processing and annotation.

Recent developments in large language models for Automatic Speech Recognition (ASR), in particular the emergence of Whisper [1], have clearly accelerated the transcription process. The business sector is the biggest beneficiary, as any kind of recording can be used for quick documentation of a meeting, compilation of notes, protocols or presentations.

From the perspective of multilingualism research, however, the situation looks different due to the different expectations of the quality of automatic transcription from business and research. This paper provides an overview of the issues that need to be addressed to enable linguists to use LLM-enhanced ASR efficiently. We describe the targeted standard of transcript in Section 2. We use Polish-German bilingualism as an example and data from the LangGener project [2] (see Section 3.2). The preliminary research presented here is based on three versions of Whisper[1]: small, medium, large, because it can be run locally (see Section 6 for ethical reasons) and is evaluated as best performing among the available transcription tools (see [3] for comparison). Section 4 describes the generally known problems related to the quality of transcripts, while Section 5 presents the issues specific to multilingualism.

2. Targeted Features of Transcript

In terms of transcripts, the main difference between business and multilingual studies is that business-oriented transcripts should ideally be clean and monolingual to ensure fluent reading. In other words, the focus is on the content's informativeness. Therefore, an ASR should automatically translate foreign elements into the company's language — English in many cases — even if the meeting was multilingual. Elements of speech that do not carry essential information, such as repetitions, disfluencies and hesitation markers, should be removed. In other words, a business transcript will resemble standard written language more closely than the actual oral communication.

Transcripts used for research purposes are different in nature. In linguistic investigations, the recording is the primary object of research and transcripts approximate all the sounds that speakers make, including hesitation markers and self-corrections, to show how and when oral communication flows.

The standards for transcripts vary in academic research, depending on the language being documented and the purpose of the research. Underresourced languages without a written tradition or standard orthographic rules are usually transcribed using phonetic transcription systems (for overview, see [4]). Most European languages can be transcribed using simpler systems that have been developed for analysing spoken discourse (for overview, see [5]). Such systems are mostly based on the orthography of the standard language. They vary in terms of the notation used for spoken language phenomena and issues related to transcript quality, as well as the number of annotation layers. The latter approach is preferable for further automatic language processing, such as lemmatisation and morphosyntactic tagging, as well as for automatic queries, which can be conducted according to standard forms. Further annotation layers, such as phonetic transcriptions, can be added afterwards to differentiate between different non-standard pronunciations, which are typical of dialects, heritage, and non-native varieties. In summary, a transcript for research purposes should be a precise representation of speech that is simple enough to be easily searched and processed further.

3. Experiment Setup

The following section focuses on the technical side of this article. First, a brief introduction into the Whisper models used is provided together with the configuration parameters. Afterwards, a brief description of the dataset is given.

3.1. Whisper Models

For this case study, we have decided to use the standard Whisper models[1]. The latest versions as described on the Whisper GitHub page of both the library and the versions of the models were employed[1]. For this first case study, our goal was to have a base case estimate of how the basic Whisper models will perform. Thus, we have decided to use the auto language detection modes for all the experiments. We have used three different model sizes to find the tradeoffs between the model size and qualitative results of the transcription. The three models used were

- Whisper base with 74 million parameters,
- Whisper medium with 769 million parameters and
- Whisper large with 1550 million parameters.

More information can be found on the model card on GitHub[2] All Whisper models were used in the multilingual versions.

3.2. Corpus LangGener as Source of Data

The LangGener Corpus [6] contains recordings of language biography interviews with Polish–German bilinguals. The sample is stratified across two generations: an older generation who lived in Poland in areas that were part of the German Reich before 1945 (called Generation Poland), and late bilinguals who were born in these areas and immigrated to Germany (nowadays known as 'Aussiedler' or in the project called Generation Germany).

This stratification principle makes the sample interesting for ASR since their speech contains various features of non-standard language, including dialectal features and non-native pronunciation.

Structurally, the corpus contains many phenomena related to multilingualism [7] such as code-switching, lexical matter and pattern replication. We describe them briefly in Section 5. For a full overview of these features in LangGener, see [8].

