

Konstruktion eines phonetisch-phonologischen Lernerkorpus für das Sprachenpaar Französisch-Deutsch



*Frank Zimmerer^a, Jürgen Trouvain^a, Yves Laprie^b,
Bernd Möbius^a, Bistra Andreeva^a, Anne Bonneau^b,
Vincent Colotte^b, Camille Fauth^b, Dominique Fohr^b,
Denis Jouvet^b, Odille Mella^b und Jeanin Jügler^a*

^aPhonetikinstitut, Universität des Saarlandes, Saarbrücken

^bÉquipe Parole, LORIA, Nancy

Zürich, 11.10.2013

> 0. Überblick über heutigen Vortrag



1. Warum?
2. Wie?
3. Erste Daten / Beispiele
4. Ausblick



Hintergrund:

Phonetik und Phonologie einer Fremdsprache (L2) sind aus verschiedenen Gründen schwer erlernbar, das phonetisch/phonologische System der Muttersprache (L1) hat einen großen Einfluss auf die **Produktion und Perzeption** der L2 (z.B. (und noch viele, viele mehr), Barry et al., 2007; Best, 1995; Best & Tyler, 2007; Bohn & Best, 2012; Darcy & Krüger, 2012; Dupoux et al., 1999; Flege, 1991, 1995; Flege et al., 1999; Hirschfeld & Trouvain, 2007; Kingston, 2003; Schmid, 2012; Strange, 1995).

-> Viel Training ist nötig, um besser zu werden

-> Eine Möglichkeit: computerbasiertes Training: z.B., Duolingo, Rosetta Stone, ...



-> Startpunkt unseres Projektes (gefördert von DFG und ANR):

IFCASL

(Individualised feedback in computer-assisted spoken language learning)

- ◆ Französisch (L1) und Deutsch (L2)
- ◆ Deutsch (L1) und Französisch (L2)

Was sind die phonetischen und phonologischen Probleme, die in diesem Sprachenpaar auftauchen?

> 1. Warum?



Bisherige Datengrundlage ist relativ dünn:

- Persönliche und anekdotische Daten
- (Theoretische) kontrastive Gegenüberstellung
- Kaum Korpusdaten vorhanden:
 - Hauptsächlich schriftliche Korpora
 - Meistens ist (eine der) Sprache(n) Englisch
 - Sehr weniger Korpora für Sprachpaare
 - Beispiele:
 - HABLA (Hamburg Adult Bilingual LAnguage) Korpus mit bilingualen Probanden (L1: Französisch & Deutsch) (Kupisch et al. 2012)
 - IPFC-allemand (Interphonologie du Français Contemporain) with L1: Deutsch L2: Französisch (sehr fortgeschrittenes Fremdsprachlevel) (Pustka 2012)



Aufbau eines neuen Lernerkorpus D-F, F-D:

Training- und Testmaterial für ein computerbasiertes
Feedbacksystem

Daten und Analyse für phonetische & phonologische
Fragestellungen

Anforderungen an Lernerkorpus (nicht nur für unsere Zwecke)

- Qualitativ hochwertige Audio-Aufnahmen
- Annotation auf segmentalem und prosodischem Level
- Zugänglichkeit (dazu später mehr)



100 Probanden:

50 Probanden mit L1 Französisch und L2 Deutsch in Nancy

50 Probanden mit L1 Deutsch und L2 Französisch in Saarbrücken

Verschiedene Fremdsprachniveaus innerhalb der Gruppe:

10 Schüler, 14-16 Jahre, Anfängerniveau (A1/A2)

20 Erwachsene, Anfängerniveau (A1/A2)

20 Erwachsene, Fortgeschrittene (C1/C2)

Die Schüler gehen auf verschiedene Schulen (Saarland und Lothringen), die Erwachsenen in Kurse an der Universität (Saarbrücken und Nancy) oder Goethe-Institut



Einteilung in Fremdsprachniveau anhand von Zertifikat(en) & Fragebogen.

Linguistische Biographie:

Alter, L1, Wohnort (der ersten 16 Lebensjahre)

Höchster (Schul-)Abschluss

Für jede Fremdsprache (L2):

Anzahl der (Schul-) Jahre des Sprachlernens

Auslandsaufenthalte

Sprachgebrauch (Eltern, Partner, Freunde, ...)

Kurse (& Zertifikate)



Selbsteinschätzung:

Wie gut sind die Fremdsprachenkenntnisse (v.a. Aussprache, Grammatik, ...)

