

# An Investigation into the Intra- and Inter-labeller Agreement in the JURISDIC Database<sup>†</sup>

Katarzyna Klessa and Jolanta Bachan

Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland  
klessa@amu.edu.pl  
jolabachan@gmail.com

## ABSTRACT

This paper reports on the labeller agreement examination in the annotation of the JURISDIC Polish speech database. One recording session (355 utterances, read and spontaneous speech) was annotated twice by four labellers: in the first step the task was to annotate the data based only on listening to the recordings, in the second step the annotators were additionally provided with the prompt command and the orthographic input text available originally for the speaker. The authors look at the influence of the input text on the annotation process. The annotation files are compared as regards to the number and type of noise labels, special events labels, spelling and segmentation differences both among labellers and also for each of the labellers individually.

## 1. Introduction

Development of new speech and language technologies requires building large speech corpora. A typical speech corpus consists of speech signal files and time-aligned annotations. The speech corpus may contain recordings of speech of either colloquial or domain specific character as regards lexical and structural features. Speech itself may be produced in a spontaneous or controlled manner, it may be monotonous or acted to imitate various expressive speech styles. The language of the speech may be native or non-native, national standard or dialectal. The assumed hierarchy of the factors depends on what kind of information the corpus creator needs to collect, in what field the corpus is to be used [1–4] and whether it is to be externally validated or shared after being completed e.g. [5–10]. In most cases manual annotation of the speech signal is required. Human work is time-consuming and thus very expensive, being at the same time prone to certain types of errors (for instance spelling errors), therefore new techniques are developed to automatise the annotation process such as grapheme-to-phoneme conversion programs (e.g. PolPhone, [11]) and automatic segmentation systems (SALIAN, [12, 13]) which deliver phone level annotation of reasonably good quality. However, humans are still indispensable to check and correct those annotations and then, in the ideal case, expert labellers are employed for final correction and ad-

<sup>†</sup>The project is supported by The Polish Scientific Committee (Project ID: R00 035 02 – “Technologies for processing and distributing verbal information in internal security systems”).

justment. After that, a post-final automatic parsing is often performed to search for bugs such as unspecified labels or spelling errors [1, 9, 10].

The present paper reports on a study carried out to investigate the similarities and inconsistencies of annotation results within the JURISDIC speech database creation for a large vocabulary continuous speech recognition (LVCSR) system. The database was designed to provide recordings of approximately 1500 voices, therefore from the very beginning it was known that a large number of people would need to be involved and that there would be a strong demand for tools supporting automatic annotation in as many respects as possible. In this study we compare the annotations among labellers, and we also examine the annotations within the material delivered by each of the labellers individually to check the consistency of their work.

## 2. Speech material

The linguistic and acoustic material for the study was one 355-utterance recording session providing approximately 40 minutes of two-channel speech data from a female speaker selected randomly from the JURISDIC database [1, 2]. The session was a typical JURISDIC session containing recordings assessed as good at the stage of the ordinary preliminary assessment performed for all recordings before including them into the database (cf. [1] for more details of the preliminary assessment procedure). The session recording scenario was composed of utterances representing three major types of text:

- Type A – (semi-)spontaneous speech: elicited dictation of short descriptions, isolated phrases, numbers or letter sequences (69 utterances).
- Type B – read speech: phonetic coverage and syntactically controlled structures. These sentences were created for research purposes to provide triphone and diphone coverage and to cover the most important syntactic structures (159 utterances).
- Type C – read speech: semantically controlled structures of utterances containing specialised vocabulary, legal and police texts, application words (127 utterances).