4. Previously Identified Problems with LLM-enhanced Transcription

One serious flaw that applies to many LLM-enhanced tools is hallucinations. In the case of ASR, this means transcriptions of text which cannot be aligned with the audio file, observed by [9] for Whisper. Although open.ai claimed improving this issue,[3] we still identify hallucinations in the studied data, as shown in (1):

(1) To jest pożyteczna rozmowa. Dowiemy się trochę więcej o sobie.[4] (Added in transcription with Whisper Large:) **Tak. Tak. To było bardzo miłe, ale bardzo miłe. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak. Tak.**
'That's a useful chat. We are learning more about ourselves. (Added in transcription with Whisper Large:) Yes. Yes. This is very nice, but really nice. Yes. Yes. Yes. Yes. Yes. Yes. Yes. Yes. Yes. Yes.'

In their review of ASR for English and German, [3] identify several types of mistake that are probably generally valid. These are:

- wrong transcription due to similar-sounding words,
- misunderstood proper nouns,
- omitting single words and sentences,
- missing sub-sentences,
- assigning wrong endings,
- spelling mistakes.

The Whisper transcription of the LangGener corpus contains such mistakes too. In particular, it omits disfluencies, broken words, repetitions and other features of conversational data. We find issues relating to incorrect endings, spelling mistakes and homophone structures that are particularly relevant for multilingual data. We therefore return to them in the next section.

5. Transcription Issues for Research on Multilingualism

5.1. Tendency to codes' unification

Problems strictly related to multilingual transcriptions have received little investigation to date. We are aware of [10], who fine-tuned Whisper for English-Chinese (i.e. biscriptual) transcription. However, this work does not provide an overview of the problems associated with such transcriptions; rather, it contains a technical description of the fine-tuned model.

---

[3]https://www.science.org/content/article/ai-transcription-tools-hallucinate-too
[4]https://langgener.ijppan.pl/OUT/KC_HAM_II_GD_PL04-149945-153750.wav

אוד דע מן קליין, דע בנדיק דע וואל דע מאז, איש קוצ'אי אושי, ממה, אוב איידאי
וטטס דא אום צ'אי סטאי, אום איש אירי איימוסטאי, איליני בילד, אין איליני, אורטטאו
מטעין קינטאו. זי חד מי דעס גנונג. אייקא, אייקא, בנאו. אייך וואיס ניץ? דא סטאוואם
איסטם ג'צקי. יא אורטטאו מטעין קליין וואי. מי נא קליין, עטטאו דע נאיין, דע בנדדיק
־ע מאז. יצ קם איס צאו לוציאם, דע קם אין קליגא, אורטטאו, דע אורטטאו מפה אופל
אינאנט, ונדע, איינא, פרק מי דע אין אמותא, פרק, סאך מי דע. יא קוויסטו דען וסי,
וואס וילסטו פון מי וסי? אייך, ופי פייי אבי ג'לט גיימד, דע איש דעפטי צאו בזאו לצאו
Pyro וסא מהן ןכל...5לוציאם. ילא קורא צאו

Figure 1: Transcript of speaker BQ RAC recognized as Yiddish by Whisper Large

When studying multilingualism, it is important to avoid translations in transcripts known from business solutions, since the focus of research is precisely on phenomena related to the mixing of two language codes. Typically, one code dominates (the matrix code), while the other is used only occasionally and is embedded within the matrix code.

Whisper does adjust the text to one language relatively rarely, but we find some problematic passages, for example (2):

(2) a. diese Rechtsanspruch to # jest każde dziecko ma Recht da drauf[5]

   b. Tak, tak. ...na szpruch. To jest... każde dziecko ma prawo na szpruch [Whisper Large]

The word *Rechtsanspruch* 'legal right' is not recognized as the embedded German code, and its second part is transcribed with Polish orthography as Polish non-existing word. /ʃ/ is represented as <sz> instead of German spelling requiring <s> at the beginning of word before <k>, <p> or <t>. Further, the phrase *Recht da drauf* 'right to it' is not transcribed but interpreted, and rendered in Polish consistently with the previous mistake, thus, as *prawo na szpruch* 'right for szpruch'.

Preferring to unify with the matrix code in the output can also lead to words being replaced with similar-sounding words from the matrix code, or entirely new linguistic units being formed according to word-building rules of the matrix code. Thus, in example (3) German word *schicken* 'here: they send' is represented as Polish *szczypią* 'they pitch', while *pressen* 'they squeeze' is represented as the made-up verb *presują*. *In eine Gruppe* receives a direct translation. Note that the German forms are ambiguous in terms of person and number, which may explain partially why switching to German in the transcription is avoided.

