ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746160

CARINA – A CORPUS OF ALIGNED GERMAN READ SPEECH INCLUDING ANNOTATIONS

Hannes Kath[1,3]*, Simon Stone*[1]*, Stefan Rapp*[2]*, Peter Birkholz*[1]

[1]Institute of Acoustics and Speech Communication, Technische Universität Dresden
[2]Fachbereich Informatik, University of Applied Sciences Darmstadt
[3]German Research Center for Artificial Intelligence (DFKI)

ABSTRACT

This paper presents the semi-automatically created Corpus of Aligned Read Speech Including Annotations (CARInA), a speech corpus based on the German Spoken Wikipedia Corpus (GSWC). CARInA tokenizes, consolidates and organizes the vast, but rather unstructured material contained in GSWC. The contents are grouped by annotation completeness, and extended by canonic, morphosyntactic and prosodic annotations. The annotations are provided in BPF and TextGrid format. It contains 194 hours of speech material from 327 speakers, of which 124 hours are fully phonetically aligned and 30 hours are fully aligned at all annotation levels. CARInA is freely available[1], designed to grow and improve over time, and suitable for large-scale speech analyses or machine learning tasks as illustrated by two examples shown in this paper.

Index Terms— CARInA, speech data, prosodic annotation

1. WHY ANOTHER SPEECH CORPUS?

The application of neural networks in speech processing requires enormous amounts of data, often hundreds of hours of speech material with associated annotations from a large number of speakers. While in other languages, such as English or Russian, there are extensive corpora (e.g. the Hong Kong Corpus of Spoken English containing 200 hours [1], or the Corpus Of Russian Professionally Read Speech containing 60 hours [2]), German-speech corpora are not extensively annotated [3] or of comparatively small size. Two examples of extensively annotated corpora are the BITS Unit Selection Synthesis Corpus (BITS-US) [4] containing 13 hours and the Kiel Corpus of Spoken German (14 hours) [5], but both are not freely available.

Manually annotated corpora are generally of higher quality than automatically created corpora, but usually expensive to create, therefore rarely under a free or permissive license, and often not large enough to train deep learning models. This paper describes the automatically annotated Corpus of Aligned Read Speech Including Annotations (CARInA), which is built on the German Spoken Wikipedia Corpus (GSWC) [6]. It uses a large amount of input data and strong selection criteria to form a carefully annotated, comprehensive German-language data set. CARInA contains the GSWC's extensive, freely accessible and thematically diverse speech material and is published under the same permissive license.

2. SPEECH MATERIAL

CARInA is based on the speech resources of the GSWC [6]. The GSWC grows over time (monitor corpus) and currently contains the unedited recordings, orthographic alignments created by an extended SailAlign-Algorithm [7], phonetic alignments created by MAUS [8] and meta data files of 1015 Wikipedia articles (386 hours of speech material) on various topics. The articles were read and recorded by 337 speakers without a standardized procedure. The GSWC is organized by articles and published under a free license.

For creating CARInA the meta data files were used to assign the articles to the speakers and to determine their self-identified gender (267 male, 36 female, 34 unspecified). To facilitate processing, the recordings were split into individual sentences. CARInA contains all sentences of the GSWC for which a start- and an end-sample were found by the alignments. Because of low audio quality, strong dialects and mislabelling of audio files, only 129 hours of complete sentences could be fully aligned on word and phone level. As the alignment works better for standard German utterances, we consider this selection beneficial, as it leads to more representative, standard German utterances of high audio quality.

3. CORPUS DESIGN

CARInA is split into two subsets contained in the folders `Complete` and `WorkInProgress`. The subset of sentences in the folder `Complete` has been fully annotated at the orthographic, morphosyntactic, broad phonetic (canonic), narrow phonetic (phone alignment) and prosodic level. The folder `WorkInProgress` contains the extracted sentences with at least one incomplete annotation level.

[1]http://dx.doi.org/10.25532/OPARA-144

Both folders contain a sub-folder per speaker represented by a number and a gender specifier (*m*ale/*f*emale/*u*nspecified), e.g. `SpeakerID0001_f`. The sub-folders contain an audio file, a BPF file and a TextGrid file for each sentence. The file names follow the pattern `article****_sentence****`, where `****` is the four digit number of the article or of the sentence in the article, respectively.