## 3. Annotation specification

At the stage of manual processing of the JURISDIC database acoustic-phonetic description a SPEECON-based specification is used [9, 10]. The specification assumes orthographic, word-level transcription. Because of the “office” conditions of the recordings (recordings were not performed in the sound-proof recording studio), apart from transcription of text, four types of noise labels are introduced: non-speech speaker noises (breaths, cough, sniffing, lip-smacking, etc.; label: [spk]), fillers (label: [fil]), intermittent noises (phone, door bell, music, cross talk, etc.; label: [int]), stationary noises of a relatively stable character over a period of time (background noises, rain, wind, silent speech in the background, etc; label: [sta]). Additionally, three types of special events are allowed in the annotation: mispronunciation (marked with an asterisk attached to the mispronounced but still intelligible word: \*word), an unintelligible stretch of speech (two asterisks inserted instead of the unintelligible fragment: \*\*),

waveform truncated due to a recording error (~word). Segmentation of the speech data is provided according to the following boundary indicators: pause, filled pause, fillers, prosody or syntax. The most important boundary indicator is an acoustic pause, and when the pause exceeds half a second inserting a boundary is obligatory, for shorter pauses the insertion of a boundary marker depends on the remaining indicators and the labeller's judgement. An important issue is that according to the above specification subjective judgements are allowed to a certain extent, thus it is crucial to clarify as much details as possible also with respect to individual decisions taken by the annotators.

#### 4. Annotation procedure

Development of the JURISDIC speech database requires employing over 20 people at a time to provide annotations. The annotation of the speech recordings in the corpus is performed in two stages: (1) regular annotation (done by students at the Institute of Linguistics of Adam Mickiewicz University in Poznań); (2) annotation verification (experienced students, expert phoneticians).

The annotators selected for the present experiment were four students from the group of verifiers, each of them with at least five months of annotation experience at the moment of performing the study tasks. All annotators worked in similar conditions of a quiet phonetic laboratory using equipment of similar standard. The computer tools involved in the annotation process was the *Annotation Database Manager* (software created specifically for purposes of the JURISDIC project [1, 2]) – a system based on MSDE 2000, and Windows 2003 Server integrated with Transcriber [14]. Keyboard shortcuts were defined in Transcriber for the special labels used in the annotation with a view to eliminate the number of possible typing errors.

#### 5. Experiment design

The experiment was focused on the comparison of annotation results provided by four labellers independently annotating exactly the same recording session. The aim was to observe the differences and similarities between the annotation provided by particular labellers with regard to the number and types of special events and boundary markers in the annotations. It was also intended to observe the possible differences in annotation while labellers were or were not provided with the original input text, a facility introduced to accelerate the annotation task and reduce misspelling. Using this facility provides each annotation file with the prompt text (e.g. “Read the following sentence” or “Tell a story about your friend’s birthday party”) or (for the read texts) the actual text which the speaker was asked to read during the recording process. In the present experiment the annotators were required to annotate the speech data twice in the following manner:

- Step 1. Annotation of the speech data based only on listening to the speech signal and visual inspection of the waveform with no additional text input.
- Step 2. Annotation of the speech data based on listening to the speech signal and waveform visual inspection and: the orthographic text input (for read speech); the prompt command provided to the speaker (for read and spontaneous speech).

## 6. Results

Figure 1 shows the types of events annotated by the four labellers in the two steps of the experiment. The total number of noise labels was greater in Step 2 of the experiment. This may result from the fact that in Step 1 the annotators were more focused on the text itself rather than on the other phenomena because they needed to understand and transcribe speech while in the Step 2 they were provided with the input text so they could pay more attention to the other types of events and categorise them higher, therefore make the decision to label those noise events. The fatigue resulting from typing the text could also affect the perception of noises in the recordings and their categorisation.

The speaker noise label [spk] has the highest number of occurrence, in both Step 1 and Step 2 of the experiment, however in Step 2 it is used more frequently. The filler label [fil] was used most consistently across labellers, suggesting that the conditions on when to use this label were clear and that the filler sounds are easily distinguishable for the labellers. In both steps Labeller's 4 insertion of the speaker noise labels and the mispronunciation labels does not follow the trend set by the other labellers. Labeller 4 used significantly fewer speaker noise labels [spk] and significantly more mispronunciation labels. This may be justified by the Labeller's different conception of what a speaker noise is and when to use the mispronunciation label. However, the other labellers seem to agree on when to insert the speaker noise labels and when to mark the word as mispronounced, which might suggest that the conditions of the usage of those labels are rather clear and Labeller 4 should preferably adapt the others' definition of the mispronunciation label in the future.