(3) a. i oni coraz więcej tych dzieci # **schicken in eine Gruppe pressen in eine Gruppe** _ żeby się wszyscy pomieścili [6]

   b. i oni coraz więcej tych dzieci... ...**szczypią w jednej grupie**, **presują w jednej grupie**. [Whisper Large]

5.2. Code detection

In all three versions of Whisper, the matrix code was usually recognised in our data, and the embedded code was a considerable source of transcription errors. However, we obtained transcripts where the output was recognised as an entirely different language, such as Yiddish, as shown in Figure 1 from the transcript of speaker BQ RAC with Whisper Large.

We assume that this may be due to the frequent code switching and unclear pronunciation in the first few seconds of the audio file. This shows that the way in which the recordings are cut before processing affects the accuracy of the automatic transcription. Whisper Base exhibits a similar issue in the same section, but incorporates an even greater variety of languages, ultimately producing a predominantly Germanic transcription, as shown in Figure 2.

---

[5]https://langgener.ijppan.pl/OUT/NF_PAD_II_GD_PL01-146207-151067.wav
[6]https://langgener.ijppan.pl/OUT/NF_PAD_II_GD_PL01-151067-158777.wav

```
 Oh, de, mein. Lymek kleyne, digging toreak Maye. ¿Iško, dzieuš? Vanmó,
buja da Ia, yo 입니다odbe' Borisayumi, hai coo se e Atlantic toreak
bees, Alayne bioje stingiën he, ail-Blirte agl▨▨ t'la' adah prescribed
at y○⊓ld campoon 6á▨▨▨ gap te strata mite aile, he, fill de aile,
biill dichu eil aile, en Alayne et A-yal Cleyne por vellchron woe eil
Q-be Niklidrumki llaaa fadesore Dyneomeiki Kate Jets kamik zum lucean.
Na kam angla lege. Auch daoch van hefol opel en hand. On sakhe ei na.
Fagma dai na motea. Sakhvala sakhe. Ja kvils tu den vessen. Was vils tu
van me vessen? Wie viel hab ich geld gegeben? Dass ich dafte zum Besuch
zum lucean. Ja, ich hab nie lakim jungen. Ich hab niks gegeben. Die
Leute konnten, dass ich begreifen, dass ich hab kein geld, die man
gesagt, die man immer. Ich hab ein Zehber zwehengut. Ja, ja, ja, ja,
ja. Ja, ja, ja, ja. Auch nie. All bohtaker, bohtaker. Ich lass jeden.
Und ja, sov, lass jeden. Und ja, sov, lass jeden.
```

Figure 2: Transcript of speaker BQ RAC by Whisper Base

5.3. Influence of language contact related phenomena

In Section 3.2 we mentioned three phenomena that occur particularly frequently in the situations of language contact: Code Switching (CS), Matter Replication (MAT) and Pattern Replication (PAT). Having them shortly explained, we point out the problems they cause for ASR.

Following [7], we classify phenomena related to the direct transfer of embedding language as CS and MAT. When a single word or phrase takes over the inflection of the matrix language and becomes fully integrated, we call it MAT. CS is similar to MAT, but it shows the lower grade of grammatical integration as shown in (4).

(4)   bo **normalerweise**$_{CS}$ zawsze miałem pół # **Sprachkursu**$_{MAT}$[7]
        'because normally, I had only half of the language course.'

CS and MAT are particularly interesting from the perspective of ASR, as they imply mixing of two language codes with distinct phonological systems, and orthographies. This leads to issues regarding orthographic choice and the identification of words as belonging to the lexicon of the matrix or embedded language. Consequently, we observe the frequent mistakes mentioned in the previous section, such as incorrect spellings, interlingual homophones and problems with ending assignment.

MAT in (4) receives Polish inflection of male noun in accusative singular *-u*. However, the embedded language orthography in the stem of the word is required by the HIAT-based transcription [11], the approach taken by the authors' of LangGener corpus. This is problematic for Whisper, as shown in (5). The word *Sprache* poses the beginning of a CS. Whisper interprets it as MAT and transcribes it following Polish orthography as <sz> and not as capital <S> according to the German orthographic norm. Note that speakers of Polish usually denasalise /ɛ̃/, rendered in Polish orthography as <ę>, in coda to /ɛ/. Therefore, it is unlikely that it could be registered in an audio file.

(5)   a.   to mi było wichtig żeby nasze dzieci jedną **Sprache** gut beherrscht he * # haben[8]

        b.   to mi było wichtig, żeby nasze dzieci jedną **szprachę** gut beherste haben. [Whisper Large]

Another frequently occurring mistake is writing words together, for example an adverb and an adjective, as in the phrase *ganz kleine Kinder* 'very small children' in example (6) is transcribed as *ganzkleinen Kindern*. Note that additionally the inflection ending *-n* is added, although it does not occur in the audio file.

