Designing a Bilingual Speech Corpus for French and German Language Learners: a Two-Step Process

Camille Fauth, Anne Bonneau, Frank Zimmerer, Jürgen Trouvain, Bistra Andreeva, Vincent Colotte, Dominique Fohr, Denis Jouvet, Jeanin Jügler, Yves Laprie, et al.

Designing a Bilingual Speech Corpus for French and German Language Learners: a Two-Step Process

Camille Fauth[2], Anne Bonneau[2], Frank Zimmerer[1], Jürgen Trouvain[1], Bistra Andreeva[1], Vincent Colotte[2], Dominique Fohr[2], Denis Jouvet[2], Jeanin Jügler[1], Yves Laprie[2], Odile Mella[2], Bernd Möbius[1]

[1]Saarland University – Germany

[2]LORIA, Nancy – France

bonneau@loria.fr

trouvain@coli.uni-saarland.de

Abstract

We present the design of a corpus of native and non-native speech for the language pair French-German, with a special emphasis on phonetic and prosodic aspects. To our knowledge there is no suitable corpus, in terms of size and coverage, currently available for the target language pair.

To select the target L1-L2 interference phenomena we prepare a small preliminary corpus (corpus1), which is analyzed for coverage and cross-checked jointly by French and German experts. Based on this analysis, target phenomena on the phonetic and phonological level are selected on the basis of the expected degree of deviation from the native performance and the frequency of occurrence. 14 speakers performed both L2 (either French or German) and L1 material (either German or French). This allowed us to test, recordings duration, recordings material, the performance of our automatic aligner software.

Then, we built corpus2 taking into account what we learned about corpus1. The aims are the same but we adapted speech material to avoid too long recording sessions. 100 speakers will be recorded.

The corpus (corpus1 and corpus2) will be prepared as a searchable database, available for the scientific community after completion of the project.

Keywords: speech corpus, phonetics, language learning

1. Introduction

We present the design of a corpus of native and non-native speech for the language pair French-German, devoted to an in-depth analysis of both segmental and prosodic aspects of the non-native production of these languages. To our knowledge there is no suitable corpus, in terms of size and coverage, currently available for the target language pair. Our design of such a corpus has three aims. (i) We will bring to the research community two non-native corpora for the French-German language pair, whereas most studies have focused on English (cf. website on "Learner corpora around the world"). (ii) The corpora will be informed by in-depth phonetic knowledge to predict the types of errors made by French and German learners. (iii) The corpora can further be used by the research community for the recognition of non-native speech, which is notoriously difficult (see e.g. Goronzy et al. 2001, van Doremalen et al. 2009, Bouselmi et al. 2011).

From this entire corpus, we created two different corpus (corpus1 and corpus2) which will be used at different stage, corpus1 being the preliminary corpus and corpus2 being the final one. Corpus1 has been recorded to test hypotheses about interferences between L1 and L2 of the speakers, as well as to test the performance of our automatic aligner software, especially with respect to L2 recognition. First results obtained from the analysis of corpus1 were incorporated into the design of corpus2.

2. Design of Corpus1

To select the target L1-L2 interference phenomena, we prepared corpus1, which was analyzed for coverage and cross-checked jointly by French and German experts. From this analysis, target phenomena on the phonetic and phonological levels were selected on the basis of the expected degree of deviation from the native performance and the frequency of occurrence.

The material covered for both languages consists of: (i) a phonetically rich design comprising all phonemes in relevant contexts, to achieve a reliable assessment of the entire phonemic inventory for each speaker; (ii) the most important phenomena in the phonetics and prosody of French and German as a foreign language, respectively (e.g., vowel quantity, consonantal articulation and word stress); (iii) phonological processes and alternations (e.g., final devoicing); (iv) minimal pairs.