The results of Labeller 1 were the most outlying ones especially as regards the intermittent noise label [int]. Moreover, Labeller 1 inserted over 100 fewer speaker noise labels in Step 1 than in Step 2. Such results may be explained by the Labeller's high sensitivity to noise, especially in Step 2, when the input text was provided.

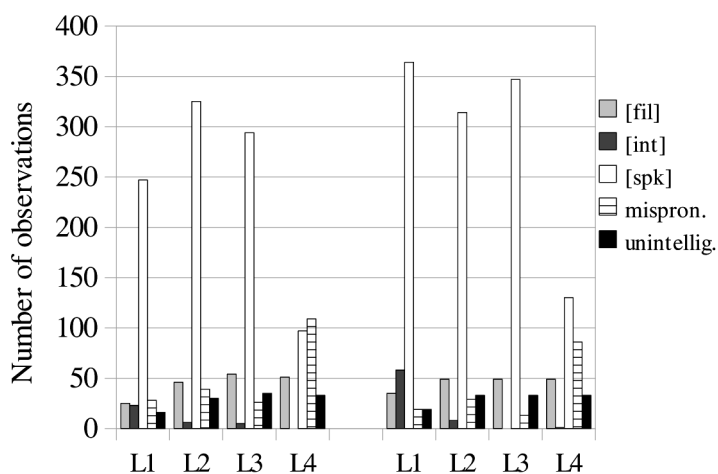


Figure 1. Occurrence of the most frequent noise labels ([fil, int, spk]) and mispronunciation markers (\*, \*\*) in Step 1 (left) and Step 2 (right) of the experiment; L=Labeller.

The total number of spelling errors was greater in Step 1 (22 errors) than in Step 2 (12 errors), confirming the facilitation effect of providing the annotators with the text input resulting in reducing the number of spelling mistakes. The stationary noise label [sta] was used only once which indicates low or no background noise in the selected session. The session turned out to be also well controlled at the recording stage which is confirmed by only one usage of the truncation label.

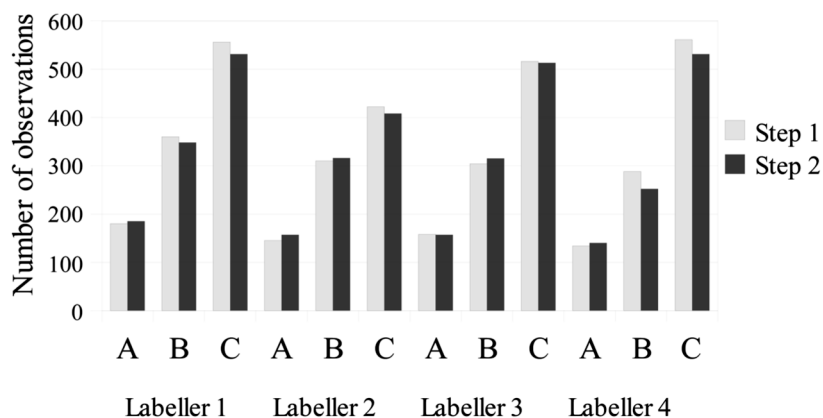


Figure 2. The number of boundary markers inserted by four labellers in the annotation of the three text types: A, B or C.

In Figure 2 the number of the boundary markers inserted by the four labellers depending on text type (A, B, C) is depicted. For all annotators the number of boundary markers appears to be strongly related to the text type by being greatest for the semantically controlled structures (type C: complex sentences with domain specific vocabulary which might have been unfamiliar to most of the speakers) and smallest for the semi-spontaneous speech (type A). Although the overall number of boundary markers inserted in Step 1 was slightly higher than in Step 2, the observed tendencies were quite consistent in the two steps of the experiment for all the subjects.