Motivation für das Erlernen der Sprache

Einstellung im Bezug zum Erwerb der L2

Einstellung zum Fremdsprachenlernen am Computer



Rechtliche Aspekte:

Probanden erhalten kleine „Entschädigung“ für die Teilnahme

Die eigenen Daten können auf Wunsch zugänglich gemacht werden,
Probanden bleiben im Besitz ihrer Rechte über die eigenen Daten

Die Daten können für folgende Zwecke genutzt werden:

- Wissenschaftliche Zwecke in anonymisierter Form (schriftlich und mündliche Daten)
- Zu Signalverarbeitungszwecken
- Zur Verbesserung der Sprachlernsoftware

Zugang zu den Daten:

- Die Daten werden (es sei denn, ausdrückliche Einverständnis vorhanden) nicht der breiten Öffentlichkeit zugänglich gemacht
- Wissenschaftler können für Forschungszwecke (mit Antrag) Zugang erhalten



Korpus besteht aus gelesener Sprache (keine Spontansprache)
Alle Probanden werden in L2 und L1 aufgenommen.

Die Aufnahmen werden in leiser Büroumgebung und guter akustischer Qualität vorgenommen:

- Headset-Mikrofon (fast unsichtbar für Probanden)
- Aufnahmesoftware: CorpusRecorder (in Nancy entwickelt)
- Alle Sätze werden auf dem Bildschirm dargestellt
- Ein Teil der Sätze ist reiner Lese-Task, ein anderer Teil ist ein Wiederholungs-Task
- Eine Aufnahmesession dauert 60-75 Minuten (~10 Minuten für Fragebogen, ~50 Minuten für die eigentlichen Aufnahmen)
- Bisherige Aufnahmen – „Minikorpus“



Alle Probanden produzieren pro Sprache 4 Konditionen:

1. Sätze lesen

Sätze Vorlesen, die auf dem Computerbildschirm erscheinen

2. Sätze wiederholen (Muttersprachler lesen lediglich die Sätze)

Sätze werden auditiv vorgespielt (von einem Muttersprachler aufgenommen) und sollen wiederholt werden – Fehler, die aufgrund von Orthographie entstehen, sollen minimiert werden.

3. Fokussätze

Frage wird auditiv vorgespielt, die Antwort darauf wird vorgelesen (zusätzlich wird Fokus durch GROSSBUCHSTABEN markiert)

4. Text lesen

Zwei Texte vorlesen:

- a) informativer Text (zu Berufen in der Umweltbranche)
- b) Geschichte (die drei Schweinchen)



Auswahl der Sätze und Texte, anhand (u.a.) folgender Phänomene:

1. Auf segmentaler Ebene

- Glottalplosiv [ʔ] und glottaler Frikativ [h]
- Liaison and enchaînement consonantique (z.B. *pot^t-au-feu*)
- Nasalvokale [ɛ̃, ã, õ]
- Auslautverhärtung von Plosiven und Stimmhaftigkeit von Frikativen
- Aspiration stimmloser Plosive [p^h, t^h, k^h]
- ich- und ach-Laut [ç, x]
- Schwa: Gerundet? Vollvokal?
- /r/ als Konsonant [ʀ, ʁ] vs. Vokal [ɐ]
- Vokallänge (und Qualität) [iː-ɪ, eː-ɛ, ɛː-ɛ, aː-a, oː-ɔ, uː-ʊ, yː-ʏ, øː-œ]
- Konsonantencluster (z.B. *Bratw^rüstchen*)
- Reduktionen (z.B. Lenisierungen, Assimilierungen)



2. Auf prosodischer Ebene

- Wortbetonung
- Kontrastiver Akzent
- Pitch Range

3. Unsichere Orthographie-Aussprache Verbindung

- Franz.: *plus tard* [ply(s)], *un loup* [lu(p)] Dt.: *hohem* [ho:(h)əm]

4. Internationale Wörter, Eigennamen

- Wird Frz. "énergie" dt. gelesen? ([enɛʁ'gi:] in L2=FR)
- Wird Dt. "Berlin" frz. gelesen? ([bɛr'li:] in L2=DE)

5. Fehlerhaftes Lesen (evtl. Einfluss anderer L2)

- Deutsch "Licht" [lɪçt] (Engl. "light") gelesen als [laɪt]

6. Minimalpaare, z.B. dt. „Paar-Bar“, oder „pont-bon“

7. Alle Phoneme jeder Sprache sollten mindestens einmal auftauchen

> 2. Wie?



Corpus Recorder

Item n° 1

Votre répertoire d'enregistrement : Recordings/GJJM
Prononcez et enregistrez la phrase en gras le plus naturellement possible.

La vaisselle propre est mise sur l'évier.

 Ecoute

Exit

Phrase suivante

Type a question for

Design New Slide

Slide Layout

Apply slide layout:

Text Layouts

Content Layouts

Text and Content

DE

> 2. Wie?