(6)   a.   ta grupa jest z tymi **ganz kleine Kinder** jest dobrze # belegt [9]

        b.   ta grupa z tymi **ganzkleinen Kindern** ist dobrze belegt. [Whisper Large]

Frequent CS seems to increase the error in the transcription, as shown in (7):

[7]https://langgener.ijppan.pl/OUT/BN_WUP_I_GD_PL07-477730-480050.wav
[8]https://langgener.ijppan.pl/OUT/NF_PAD_II_GD_PL01-237220-247670.wav
[9]https://langgener.ijppan.pl/OUT/NF_PAD_II_GD_PL01-58850-64500.wav

(7) a. my to są trzy # **Personalschlüssel ist** # **ich glaube drei** # **zwei Vollzeitkräfte** i chyba jeszcze trzydzieści # godzin jedna **Kraft**[10]

    b. To są trzy **personalsche SEL**, **jakieś global drive**, **zwei Vollzeitkräfte** i chyba jeszcze 30 o godzin, bo jednak **kraft**. [Whisper Large]

The long German compound noun *Personalschlüssel* 'staffing ratio, personnel counting' is most likely recognised as German since the first part of the word is transcribed following German spelling, the German clause *ich glaube drei* 'I guess three' is transcribed as a mix of Polish and English *jakieś global drive*. Although the phrase *zwei Vollzeitkräfte* 'two full positions' is entirely correctly transcribed, the final part of the utterance *trzydzieści godzin jedna Kraft*, which is a mix of Polish and German is transcribed (partly incorrectly) with Polish spelling, since *kraft* is not capitalised.

We also observe here that if a word is spelled incorrectly, this error is consistent throughout the transcript. The word *Personalschlüssel*, which appears four times in one recording, is consistently misspelled with a made-up expression *personalsche SEL* each time.

The third mentioned category, PAT refers to the situation where only structures are borrowed into matrix language. In example (8), the Polish age construction with a habere verb (*mieć*) is replicated into German, instead of copula construction *sein* + Numeral + *Jahre alt*.

(8) meine Mutter die **hat dreizehn Jahre gehabt**$_{PAT}$[11]
    'My mother, she was thirteen years old.'

The influence of pattern replication is difficult to assess using purely qualitative analysis. Often, PAT violates the idiom of the matrix languages, or its syntax, e.g., as demonstrated by the famous sentence uttered by the footballer Lothar Matthäus *again what learned*, which copies the German syntax of 'wieder was gelernt'. Therefore, we leave the analysis of PAT for future research.

Nevertheless, the syntax-based prediction may play a role ASR. For example, the Polish modal verb *może* 'can' is often followed by an infinitive complement. In (9), however, the subject and predicate are inverted. Therefore, the subject *rodzic* 'parent' linearly following the modal predicate seems to be interpreted as similar in pronunciation infiniitve *rodzić* 'give birth', while the actual infinitive is spelled-out by the speaker in German and, therefore, unrecognized.

(9) bo takie dziec * jak dziecko nie dostanie **może rodzic** verklagen die Stadt[12]

    a. jak dziecko nie dostanie, **może rodzić** swe klage w tej szczypie [Whisper Large]

Therefore, we believe that the word-order related PAT are worth inspections in the future. For example, we formulate a working hypothese that the strict German word-order rules regarding the position of the verbal predicate – in the second position in the main sentence or in the last position in the subordinate sentence, frequently copied by bilingual speakers from Polish to German – could pose a potential source of error for ASR too.

6. Ethical Issues

Due to protection policies and laws, such as the General Data Protection Regulation in the European Union, personal and sensitive data must be protected and inaccessible to external entities during and after the preparation and publication of data. Nonetheless, data obtained is expected to be publicly available to other researchers in accordance with the current research standards. Typically, the personal and sensitive data are accessible only to authorised project members. To enable LLM-based processing of data that could contain sensitive information, such information would need to be deleted or hidden before

---

[10]https://langgener.ijppan.pl/OUT/NF_PAD_II_GD_PL01-49260-57407.wav
[11]https://langgener.ijppan.pl/OUT/XL_PIL_GP_DE27-77300-82110.wav
[12]https://langgener.ijppan.pl/OUT/NF_PAD_II_GD_PL01-158897-162417.wav

processing on external servers. Alternatively, it can be conducted exclusively locally while maintaining the data protection standards. Of these two options, only local processing is realistic. First, manually processing transcripts to pseudonymise or erase information would be economically unreasonable. Second, an ASR with integrated LLMs benefits from the syntactic and semantic information encoded by name entities, which are usually the subject to pseudonymisation.

