Edalatishams, I. (2017). LeaP corpus (Review). In M. O'Brien & J. Levis (Eds). *Proceedings of the 8th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Calgary, AB, August 2016 (pp. 236-240). Ames, IA: Iowa State University.

CORPUS REVIEW

LeaP Corpus

Idée Edalatishams, Iowa State University

INTRODUCTION

The LeaP corpus is a collection of speech by L2 learners of German and English, annotated mostly for phonetic features that contribute to prosody (Gut, 2014a). It was created as part of a larger project titled *Learning Prosody in a Foreign Language* aimed at describing learners' acquisition of prosody at both phonetic and phonological levels as well as the learner characteristics that affect the process of learning prosody (Gut, 2014b). The corpus was developed at the University of Bielefeld, Germany, from 2001 to 2003.

The Speakers

The LeaP corpus consists of speech from a total of 131 speakers varying in age, gender, L1, proficiency, age of first L2 exposure, length of L2 exposure, and non-linguistic factors such as motivation or musicality. Especially proficient L2 learners, or "superlearners" as Gut (2014a) calls them, were included among the speakers in order to provide data for exploring "ultimate phonological attainment" and the non-linguistic factors that contribute to native-like achievement (Gut, 2012). Some learners were recorded before and after receiving prosody training, so as to provide evidence for the effects of guided learning on pronunciation. Other learners were recorded before and after going abroad in order to provide information on the effects of unguided training on pronunciation. In addition to L2 learners, native speakers of English and German were also included to serve as control groups (Gut, 2014a).

Corpus Development

The speakers were recorded reading aloud a list of non-words and a short story, retelling the same story in their own words, and responding to informal interview questions. The resulting 12 hours of recordings were transcribed and annotated using manual and automatic methods and are available for free download in form of 359 annotated and time-stamped *Praat* TextGrids (Boersma & Weenink, 2016). The annotations include linguistic features of speech in 8 different tiers: phrase, word, syllable, segment, tone, pitch, part of speech, and lemmata. Additionally, the files are annotated for non-linguistic information—metadata—such as various speaker characteristics, date and place of the recording, and the language of the interview. After receiving intensive training, a total of six annotators carried out all the annotations and transcriptions. Inter-annotator and intra-annotator reliabilities were calculated and are reported to have differed considerably based on the complexity of each annotation task (Gut, 2012).

Corpus Use

The corpus is available for free download at https://sourceforge.net/ as well as for online access as part of the language archive at https://corpus1.mpi.nl/. The corpus is located on the side menu under “TLA corpora > donated corpora.” Users can download the German and the English corpora separately. Each package is composed of all the sound files along with the annotations and metadata in *Praat* TextGrid and *XML* formats, resulting in 3 data files for each recording. The sound files can be played using any audio player application, but *Praat* is required for accessing the annotations along with the sound. In order for that, users need to open both the sound file and the *Praat* TextGrid file for each recording. After opening an annotated file in *Praat,* users can see 8 types of annotations at the phrase, word, syllable, segment, tone, pitch, part of speech, and lemma level. These tiers are not necessarily lined up in the same order, and some recordings do not include the last two tiers of annotation. Figure 1 illustrates a portion of an annotated file opened in *Praat* with all the 8 tiers of annotation for the sentence *I think it was very helpful for me*.