Each speaker had to perform four tasks in both languages L1 and L2. In the first task, the speaker was asked to read aloud a sentence (25 different sentences). In the second, the speaker heard the sentence pronounced by a native speaker and then was asked to read the same sentence aloud (25 different sentences). The third task, the focus condition, consisted of sentences that had to be produced as an answer to a question pronounced by a native speaker, including broad and narrow focus conditions. The part of the sentence intended to be in focus, was additionally indicated by upper-case letters. Six sentences were presented, each of which also occurred with different constituents in focus, that correspond to 24 different sentences (6 sentences X n focus conditions). The fourth condition was to read aloud two short stories: *The three little pigs* and a more technical text (about ten sentences long – 2 minutes recordings).

2.1. Subjects and Recordings

For corpus1, we recorded seven native subjects for each language (see Table 1). We included beginners (A2 level according to the CEFR) as well as advanced (C1 level) second language learners. Among the beginners we aimed at teenage learners with two to three years of L2 instruction in school, and university students. Informed consent was obtained from subjects (and parents for the children) allowing us to use the recorded data for scientific purposes. For corpus1 German and French subjects were recorded in their L2 (FG and GF respectively) as well as in their native language. They all recorded the non-native part first, and then they also produced the sentences in their L1 (FF and GG respectively – see Table 1).

# subjects	L1	L2	level	age
4	F	G	beginners	18-30 years old
2			advanced	
1			beginners	15-16 years old
3	G	F	beginners	18-30 years old
2			advanced	
2			beginners	15-16 years old

Table 1: Groups of subjects recorded for corpus1 pooled across L1 (F=French, G=German), L2 (G=German F=French,), level of proficiency and age range.

High-quality recordings were made, using the software JCorpusRecorder (see raweb.inria.fr) on a Windows laptop. As recording device, we used a headset microphone (AKG C520) and an Audiobox (M Audio Fast track). The gain was automatically controlled during the recording to avoid clipping. Sometimes, the gain was also manually adjusted. The subject was seated in a quiet room. Recording sessions were carefully monitored for consistent quality. Subjects could listen to their recordings after each sentence and decide whether they wanted to make a new recording or take the one they just performed. A recording session lasted between 50 and 75 minutes.

2.2. Annotation and Automatic Alignment

Annotation was performed in two stages:

First, the entire corpus1 was automatically segmented and annotated by our speech-text alignment tool (Jouvet et al. 2011, Fohr & Mella 2012) via the use of a two-step approach for automatic phone segmentation. The first step consisted in determining the phone sequence that best represents the learner's utterance. This was achieved by force-aligning the learner's utterance with a model representing pronunciation variants of the sentence. It was crucial to consider both native and non-native variants. In this step, detailed context-dependent acoustic Hidden Markov Models (HMMs) were used, with a rather large number of Gaussian components per mixture density. This kind of detailed acoustic model was the one that provided the best performance in automatic speech recognition.

The second step consisted in determining the phone boundaries. This was also achieved by means of a forced alignment process, but this time, the sequence of phones was known (as determined in the first step), and context-independent acoustic phone models with only a few Gaussian components per mixture density were used because they provide a better temporal precision than detailed acoustic models (Toledano & Gomez 2003). For training the models in both forced alignment steps, the speech of native and non-native speakers was used, either directly or by MLLR (Maximum Likelihood Linear Regression) adaptation. Furthermore, having speakers produce both native and non-native speech allows for better adaptation.

In the second stage, the entire corpus1 was checked with respect to the orthographic transcription. It was manually checked at the levels of phones and words (phonetic transcription) and corrections were made if necessary. This procedure enables us to investigate how automatic forced alignment performs on non-native speech (Fohr & Mella 2012) and to improve the models.

This manual control and (re)labeling was performed using the software PRAAT (Boersma & Weenink, 2009) from which it was possible to add various lines of information (or tiers) for an alignment at the sentence, the word and phoneme level (see Figures 1 & 2).

The automatic alignment generated several Tiers which two are exactly the same: the Align and the Real Tier. The annotator was responsible for verifying whether the aligner correctly recognized the segments that were produced and also, whether the boundaries between the segments were correct. Otherwise, the labels were changed, or the boundaries were move in the so-called RealTier, using the acoustic signal and its spectrogram associated. If sounds had been changed, compared to a canonical pronunciation, annotators had in some cases also the choice to use a set of diacritical labels adapted and extended from Kiel corpus (IPDS, 1994, Kohler et al., 1995; or see http://www.ipds.uni-kiel.de/forschung/kielcorpus.de.html). This list may be further extended as the project progresses. The annotator also had a comment line (CommentTier) to add further notes. However, this line is used sparingly. It is also useful for remarks concerning the entire sentence (in particular prosody).