Figure 3 shows the total number of speech events annotated for various text types by the four labellers. Again, it is clearly visible that the speaker noises are most frequent noise labels in the annotations for all three text types. The speaker noise labels are relatively more frequent in the semantically controlled read speech (type C) than in the two other types, possibly resulting from the texts' lexical and grammatical difficulty as well as from the length of sentences leading to more breath pauses, lip smacking etc. expressing speaker's fatigue.

Another observation is that the filler label was found significantly more common in the semi-spontaneous A text type speech which may be explained by the more natural process of speech production characterized by hesitations and disfluencies. On the other hand, fillers appeared to be practically absent in the text type B with well-formed utterances of medium length created specifically for phonetic and grammar structure coverage. Hesitations might be anticipated in those sometimes tricky sentences,

however they seem to have been eliminated in the recording process where the sentences could have been repeated by the speaker several times to obtain utterance of good quality and as few disturbances as possible during the recording session. The B-text type sentences were typically not very long, so repetitions might improve the recording significantly.

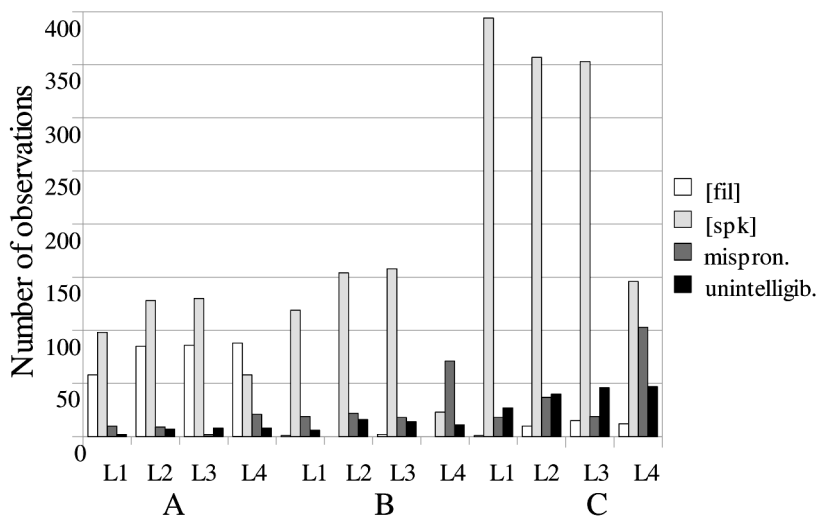


Figure 3. Speaker noise and mispronunciation labels in the three text types A, B or C in the annotation files of four labellers; L=Labeller.

Table 1 presents the number of utterances in the three main A, B, C subsets and the average numbers of all types of events per utterance. The results show that A-text type recordings, although small in number of utterances, turned out to be very rich in special event labels, mainly fillers. The high number of fillers could affect the higher number of boundaries per utterance. B-text type recordings had a very low proportion of special event labels, apparently resulting from the recording procedure which eliminated the slips of the tongue and hesitations, but preserved the need to making pauses. The need of making pauses in quite long sentences is showed in the number of boundaries. Finally, the results of C-text type recordings confirm the higher rate of special events per utterance, mainly the speaker noises, and the utterance length, indicated by the

Table 1. The number of events per utterance in the three text types: A, B or C

| Text type | Number of utterances | Events per utterance | Boundaries per utterance |
|-----------|----------------------|----------------------|--------------------------|
| A         | 69                   | 1.47                 | 2.28                     |
| B         | 159                  | 0.53                 | 1.96                     |
| C         | 127                  | 1.64                 | 3.97                     |

number of boundaries per utterance. The latter numbers are slightly higher for C-type texts and lowest for B-type texts, which may confirm, to some extent, the dependency of the event number on the type of text.