The audio files are in WAV format with one channel, a sampling rate of 44 100 Hz and a resolution of 16 bit per sample. The BPF files (extension `.par`) are created according to the standard of the *Bayrisches Archiv für Sprachsignale* (BAS) [9]. The TextGrid files were created with the MATLAB tool mPraat [10] and can be used for speech processing in Praat [11]. The BPF and TextGrid files contain information about the word alignment, the canonic pronunciation, the part-of-speech tags and the phone alignment.

Prosodic information was generated in four ways for all sentences in the folder `Complete`. Since no conclusive statement could be made about the quality of prosodic annotations, all results were stored in the folder `WorkInProgress` and provided in so-called `.snippet` files which can be inserted into the BPF or TextGrid files by simply appending the snippet to the respective file. The snippet file names are composed of the original file name, the name of the prosodic annotation system used and the extension `.snippet`.

In addition, CARInA also contains three meta files: `ContentStatus.txt` gives an overview of the annotation status of every sentence. This file can be used for clustering the data for specific applications, e.g. for selecting all sentences with complete word-level annotation, but not necessarily complete phoneme-level annotation. `MissingSentences.txt` contains the written content of all sentences from the original articles, for which the audio could not be extracted. Finally, the file `README.txt` contains much of the information from this paper regarding the corpus contents and structure, as well as the label sets for the POS tags and the prosody annotation.

4. AUTOMATIC ANNOTATION PIPELINE

Our annotation pipeline works as follows: First, the data of the GSWC were processed and organized by speaker IDs and sentences. Then, dictionaries were created, which mapped every unique word from the GSWC to its canonic and morphosyntactic annotations. All sentences, which were fully annotated at these levels, were prosodically annotated with the automatic prosody tagging systems PyToBI [12] and Prosody Recognition Revisited (PRR) [13, 14]. The following describes the individual steps in greater detail.

4.1. Orthographic and phonetic annotation

The first step of the automatic annotation pipeline was the preprocessing of the GSWC data. The GSWC is organized into articles, which for most applications is less convenient than splitting into speaker and sentence. To assign the articles to the user names and generate the speaker IDs used in CARInA, the `audiometa.txt` files were used. For each article, the GSWC contains one XML-file named `aligned.swc` containing the orthographic and phonetic alignment.

Due to the different procedures for aligning the orthographic boundaries (SailAlign) and phonetic boundaries (MAUS), the word boundaries did not agree (compare Figure 1). Spot checks conducted by the original authors of the GSWC revealed more accurate word boundaries for the phonetic alignments [6]. To increase consistency, the word-level alignments found by SailAlign were not adopted in the CARInA, and replaced by the word boundaries according to MAUS. An investigation using random sampling revealed an increased error rate of phonetic alignments in sentences with word boundary differences of more than 1500 ms. These large differences are to a significant extent due to different contents of the speech recording and the associated article. Sentences with a difference in word boundaries of more than 1500 ms were therefore marked as phonetically incomplete. The remaining phonetically fully aligned subset contained 54 141 sentences with 124:15 hours of speech material and formed the basis of the sub-corpus Complete, which was further refined in the next processing stage.

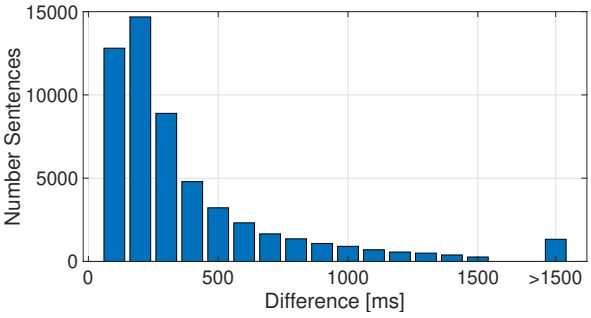

Fig. 1. Number of completely phonetically aligned sentences versus the time difference between the word boundaries given by the annotations at the phone-level (MAUS) and the word-level (SailAlign).

4.2. Canonic and morphosyntactic annotation

Canonic and morphosyntactic annotations were added using the free online dump of the dictionary Wiktionary[2] provided by the Umeå university. The German Wiktionary is the most comprehensive freely available German-language dictionary [15]. It contains over 1 million words and was used to create three dictionaries mapping full-form orthographic words to their canonical transcription, POS-tag and their syllabification (compare Table 1). Since the POS-tags were assigned

[2]http://ftp.acc.umu.se/mirror/wikimedia.org/dumps/dewiktionary/

without context words, different syntactic functions cannot be distinguished. 102 766 different words were missing in the dictionary for total coverage with composites, proper nouns and numbers representing the major groups. Using semi-automatic and manual procedures the numbers from one (German "eins") to ninety-nine (German "neunundneunzig") with declinations, the years from 1100 to 1999 and the most frequently occurring proper nouns and composites were added.