As you can see on Figure 1, the Realtier has been changed at the phonetic level to indicate that the expected sound /@/ has been pronounced as /ø/ (noted /2/). The annotator also signaled the presence of glottalisation, which is coded by the symbol /q/ and preceded by the symbol /-/ used in case of insertion. On Figure 2, we can see that there is an important mistake concerning the end of the automatic boundary of the consonant /Z/ and therefore the boundary has been moved. Another correction applied to the same sound, signaled, though the use of the diacritics (_0), that it has been devoiced. This devoicing is due to the influence of the speaker's L1, voiced obstruent consonants /Z,z,v,b,d,g/ being devoiced in word final position in German.

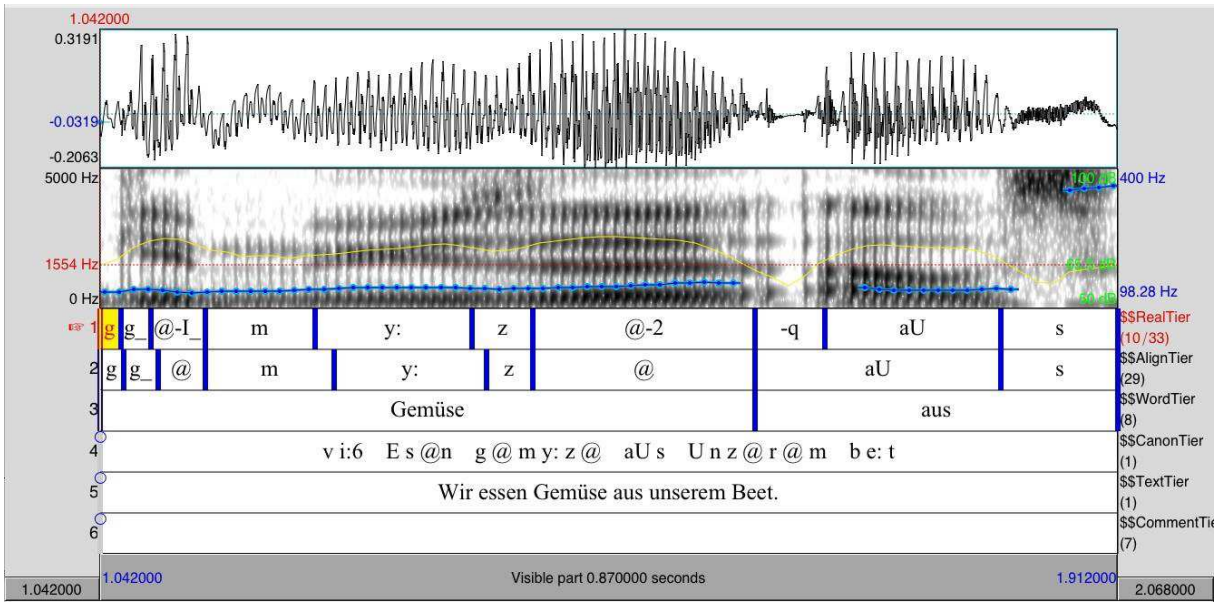

Figure 1: Automatic alignment and corrections of a German sentence produced by a French speaker learning German (FG) *"Wir essen Gemüse aus unserem Beet" – "We eat vegetables from our patch"*