## 7. Conclusions and future work

The present study aimed at verifying the consistency of speech signal annotation by human experts and investigating the influence of the annotation facilitation tool, the input text when seen together with the speech signal, on the annotation quality. The obtained results showed that experienced annotators do agree in their categorisations of the special events, however, some justified exceptions may happen and are labeller-dependent. The high consistency on the boundary insertion is also encouraging, especially that the segmentation conditions are specified as dependent on subjective judgement to a certain degree. Some of the above observations do not directly concern the inter or intra-speaker comparisons, namely the statistics on the dependency of the annotation on text types or the influence of the input text. However, they may well serve as an addition to the annotation specification, providing a preliminary insight into the complexity of the acoustic-phonetic description of speech. An important future work will be to investigate the exact placement of the special labels and boundary markers in order to obtain a fully informative measure of labeller agreement. It is planned to provide such analyses based on bigger samples of speech data along with a large-scale annotation mining across the whole JURISDIC database.

### REFERENCES

- [1] Demenko, G., Grochowski, S., Klessa, K., Ogórkiewicz, J., Lange, M., Śledziński, D., Cylwik, N. 2008. Jurisdic–Polish Speech Database for taking dictation of legal texts. In: *Proceedings of the Sixth International Language Resources and Evaluation*, Ed. European Language Resources Association (ELRA), Marrakech, Morocco.
- [2] Demenko, G., Grochowski, S., Klessa, K., Ogórkiewicz, J., Wagner, A., Lange, M., Śledziński, D., Cylwik, N. 2008. LVCSR Speech Database – JURISDIC. In: *Proceedings of NTAV / SPA 2008, Signal Processing: Algorithms, Architectures, Arrangements, and Applications, New Trends in Audio and Video (AES)*, Poznań University of Technology, September 25–27, 2008.
- [3] Mostefa, D., Hamon, O., Choukri, K., Van den Heuvel, K., H., Choukri, K., Gollan, Chr., Moreno, A. 2006. TC-STAR: New language resources for ASR and SLT purposes. In: *Proceedings of the LREC 2006*, Genoa, Italy.
- [4] Sundermann D. 2005. A language Resources Generation Toolbox for Speech Synthesis, TC\_STAR publication, [http://www.tc-star.org/publicazioni/scientific\\_publications/Siemens/2005/ast2005.pdf](http://www.tc-star.org/publicazioni/scientific_publications/Siemens/2005/ast2005.pdf).
- [5] Docio-Fernandez, L., Cardenal-Lopez, A., Garcia-Mateo, C., 2006. Automatic Speech Recognition Evaluation: The UVIGO System, TC-STAR Workshop on Speech-to-Speech Translation, Barcelona.
- [6] Gibbon, D., Moore, R. and Winski, R. 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.
- [7] Gibbon, D., Bachan, J. 2008. An automatic close copy speech synthesis tool for large-scale speech corpus evaluation. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Ed. European Language Resources Association (ELRA), 28–30 May 2008 Marrakech, Morocco.

- [8] Henk van den Heuvel, Sanders, E. 2006. Validation of language resources. In: *TC-STAR*, TC-STAR Workshop on Speech-to-Speech Translation, Barcelona, June 19–21, 2006.
- [9] Fischer, V., Diehl, F., Kiessling, A., Marasek, K. 2000. Specification of Databases – Specification of annotation. SPEECON Deliverable D214.
- [10] SPEECON: <<http://www.speechdat.org/speecon/index.html>>, accessed on 2008-11-05.
- [11] Demenko, G., Wypych, M., and Baranowska, E. 2003. Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis. In: *Speech and Language Technology*, vol. 7. [Ed.] PTFon, 79–97, Poznań.
- [12] Szymański, M., Grocholewski, S. 2005. Semi-Automatic Segmentation of Speech: Manual Segmentation Strategy. Problem Space Analysis. In: *Advances in Soft Computing, Computer Recognition Systems*, 747–755, Springer Berlin.
- [13] Szymański, M., Grocholewski, S. 2008. Error prediction-based semi-automatic segmentation of speech databases. In: *Proceedings of TSD 2008 Conference*. September 8–12, 2008, Brno, Czech Republic.
- [14] Transcriber: <<http://trans.sourceforge.net/>>, accessed on 2008-12-05.