Table 1. Number of words in each dictionary, extracted from Wiktionary and total with words added manually.

Dictionary	Words Wiktionary	Words total
Canonical transcription	764 185	765 847
Part of speech	915 648	917 303
Syllabification	825 888	827 536

Using these dictionaries, the speech data annotation was enriched with canonic, morphosyntactic and syllabic information. The canonical transcription was converted from IPA to SAMPA and phones were separated by spaces for more convenient processing. Stress marks from the canonical transcription are used to find and mark word stress in the phonetic annotation. 29:47 hours of speech material were annotated at all of these levels and formed the sub-corpus Complete.

4.3. Prosodic annotation

Prosodic labels cannot be easily assigned based on segments (such as words or phones), since prosody describes features such as intonation or accentuation at the supra-segmental level [16]. Therefore, the prosodic labels could not be directly derived from existing annotations. The sub-corpus Complete was instead automatically annotated at the prosodic level using ToBI labels [17] generated by two different programs: PyToBI and Prosody Recognition Revisited.

The input of the Python-based prosodic annotation system PyToBI [12] is the speech audio and a TextGrid file with aligned words and phones. Prosodic contours are computed using Praat and each word is labeled from a subset of ToBI labels containing deaccentuation, six pitch accents, seven boundary tones and four break indices. The results were saved as BPF files (tier label `PRB:`) and as TextGrid files containing the two tiers `tones` and `breaks`.

Prosody Recognition Revisited (PRR) is a revision of the Prosody Recognition System, which was created for automatic prosodic labelling of German speech [13, 14]. The PRR input consists of a speech audio file and the corresponding graphemic text file. In several processing steps the fundamental frequency contour of the audio file is parameterized and the text file is segmented in syllables and phones (using the created canonic transcription dictionary) and aligned with the audio file. Using a pretrained decision tree, prosodic labels are generated for each syllable. To annotate the sub-corpus Complete three different decision trees were trained resulting in three different sets of labels. The three corpora used for training were the Stuttgart Radio News Corpus (SRNC) [13], the BITS-US, and the Kiel Corpus of Spoken German Read Speech (KCSGrs) [5]. In collaboration with the original authors of the KCSGrs, the contour-based KIM labels [18] were converted to a subset of ToBI labels (H*, H*L, L*H, H% and %) with loss of information. The file `README.txt` contain these conversion rules.

The prosodic annotation systems were validated using the KCSGrs. The level of compliance [19] between the predicted labels and the converted KCSGrs labels was calculated using 10-fold cross-validation (compare Table 2).

Table 2. Mean level of compliance between the converted KCSGrs and the prosodic annotation systems using 10-fold cross-validation. The results for correct location (Loc) and label (Lab) of tones and breaks are presented.

	Level of compliance [%]			
Score	PRR KCSGrs	PRR SRNC	PRR BITS-US	PyToBI
Loc Tones	82.14	72.46	70.52	62.14
Lab Tones	81.29	66.74	65.47	59.48
Loc Breaks	89.88	88.33	87.1	94.65
Lab Breaks	88.96	85.93	84.75	92.96

As a baseline, a "naive" system predicting only deaccentuation would have 76.6 % compliance for tones and 79.4 % for breaks. Given that the accuracy of prosodic labels is difficult to determine in subjective terms, no final system for the prosodic annotation was picked and instead results were made available and can be chosen by inserting the respective snippet file into the annotation file.

5. FINAL DATA SET

CARInA contains 194:20 hours of speech material from 327 speakers (34 female, 259 male and 34 unspecified). 124:15 hours of speech material are fully orthographically and phonetically aligned using SAMPA. The sub-corpus Complete contains 29:47 hours and is fully annotated at the orthographic, canonic, morphosyntactic, phonetic and prosodic level. The distribution of speech material is inhomogeneous (compare Figure 2): 61 % of the speech material in the sub-corpus Complete is spoken by 4 % of the speakers with one speaker (`SpeakerID0041_m`) contributing over 8 hours of the fully aligned speech material.