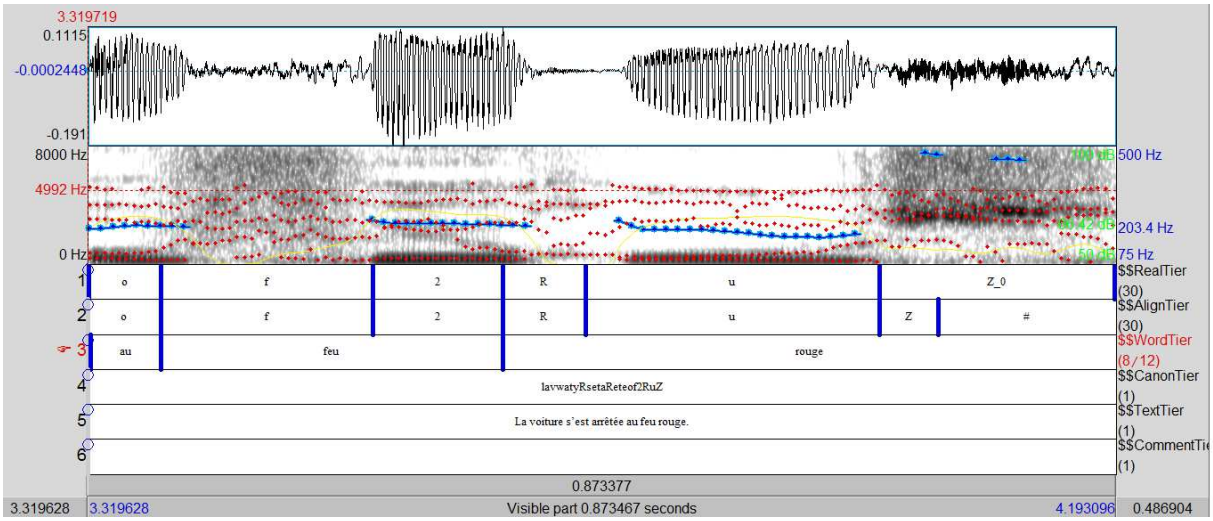

Figure 2: Automatic alignment and corrections of a French sentence produced by a German speaker learning German (GF) "*La voiture s'est arrêtée au feu rouge" – "The car stopped at the red light*"

3. Conclusions Concerning Corpus1

The purpose of corpus1 was to verify the relevance of the selected phonetic phenomena and detect possible unexpected phenomena and/or deviations. The data suggested that several observed phenomena are crucial for improving the pronunciation of L2, for both German and French native speakers, such as vowel quantity, final devoicing and the aspiration or non aspiration of stop consonants.

Concerning vowel quantity, we found that French native speakers inconsistently produced tense (long) and lax (short) vowels in their L2, whereas German speakers tend to maintain vowel length differences (vowel quantity) in French where such as difference does not exist. Concerning final devoicing of obstruents, we found that German speakers (beginners and advanced) tend to maintain this devoicing when speaking French (Fauth and Bonneau, 2014) whereas most French speakers does not devoice obstruents in final position.

Also, the analysis of voiceless stops indicated that German speakers produced stop segments in French (GF part of the corpus) as they do in German, without adjusting for the French standard production with less aspiration, whereas the opposite phenomenon is observed for French speaking German.

Among other interesting phenomena, as expected, there were influences of orthography, as for example, the production of the word *loup* ('wolf') where the final consonant was mistakenly produced by many native German speakers. French speakers showed some mistakes in the production of the German glottal fricative /h/. On a general level, French learners of German produced fewer /h/s than native German speakers. For the French speakers, the data revealed an interaction with task, when repeating sentences, the /h/ was produced significantly more often (about 77%) than in a pure reading task (about 52%), despite the influence of orthography, where <h> is written in both cases (Zimmerer & Trouvain, submitted).

These results emphasize the possible usefulness of having an audio model (or 'golden speaker') in language teaching contexts. Interestingly, in cases where French speakers did produce /h/, the segment seemed to be hyper-articulated. French speakers' /h/s were longer (in relation to the rest of the word) and articulated with more intensity (in relation to the rest of the word) than the native speakers' productions.
Based on our findings, we were able to deduce important aspects in the production of L1 and L2, in order to further improve for instance the contexts in which the phenomena occurred in our data. We slightly modified some sentences in the preparation phase of corpus2, which is discussed in the next section.