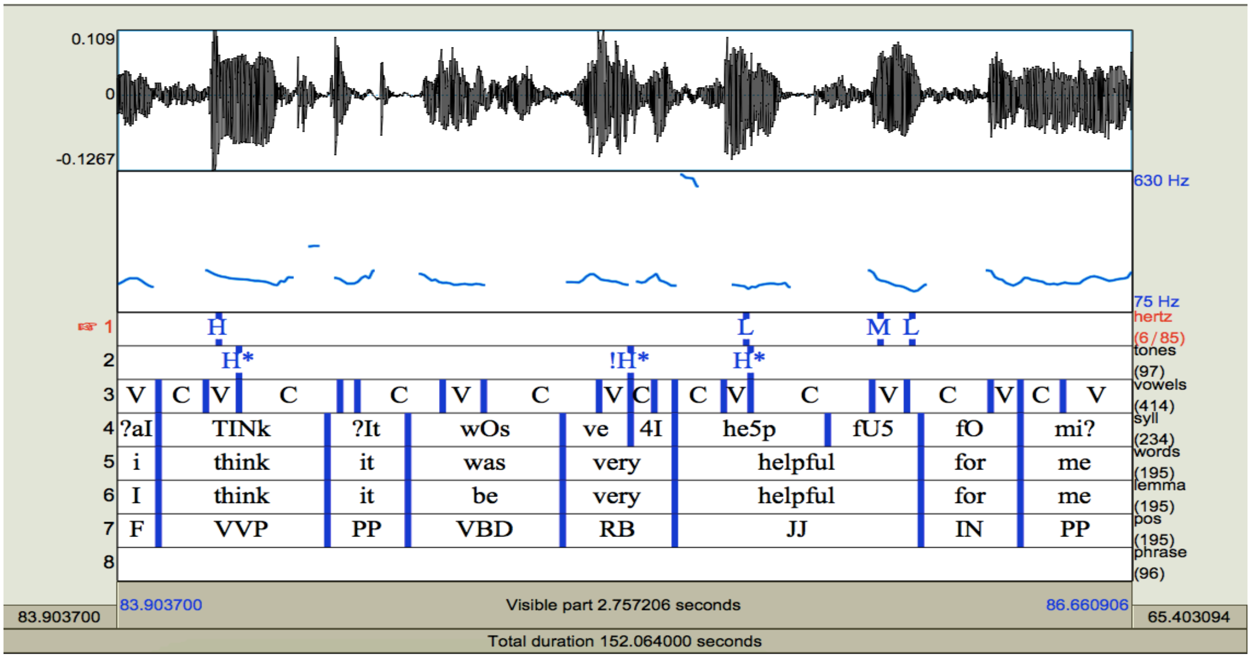
0.109
0
-0.1267
630 Hz
75 Hz
hertz
(6/85)
1 H L M L
2 H* !H* H*
tones
(97)
3 V C V C C V C V C C V C V C V C V
vowels
(414)
4 ?aI TINk ?It wOs ve 4I he5p fU5 fO mi?
syll
(234)
5 i think it was very helpful for me
words
(195)
6 I think it be very helpful for me
lemma
(195)
7 F VVP PP VBD RB JJ IN PP
pos
(195)
8
phrase
(96)
83.903700
83.903700
Visible part 2.757206 seconds
86.660906
65.403094
Total duration 152.064000 seconds

Figure 1. Sample Annotated Speech in the LeaP Corpus.

At the phrase level, in addition to the marked “quasi-intonation phrases,” non-speech events such as laughter, noise, or breath are marked (Gut, 2012). The word tier includes the beginning and end of each word along with a manual transcription. At the syllable level, beginning and end of syllables are marked, while SAMPA symbols are used for a broad transcription of the syllable including only a few articulatory and coarticulatory phenomena such as aspiration, unreleased stops, and nasalization (Gut, 2012). At the segment level, vocalic intervals, consonantal intervals, and pauses are annotated. The tone tier includes annotations for pitch accents and boundary tones using a modified version of ToBI transcription system that accounts for phonetic—rather than phonological—realizations of tones (Gut, 2012). The pitch annotation tier includes markings for initial high pitch, final low pitch, pitch peaks, and pitch valleys. The two additional tiers have been automatically annotated and provide users with the part of speech and the lemma for each word.

In addition to manual search in the data files, Gut (2009) states that analysis of the corpus is also possible through *TASX* corpus browser. *TASX* is an XML-based data format specifically designed for this corpus. Browsing this type of data is possible using a set of scripts that can be accessed upon request sent to the corpus developer. According to Gut (2009), the user-friendly environment of *TASX* allows for searching and browsing in the data, running some statistical analyses, and converting to and from different file formats commonly used in phonetic research.