Since the speech recordings were made in an uncontrolled environment, the quality of the recordings was evaluated in terms of the Signal to Noise ratio (SNR). The noise power was calculated over segments with no speech (as indicated by the annotation) and the signal power was calculated over all remaining segments [20]. The calculated SNR must thus be seen as a lower bound because the noise power was likely

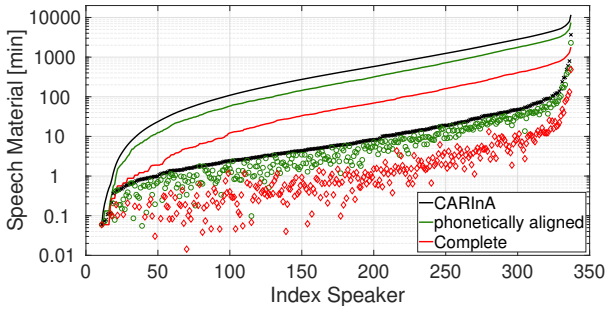

Fig. 2. Speech material of CARInA, the fully phonetically aligned part of CARInA and the sub-corpus Complete. The speech material is plotted per speaker and cumulatively.

to be overestimated due to non-speech sounds like breathing occurring in the "silent" segments. With an average SNR of 26.8 dB, the audio quality of the sub-corpus Complete can be considered high [20].

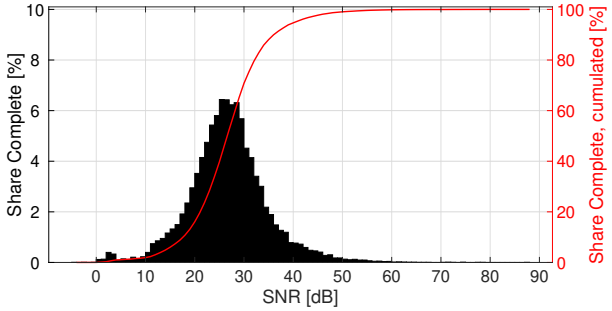

Fig. 3. Distribution of the SNR of all sentences of the sub-corpus Complete, plotted as a histogram and cumulatively.

6. VALIDATION

To illustrate the corpus' suitability for big data analysis tasks, a formant map was calculated for all vowels in the sub-corpus Complete. To that end, the central third of each annotated vowel segment was extracted from the audio files and the first two formants were calculated in Praat (compare Figure 4) using default settings. Due to the limitations of automatic formant measurement, the variances in Figure 4 are strongly overestimated. The resulting formant maps largely correspond to the manually measured reference in [21].

As another typical application for an annotated speech corpus, a command word recognition system[3] was devised using a CNN with 24 layers (5 convolutional layers). Using spectrograms calculated from the audio signals, the CNN was trained over 25 epochs with a learning rate of $\eta = 10^{-4}$, reduced by a factor of 10 after 20 epochs. The accuracies in Table 3 were obtained without adjusting the hyperparameters and show that CARInA is suitable as training material for speech processing.

[3]https://de.mathworks.com/help/deeplearning/ug/deep-learning-speech-recognition.html

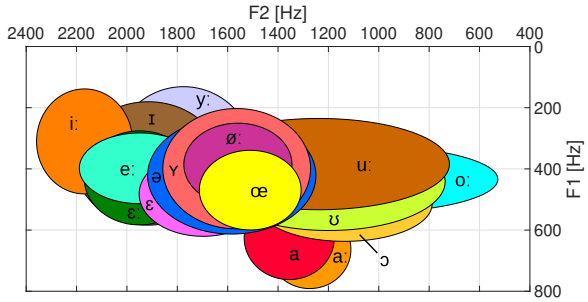

Fig. 4. Formant values and standard deviations calculated by Praat of the sub-corpus Complete, male speakers.

Table 3. Compilation and performance of data sets for speech command recognition. Set 1 is not uniformly distributed. There was no overlap between the speakers in the training and validation sets.

		Utterances per word		
Set	Words	Training	Validation	Accuracy
1	373	21 – 1348	3 – 182	81 %
2	282	25	4	61 %
3	92	50	7	59 %
4	20	100	15	89 %

7. CONCLUSION AND FUTURE WORK

We presented CARInA, an open-source speech corpus with high audio quality and multi-level annotations. It is based on the GSWC's speech data and added automatically generated, but rigorously vetted, canonic, morphosyntactic and prosodic annotation. Due to the increase of speech material of the German Wikipedia from 32-33 hours per year [6], the CARInA can be used as a monitor corpus.