4. Corpus2

4.1. Design

The general idea, to obtain productions from speakers both in their native language and their L2, is identical to corpus1. Again, speakers will first produce the L2, and then they will perform the L1 tasks.
We learned from the recordings of corpus1 that the total recording time should not exceed one hour per subject, to avoid the effects of vocal fatigue, especially for beginners. Therefore, we adjusted the overall size of the corpus.
The number of sentences for reading and repetition has been slightly increased (from 50 to 60), to be able to exploit phonetic and phonological phenomena of interest in more depth. Furthermore, the presentation has been changed to some extent. While the order of the tasks was kept constant, the sentences of every task are now randomly presented to the speakers. 31 out of 60 sentences are now in the reading task, 29 were heard before reading.
The sentences have been built to include phonetic target phenomena mentioned for corpus1. But if for corpus1, our approach was more general and wanted to cover all the difficulties that could be encountered in learning German or French as L2, for corpus 2, we want to focus on certain phenomena and finely control the context in which they may appear. These phenomena have been highlighted by the analysis of the corpus1. This approach led us to "select" phenomena rather than others. These selected phenomena have been carefully controlled, such as the pronunciation of stop consonants.
For stops consonants, the sentences been adapted to create quasi-minimal pairs also in the sentence structure, when this was possible.
For the pronunciation of stop consonant, some of these sentences are:Braucht man in einer Band einen **Bass**? ('Do you need a bass in the band?')
Braucht man zum Reisen einen **Pass**? ('Do you need a passport to travel?')
Wieviele Kinder fahren mit dem **Rad**? ('How many children are driving the bike?' – in this example, the final /d/ is devoiced - *Auslautverhärtung*)
Braucht sie deinen **Rat**? ('Does she need your advice?')
Bring bitte noch Teller in den **Garten** ('Bring the plates to the garden, please.')
Auch Täler findet man in **Karten**. ('You find valleys in maps, too.')
Le parrain a quitté le **bar** ce matin. ('The godfather left the bar this morning')
Le garçon mange deux **parts** de gâteau. ('The boy eats two pieces of cake.')
L'abeille n'a pas de **dard** sur le corps. ('The bee does not have a stinger on the body.')
Les poids ont fait la **tare** sur la balance. ('The weighs have made the tex weight on the scale.')
Le train a quitté la **gare** de Paris. ('The train has left the station of Paris.')
Le garçon a pris le **car** à Berlin. ('The boy has taken the bus to/in Berlin.')

German, like English, has long and short vowels. Short vowels are usually produced when they precede a sequence of two consonants and not starting with a /h/. Moreover, the vowel quality in German is also related to duration. The following sentences are examples with minimal pairs that focused on this phenomenon.
Ich wüsste nicht, wie der schnellste Weg nach **Polen** ist. ('I would not know the shortest way to Poland)
Im Frühling fliegen **Pollen** durch die Luft. (In springtime, pollen are flying through the air)
Wir essen Gemüse aus unserem **Beet**. (we eat vegetables from our vegetable patch)
Bei hohem Fieber legst du dich ins **Bett**. (With high fever, you lie in bed)

In the following sentence, the « o » of Rome is pronounced /o/ by German and /ɔ/ by French.
Ils ont acheté leur armoire **à Rome.** ('They bought their wardrobe in Rome.')

Some other phenomena are not investigated systematically, but remain present in corpus 2 (such as the production of post-vocalic /R/ produced by GF). In addition compared to corpus1, we included numerals and abbreviations. These will increase the cognitive load for some of the sentences and maybe show interesting differences between beginners and advanced learners. Besides, they are also intended to reveal further segmental variations, and especially numerals which occur in everyday situations are different in their structure in French and German. For instance, in German, the number "*96*" is literally "*6 and 90*", whereas in French, it is "*4 20 16*".
As a consequence of the analysis of corpus1, the number of sentences in the focus condition was reduced to two (instead of six). This reduces the total number of sentences, which leads to an overall size of corpus2 that is smaller than corpus1, but it also minimizes sentences that are rather not very natural for speakers. Focus sentences occurred in up to four focus conditions that have to be repeated, including contrastive and non-contrastive focus. If there are too many sentences, such repetition effects may lead to somewhat unnatural utterances.
For this task, sentences were created that would allow for a smooth detection and automatic tracking of the fundamental frequency, which is one of the principal clues in this type of task. We also selected simple sentences in terms of syntactic structure and vocabulary. As far as possible, they also have the particularity to contain as many voiced sounds and sonorant consonants as possible, which should facilitate the automatic detection of the fundamental frequency. This part therefore is different in corpus2 compared to corpus1.
Concerning the text condition, we selected only one story,