Überblick über die verschiedenen Tasks

Tasks	L2=FR*		L1=DE*	
	Anz. Sätze	Anz. Wörter	Anz. Sätze	Anz. Wörter
Sätze lesen	25	183	51	359
Sätze wiederholen	29	207	-	-
Fokussätze**	24	291	25	144
Text lesen***	8 + 12	154 + 205	10 + 13	127 + 215
Insgesamt	98	1040	101	845

* L1=FR, L2=DE Anz. Sätze/Wörter vice versa

** FR und DE Versionen sind sehr ähnlich

*** Übersetzung desselben Textes in beiden Sprachen



Was sind häufige Abweichungen?

- Frequenz: Manche Abweichungen erwarten wir häufiger als andere
- Hartnäckigkeit: Manche Abweichungen erwarten wir auch bei hohem L2-Niveau
- Task-abhängige Fehler? (z.B. Orthographieinflüsse)

Welche Abweichungen sind „schwerwiegender“?

- Nicht alle Abweichungen führen zu verminderter Verständlichkeit
- Nicht alle Abweichungen führen zu starkem Akzent

Welche Abweichungen können automatisch erkannt und repariert werden?

- Manche Abweichungen sind leicht erkennbar
- Manche Abweichungen können (leichter) „repariert“ werden.



Beispielsätze:

Deutsch: „In jeder Bank gibt es eine Kasse.“

- (Nasal)Vokal in „*Bank*“
- Keine Aspiration bei [k] in „*Kasse*“
- Keine Vokalisierung von /r/ in „*jeder*“
- Gerundetes Schwa in „*es, eine, Kasse*“

Französisch: „Marie a rangé ses lunettes dans son étui.“

- Nasalvokale in „*rangé, dans, son*“
- (Fehlende) Liaison bei „*son étui*“
- Eingefügte Aspiration bei „*lunettes*“
- Eingefügte Glottalverschlüsse bei „*Marie a*“ und „*son étui*“