Access to data sets appropriate for the fine-tuning ASR is also more problematic than for other LLM-enhanced applications. Such data is most frequently subject to data protection regulations, meaning the data sets cannot be shared without major changes being undertaken.

7. Discussion and Future Work

It appears that the quality of state-of-the-art ASR systems is still far from enabling instant multilingual transcriptions without considerable expenditure on post-editing. It is important to point out that grammatical and orthographic errors are easily detectable. Problems with translation and the removal of embedded language from transcriptions are more serious, as they are harder to identify. Therefore, these two issues are given higher priority in additional training. Based on the results of the survey, our future work will focus on fine-tuning Whisper using the bilingual LangGener corpus. Additionally, we will explore ways to incorporate ASR into the Research Data Repository of the University of Hamburg, using the Polish-German example as the prototype. This would enable users to search for recordings based on the words used in them without requiring the uploader to provide a transcript upfront.

Limitations

This paper provides a very preliminary overview of problems related to ASR in multilingualism research. It is evident that trials on fine-tuning could offer solutions to these issues. Although multilingual transcripts are the focus of this paper, some of the problems (and therefore potential solutions) may also be relevant to monolingual transcriptions with dialectal features or other vulnerable groups.

Ethics Statement

This work complies with the ACL Ethics Policy. Prior to the current study, we had not taken any actions to pre-train the systems for the needs of the current task.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL in order to: Grammar and spelling check. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Acknowledgements

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, 2022. `arXiv:2212.04356`.

[2] B. Hansen, A. Zielińska (Eds.), Soziolinguistik trifft Korpuslinguistik: Deutsch-polnische und deutsch-tschechische Zweisprachigkeit, Universitätsverlag Winter, 2022. doi:`doi.org/10.33675/2022-82538591`.

[3] S. Wollin-Giering, M. Hoffmann, J. Höfting, C. Ventzke, Automatic transcription of english and german qualitative interviews, Forum Qualitative Sozialforschung / Forum: Qualitative Social Research 25 (2024). doi:`10.17169/fqs-25.1.4129`.

[4] C. Anderson, T. Tresoldi, T. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, J.-M. List, A cross-linguistic database of phonetic transcription systems, Yearbook of the Poznan Linguistic Meeting 4 (2018) 21–53. doi:`10.2478/yplm-2018-0002`.

[5] R. J. Kreuz, M. A. Riordan, The art of transcription: Systems and methodological issues, in: A. H. Jucker, K. P. Schneider, W. Bublitz (Eds.), Methods in Pragmatics, De Gruyter Mouton, Berlin, 2018, pp. 95–120. doi:`doi:10.1515/9783110424928-003`.

[6] B. Hansen, M. Nekula, Die LangGener-Korpora als multifunktionale Ressourcen der Mehrsprachigkeitsforschung zwischen Sozio- und Korpuslinguistik., in: B. Hansen, A. Zielińska (Eds.), Soziolinguistik trifft Korpuslinguistik: Deutsch-polnische und deutsch-tschechische Zweisprachigkeit, Winter Universitätsverlag, Heidelberg, 2021, pp. 175–191.

[7] Y. Matras, J. Sakel, Investigating the mechanisms of pattern replication in language convergence, Studies in Language 4 (2007) 829–865.

[8] C. Centner, Lexikalische Replikation bei deutsch-polnisch Bilingualen in zwei Generationen, Ph.D. thesis, Universität Regensburg, 2024. doi:`10.5283/epub.58164`.

[9] A. Koenecke, A. S. G. Choi, K. X. Mei, H. Schellmann, M. Sloane, Careless whisper: Speech-to-text hallucination harms, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1672–1681. doi:`10.1145/3630106.3658996`.

[10] J. Zhao, H. Shi, C. Cui, T. Wang, H. Liu, Z. Ni, L. Ye, L. Wang, Adapting whisper for code-switching through encoding refining and language-aware decoding, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025. doi:`10.1109/ICASSP49660.2025.10889634`.

[11] K. Ehlich, J. Rehbein, Halbinterpretative Arbeitstranskriptionen (HIAT), Linguistische Berichte (1976) 21–41.