the three little pigs for corpus2. We chose this story because it is well-known, its style is very casual and the words are repeated several times, which facilitates testing the consistence of deviant productions. We concentrated on one story to decrease the recording time. Reading the story, especially in the L2, was rather time consuming, because speakers were reading the stories several times to familiarize themselves with the text before they were recorded. Also, we know from the recordings of corpus1 that subjects tend to rerecord the stories. The length of the text increases the likelihood of mistakes, which made subjects feel unsatisfied with their renditions.

4.2. Subjects and Recordings

We plan to record a total of 100 subjects for corpus2 (see Table 2). This will comprise 40 university students who are beginners (A2 level) as well as 40 advanced second language learners (C1 or C2 level) and 20 teenagers. The different groups are defined as in corpus1. Speaker gender will be balanced in each of the six groups. So far, 80% of the speakers have been recorded.

# subjects	L1	L2	level	age
20	F	G	beginners	18-30 years old
20			advanced	
10			beginners	15-16 years old
20	G	F	beginners	18-30 years old
20			advanced	
10			beginners	15-16 years old

Table 2: Groups of subjects that will be recorded for corpus2 pooled across L1 (F=French, G=German), L2 (G=German F=French,), level of proficiency and age range.

The recording conditions of corpus1 were assessed to be satisfying. Therefore, they will be identical for the recordings of corpus2.

4.3. Annotation and Automatic Alignment

The entire corpus will be automatically segmented and annotated by our speech-text alignment tool. We want to improve the automatic alignment by including a third step to refine the boundaries occurring between specific phonetic classes, by computing and using ad-hoc acoustic cues, and associated decision processes (Adell et al. 2005).

A part of the corpus (50-60%) will be manually checked at the levels of phones, words (orthographic transcription), and sentences (orthographic transcription), and corrected if necessary. The complete corpus (corpus1 and corpus2) will be prepared as a searchable database, available for the scientific community after completion of the project. Researchers will get access to the corpus after sending an application form, specifying the terms and conditions for the use of the corpora.

5. Conclusions

A bilingual speech corpus consisting of speech of French learners of German (FG and FF) and German learners of French (GF and GG) speaking both in the L2 as well as in the L1, represents an optimal starting point for detailed phonetic and phonological analyses on the segmental as well as on the prosodic level. In addition, such a phonetic learner corpus provides a rich source for first-hand information and illustration for the interested public, e.g. foreign language teachers. This corpus is also the prerequisite for the development of a computer-assisted language learning software by adapting content, feedback and exercises to individual learners in the speech dimension of a foreign language

6. Acknowledgments

This work has been supported by an ANR/DFG Grant “IFCASL” to the Speech Group LORIA CNRS UMR 7503 – Nancy France and to the Phonetics Group, Saarland University – Saarbrücken Germany, 2013 – 2016.

7. Bibliographical References

Adell, J., Bonafonte, A., Gómez, J. A., & Castro, M. J. (2005). Comparative study of automatic phone segmentation methods for TTS. In Proceedings of ICASSP, pp. 309-312.

Andreeva B. and Barry W. (2012). Fine phonetic detail in prosody. Cross-language differences need not inhibit communication. In: O. Niebuhr (ed.), Prosodies - context, function, and communication. Berlin/New York: de Gruyter. 259-288

Bartkova, K. and Jouvet, D. (2007). On using units trained on foreign data for improved multiple accent speech recognition. Speech Communication 49, pp. 836–846.