Fokus (Dt.): „Der Tourist besucht Paris.“

(F: Der Fremde besucht Paris?) – A: *Der TOURIST besucht Paris.*

(F: Der Tourist besucht München?) – A: *Der Tourist besucht PARIS.*

(F: Der Tourist geht in Paris spazieren?) – A: *Der Tourist BESUCHT Paris.*

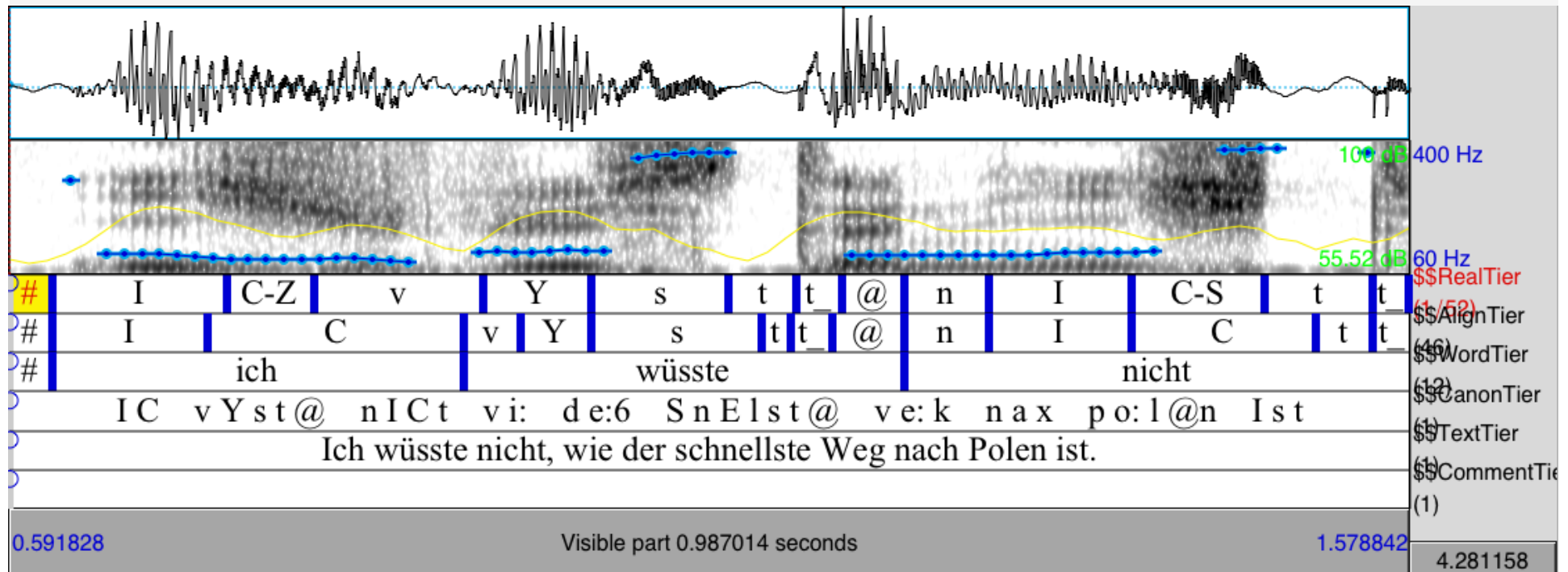
> 2. Wie? / 3. Erste Daten



Korpus besteht aus:

- 14 Probanden (7 L1: Französisch, 7 L1: Deutsch – jeweils: 2 Schüler (A1/A2), 2 Fortgeschrittene Lerner (C1) & 3 Anfänger (A1/A2))
- Audio Dateien
- Labeling Dateien (PraatTextGrids):
 - Satzebene (orthographisch und in SAMPA)
 - Wortebene „kanonische Form“
 - Phonebene (in SAMPA)
 - Best-Match von Aligner (entwickelt in Nancy)
 - Realisierte Form
 - Substitutionen
 - Einfügungen
 - Löschungen
 - Prosodische Ebene:
 - Abweichungen von Wortbetonung und Satzakzent (wenn es vorkommt)

> 3. Erste Daten





- Nicht alle Fehler sind zentral -> Konzentration auf „wichtige“ Fehler
 - L2: Deutsch
 - Vokallänge
 - Wortbetonung
 - Schwa
- Weniger „schwerwiegende“ Fehler können vernachlässigt werden (v.a. im Saarland)
 - L2: Deutsch
 - ich-Laut [ç] vs. sch-Laut [ʃ]

> 3. Erste Daten



- L2: FR – Aspiration von stimmlosen Plosiven

- "... sur la **p**etite chèvre."

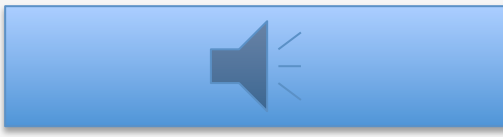


- L2: DE – Vowel length

- "... Frühling fliegen **P**ollen durch die Luft."



- "... der schnellste Weg nach **P**olen ist."





- L2: FR & D – Fokus

Petra isst einen Kuchen in der Bäckerei.

Petra isst einen Kuchen in der Bäckerei.

Petra isst einen Kuchen in der Bäckerei.



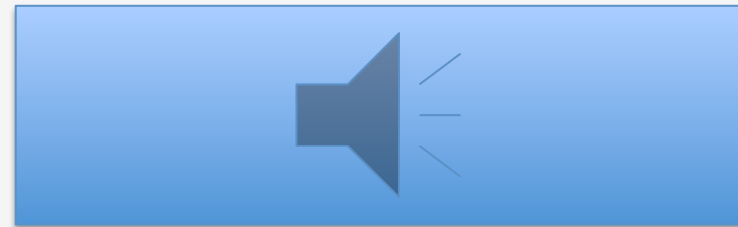
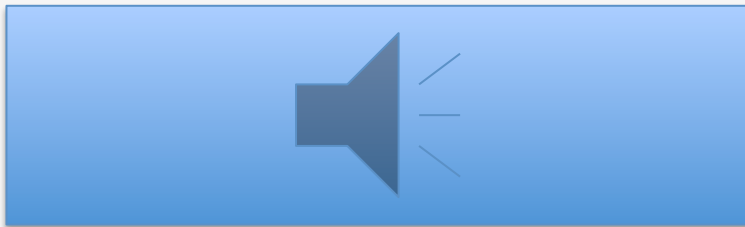


Einfluss von Task:

Lesetask vs. Wiederholungstask, (nicht-) Aussprache von /h/:

Lesetask: Frankfurt liegt in Hessen.

Wiederholungstask: Du hast eine schöne Hütte.





Der „große“ Korpus:

- Beginn der Aufnahmen in diesem Monat
- Leichte Anpassungen der Sätze
- Leichte Anpassungen des Umfangs
- Leichte Anpassungen der Prozesse, die wir untersuchen wollen
- Erneuerte Aufnahmesoftware (teilweise auch Hardware)
- Perzeption (sowohl L2 betreffend, als auch Verständlichkeit von L1-Hörern)

-> Entwickeln der Lernersoftware mit individuellem Feedback



[mɛ^e'si bo'k^hu:]

[fi:'lœn 'dãk]

Merci viumou!



> 3. Erste Daten