Besides browsing *TASX* data files, Gut (2012) states that the files can be converted to *XML*-annotated *NITE* format to be used within the *NXT* search tool, *NXT Search*, available at http://groups.inf.ed.ac.uk/nxt/index.shtml. *NXT* provides tools and libraries that enable "native representation, manipulation, query and analysis of multimedia language data" (Kilgour, 2017). This tool enables attribute tests as well as structural and temporal relations. For instance, searches can be done in words with a syllable containing a specific vowel (Slavianova, 2007) or for pitch accents on non-content words in non phrase-final position (Gut, 2014b). In the user manual, downloadable with the LeaP package, Gut (2014b) refers to the LeaP database as another tool for searching the corpus. This database, she states, enables easy generation of subcorpora. She also mentions that a user interface for online use of the corpus is under development.

Evaluation

In line with the purpose of the LeaP project, the LeaP corpus is expected to focus on learners' acquisition of prosody and learner characteristics that influence this process. Prosody is concerned with "parameters such as duration, intensity, and f_0 that contribute in various combinations to the production and perception of stress, rhythm and tempo, lexical tone, and intonation of an utterance" (Fletcher, 2012, p. 523). The LeaP corpus provides annotations for several prosodic features such as pitch measurements, pitch accents, and boundary tones. Parameters like fundamental frequency (f_0) and other vowel quality measures such as f_1, f_2, and f_3 can also be accessed through scripts that can be run on the corpus files. Opening the TextGrids along with the audio files in *Praat* will also enable users to access features such as vowel duration and energy level. Accounting for all these suprasegmental features, the corpus allows for extracting information about a whole variety of prosodic characteristics of learner language.

However, there exists an arguably significant problem with respect to the usability of the corpus. While the user manual introduces and provides links to three tools for semi-automatic searching and browsing of the corpus (*TASX* corpus browser, *NXT Search,* and the LeaP database), none of these tools seem to be publicly available. I personally contacted the corpus developer, Ulrike Gut, and she kindly shared the *TASX* browser scripts as well as scripts for converting *TASX* files into XML, required for using *NXT* search. Nevertheless, some expertise in programming in *Perl* language seems to be needed for working with these tools. Without such expertise, users are left with the rather inconvenient option of manually looking at the data, using the *Praat* TextGrid files along with the sound files. Needless to say, manually opening and reading each file in the corpus requires an extensive amount of time and does not provide a reliable approach to analyzing the data. As Rohlfing et al. (2006) have mentioned, readability of annotations through a limited number of tools is a drawback for multimodal annotation tools. In the case of this corpus, not only are the annotations not conveniently and accurately searchable, but also only *Praat* can be used for manual searching of the data.

Provided that the user has the necessary expertise to work with Perl scripts, the *NITE* tool, as one method to browse this corpus, would still face criticism. Slavianova (2007) has voiced concerns over the extensive amount of time the tool will require if all data files of the corpus are loaded into it. A corpus-wide query seems almost impossible as the *NITE* tool has not been designed to process large amounts of data in a reasonable amount of time. In response to these shortcomings, Slaviaona (2007) has developed the LeaP database as an alternative browsing tool for this corpus. However, the LeaP database is also not publicly available.

Apart from usability, reliability of the annotations of the LeaP corpus could also be improved. As Gut (2009) has reported, perfect inter-annotator and intra-annotator agreements were only achieved at the word, segment, and pitch tiers, whereas the syllable and tone tiers showed very low reliabilities. Gut attributes the low reliability for the syllable tier to the fact that the annotators had to carry out syllable segmentation and transcription simultaneously. This defect could be addressed by simply separating the two tasks at the syllable level into two different ones. Gut (2009) has further reported that the highest intra-rater reliability was achieved by only one of the six annotators, who had had previous experience with prosody annotation. She also associates the low reliability values for the inexperienced annotators to their lack of experience with and/or extensive training in annotation. While these weaknesses are acknowledged by the corpus creator, further improvements could be made in this regard by having experienced annotators check and recheck the annotations.