With 30 hours of fully aligned sentences the database is of considerable size. The size of the sub-corpus Complete could be increased (up to quadrupled) by adding more words to the dictionaries. Taken the positions of word stress into account, composites in particular can be added automatically by the components contained in the Wiktionary. The quality of POS-tags could be increased by using context-based methods.

Although they were included as optional snippets, the prosodic labels generated by PyToBI and PRR need to be evaluated before further use. This could be accomplished, for example, by a re-synthesis study where different versions of the same sentence are synthesized using different prosodic features as annotated by the various systems.

8. REFERENCES

[1] W. Cheng, C. Greaves, and M. Warren, "The creation of prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic)," *International computer archive of modern English*, vol. 29, pp. 47–68, 2005.

[2] P. Skrelin, N. Volskaya, D. Kocharov, K. Evgrafova, O. Glotova, and V. Evdokimova, "CORPRES," in *Text, speech and dialogue*. 2010, pp. 392–399, Springer.

[3] P. Puchtler, J. Wirth, and R. Peinl, "HUI-Audio-Corpus-German: A high quality TTS dataset," *ArXiv*, vol. abs/2106.06309, 2021.

[4] T. Ellbogen, F. Schiel, and A. Steffen, "The BITS Speech Synthesis Corpus for German," *AGE*, vol. 47, pp. 40 – 43, 2004.

[5] K. Kohler, B. Peters, and M. Scheffers, "The Kiel Corpus of Spoken German read and spontaneous speech," Tech. Rep., Christian-Albrechts-Universität, 2017.

[6] T. Baumann and A. Köhn, "The Spoken Wikipedia Corpus collection: harvesting, alignment and an application to hyperlistening," *Language resources and evaluation*, vol. 53, pp. 303–329, 2016.

[7] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Workshop on new tools and methods for very-large scale phonetics research*, 2011.

[8] F. Schiel, "MAUS goes iterative," in *Proceedings of the 4th international conference on language resources and evaluation*. 2004, ELRA.

[9] F. Schiel, S. Burger, A. Geumann, and K. Weilhammer, "The partitur format at BAS," Tech. Rep., Ludwig-Maximilians-Universität München, 1997.

[10] T. Bořil and R. Skarnitzl, "Tools rPraat and mPraat," in *Text, speech and dialogue*. 2016, pp. 367–374, Springer.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 6.1.53," retrieved 8 September 2021 from `http://www.praat.org/`.

[12] M. Domínguez, P. Rohrer, and J. Soler-Company, "PyToBI: a toolkit for ToBI labeling under python," in *Interspeech 2019*. 2019, pp. 3675–3676, ISCA.

[13] S. Rapp, *Automatisierte Erstellung von Korpora für die Prosodieforschung*, Ph.D. thesis, Universität Stuttgart, 1998.

[14] S. Rapp, "Automatic labelling of German prosody," in *International conference on spoken language processing 1998*. 1998, pp. 1267–1270, ISCA.

[15] F. Lin and A. Krizhanovsky, "Multilingual ontology matching based on Wiktionary data accessible via SPARQL endpoint," *CEUR workshop proceedings*, vol. 803, pp. 1–8, 2011.

[16] S. Nooteboom, "The prosody of speech: Melody and rhythm," *The handbook of phonetic sciences*, vol. 5, pp. 640–673, 1997.

[17] A. Agarwal and A. Jain, "Tones and Break Indices for speech processing – A review," in *Proceedings of the 4th national conference: Computation for nation development*. 2010, pp. 319–324, New Delhi.

[18] B. Peters and K. Kohler, "Trainingsmaterialien zur prosodischen Etikettierung mit dem Kieler Intonationsmodell KIM," Tech. Rep., Christian-Albrechts-Universität, 2004.

[19] J. Pitrelli, M. Beckmann, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," in *International conference on spoken language processing 1994*. 1994, pp. 123–126, ISCA.

[20] H. Hirsch, "Estimation of noise spectrum and its application to SNR-Estimation and speech enhancement," Tech. Rep., International computer science institute, 1993.

[21] W. Sendlmeier and J. Seebode, "Formantkarten des deutschen Vokalsystems," Tech. Rep., Technische Universität Berlin, 2007.