Barry W.J., Andreeva, B. and Steiner I. (2007). The Phonetic Exponency of Phrasal Accentuation in French

Bonneau, A. and Colotte, V. (2011). Automatic feedback for L2 prosody learning. In Intech “Speech Technology”. Available from: http://www.intechopen.com/books/speech-and-language-technologies/ automaticfeedback-for-l2-prosody-learning

Bonneau, A. & Laprie, Y. (2008). Selective acoustic cues for French voiceless stop consonants. Journal of the Acoustical Society of America 123, pp. 4482-4497.

Best, C. T. 1995. A direct realist view of cross-language speech perception. Cross-language studies of speech

Boersma, P., & Weenink, D., (2009). Praat : Doing phonetics by computer (version 5.1.20) [Computer program] http://www.praat.org/.

Bouselmi, G., Fohr, D. & Illina, I. (2011). Multilingual recognition of non-native speech using acoustic model transformation and pronunciation modeling. International Journal of Speech Technology 15(2), pp. 203-213.

van Doremalen, J., Strik, H. & Cucchiarini, C. (2009). Optimizing non-native speech recognition for CALL applications. Proc. Interspeech, Brighton, pp. 592-595.

Dupoux, E., Sebastián-Gallés, N., Navarrete, E. & Peperkamp, S. (2008). Persistent stress 'deafness': The caseof French learners of Spanish. Cognition 106.682-706.

Fauth C., Bonneau A. (2014). L1-L2 interference: the case of devoicing of French voiced obstruents in final position by German learners - Pilot study. International Workshop on Multilinguality in Speech Research: Data, Methods and Models. Dagstuhl - Germany

Flege, J. E. (1988). Effects of speaking rate on tongue position and velocity of movement in vowel production. Journal of the Acoustical Society of America 84.901 – 916.

Flege, J. E., MacKay, I. R. A. & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. Journal of the Acoustical Society of America 106.2973-2987.

Fohr, D. and Mella, O. (2012). CoALT: A software for comparing automatic labelling tools, Proceedings of LREC

Goronzy, S., Sahakyan, M. & Wokurek, W. (2001). Is non-native pronunciation modeling necessary? Proc. Eurospeech, Aalborg, pp. 309-312

Hirschfeld, U. and Trouvain, J. (2007). Teaching prosody in German as a foreign language. In: Trouvain, J. & Gut, U. (eds) Non-Native Prosody. Phonetic Description and Teaching Practice. (Trends in

IPDS. (1994). The Kiel Corpus of Spontaneous Speech. Kiel: IPDS.

Kohler, K. J., Pätzold, M. & Simpson, A. P. (eds) (1995). From scenario to segment - the controlled elicitation, transcription, segmentation and labelling of spontaneous speech. Kiel: IPDS. Linguistics. Studies and Monographs [TiLSM] 186) Berlin/New York: Mouton de Gruyter. 171-187

Jilka, M. (2007). Different manifestations and perceptions of foreign accent in intonation. Non-Native Prosody-Phonetic Description and Teaching Practice, ed. by J. Trouvain & U. Gut, 77-96. Berlin: Mouton De Gruyter.

Jouvet, D., Mesbahi, L., Bonneau, A., Fohr, D., Illina, I. & Laprie, Y. (2011). Impact of pronunciation variant frequency on automatic non-native speech segmentation. Proc.5th Language & Technology Conference (LTC'11), Poznan, pp. 145-148.

Kingston, J. (2003). Learning foreign vowels. Language and Speech 46.295-349.

Toledano, D. and Gomez, L. (2003). Automatic Phonetic Segmentation. IEEE Trans. on Speech and Audio Processing, v11, n6, pp. 617—625.

Zimmerer, F., Jügler, J., Andreeva, B., Möbius, B. & Trouvain, J. (2014). Too cautious to vary more? A comparison of pitch variation in native and non-native productions of French and German speakers. Proc. Speech Prosody 7, Dublin, Ireland.

http://www.uclouvain.be/en-cecl-lcworld.html
http://raweb.inria.fr/exploraweb/static/2011/parole/uid63.html
http://www.etsglobal.org/Fr/Eng/Research/CEFR
http://www.ipds.uni-kiel.de/forschung/kielcorpus.de.html