Despite all the aforementioned shortcomings and its relatively small size compared to most English learner corpora, the LeaP corpus offers rich annotations useful for a wide range of studies on L2 prosody. In fact, refuting the consideration of sample size as the most important factor in achieving representativeness in a corpus, Biber (1993) maintains that emphasis should instead be placed on the range of text types and linguistic distributions. The LeaP corpus achieves this goal to a sufficient degree by accounting for different speech types, diversity in L1s, and guided and unguided pronunciation training the speakers had received. Not only are these factors among the most frequently studied in phonetics research (Colantoni, Steele, & Escudero, 2015), but also considering the issue of practicality in spoken corpus development (Reppen, 2010; Adolphs & Knight, 2010), the amount of data and type of annotations in the LeaP corpus are adequate for it to be utilized in addressing many questions on the relationship between L2 prosody and different learner characteristics, linguistic contexts, or speech events. The LeaP corpus can be a starting point for filling what Colantoni, et al. (2015) refer to as the "theoretical lacuna" existing in L2 speech learning models with regard to prosody.

REFERENCES

Adolphs, S. & Knight, D. (2010). Building a spoken corpus. In M., McCarthy & A. O'Keeffe, *The Routledge Handbook of Corpus Linguistics*. New York, NY: Routledge.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4). 243-257.

Boersma, P. & Weenink, D. (2016). *Praat: doing phonetics by computer*. Retrieved from http://www.fon.hum.uva.nl/praat/

Colantoni, L., Steele, J., & Escudero, P. (2015). *Second language speech: Theory and practice.* Cambridge, UK: Cambridge University Press.

Fletcher, J. (2012). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.) *Blackwell handbooks in linguistics: The handbook of phonetic sciences* (pp. 523-602). Hoboken, GB: Wiley-Blackwell.

Gut, U. (2009). *English Corpus Linguistics: Vol. 9. Non-native Speech: A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Frankfurt, DE: Peter Lang.

Gut, U. (2012). The LeaP corpus: A multilingual corpus of spoken learner German and learner English. In T. Schmidt and K. Wörner (Eds.) *Hamburg studies on multilingualism: Vol. 14. Multilingual corpora and multilingual corpus analysis* (pp. 3-23). Amsterdam, NL: John Benjamins Publishing Company.

Gut, U. (2014a). The LeaP corpus: A phonetically annotated corpus of non-native speech. Retrieved from: https://sourceforge.net/projects/leapcorpus/files/?source=navbar

Gut, U. (2014b). *The LeaP corpus Manual.* Retrieved from: https://sourceforge.net/projects/leapcorpus/files/?source=navbar

Kilgour, J. (2017, April 10). *NITE XML toolkit homepages*. Retrieved from http://groups.inf.ed.ac.uk/nxt/index.shtml

Reppen, R. (2010). Building a corpus: What are the key considerations? In M. McCarthy & A. O'Keeffe. *The Routledge handbook of corpus linguistics* (pp. 59-65). New York, NY: Routledge.

Rohlfing, K., Loehr, D., Duncan, S., Brown, A., Franklin, A., Kimbara, I., Milde, J.-T., Parril, F., Rose, T., Schmidt, T., Sloetjes, H., Thies, A., & Wellinghoff, S. (2006). Comparison of multimodal annotation tools: Workshop report. *Discourse and Conversation Analysis*, 7. 99-123. Retrieved from: www.gespraechsforschung-ozs.de

Slavianova, E. (2007). *The LeaP corpus - Generating a relational database for linguistic query support* (Unpublished master's thesis). Institute of Computer Science, University of Freiburg.