UNIVERSITÄT DES SAARLANDES

Foundations of Language Science and Technology

Speech Corpora

Jan 30th, 2015

Frank Zimmerer

Definition

Speech corpora = collection of linguistic or phonetic data which has been constructed, edited and analyzed by specific scientific criteria

Types of corpora

- *Collection of a special field*, e.g. Thesaurus Linguae Graecae, TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien)
- *Representative data set*, e.g. LIMAS (LIMAS = Linguistik und Maschinelle Sprachbearbeitung)
- *Task-specific*, e.g., Kiel corpus (of spontaneous speech)

Corpora of spoken language - What for?

- Speech technology
- Phonetic fundamental research
- Synchronic picture of language

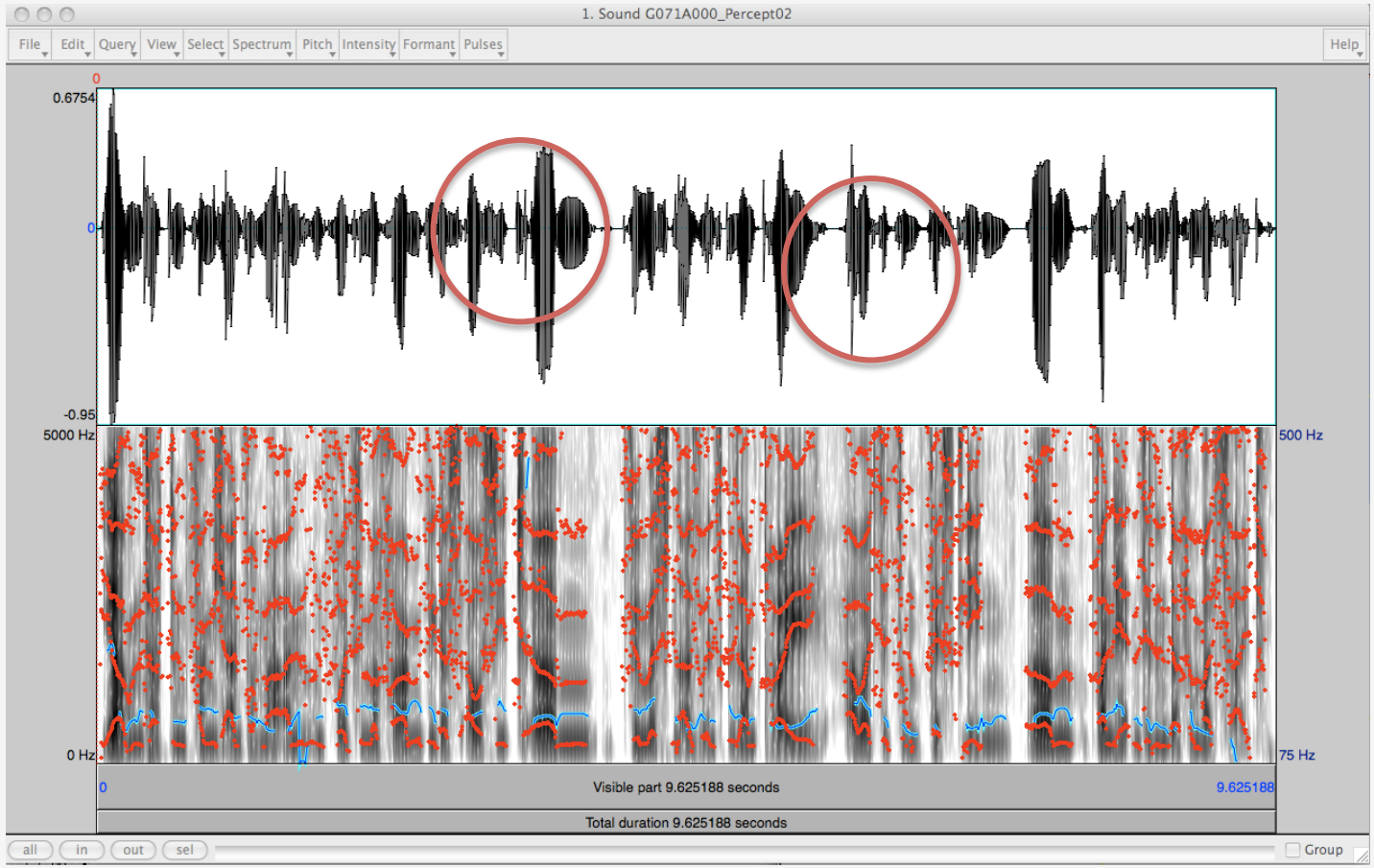
Relevant differences between spoken & written language

- Orthographic uniqueness vs. phonetic variability
- String vs. data stream
- Structure
- Corpora of spoken language are multimedia by nature

Example Text: Usually no problem, however, only if it is written like the sentences you are just reading now. However, if things happen to the text like in the next paragraph, it‘s not that easy anymore (the next text, however, is not a standard example (taken from a file where automatic text recognition was performed on a scanned text)).

labelingfu nctionsa nd categorical-likdeis criminatiofnu nctionsfo r syntheticallpyr oduceds peechs timuli differingi n voice-onsetti me (VOT). Other researchh as found somewhatc omparabler esultsf or young infantsa nd chinchillasa s well as cross-languagdeif ferencesin the perceptiono f theses ame synthetic

Spoken language example:



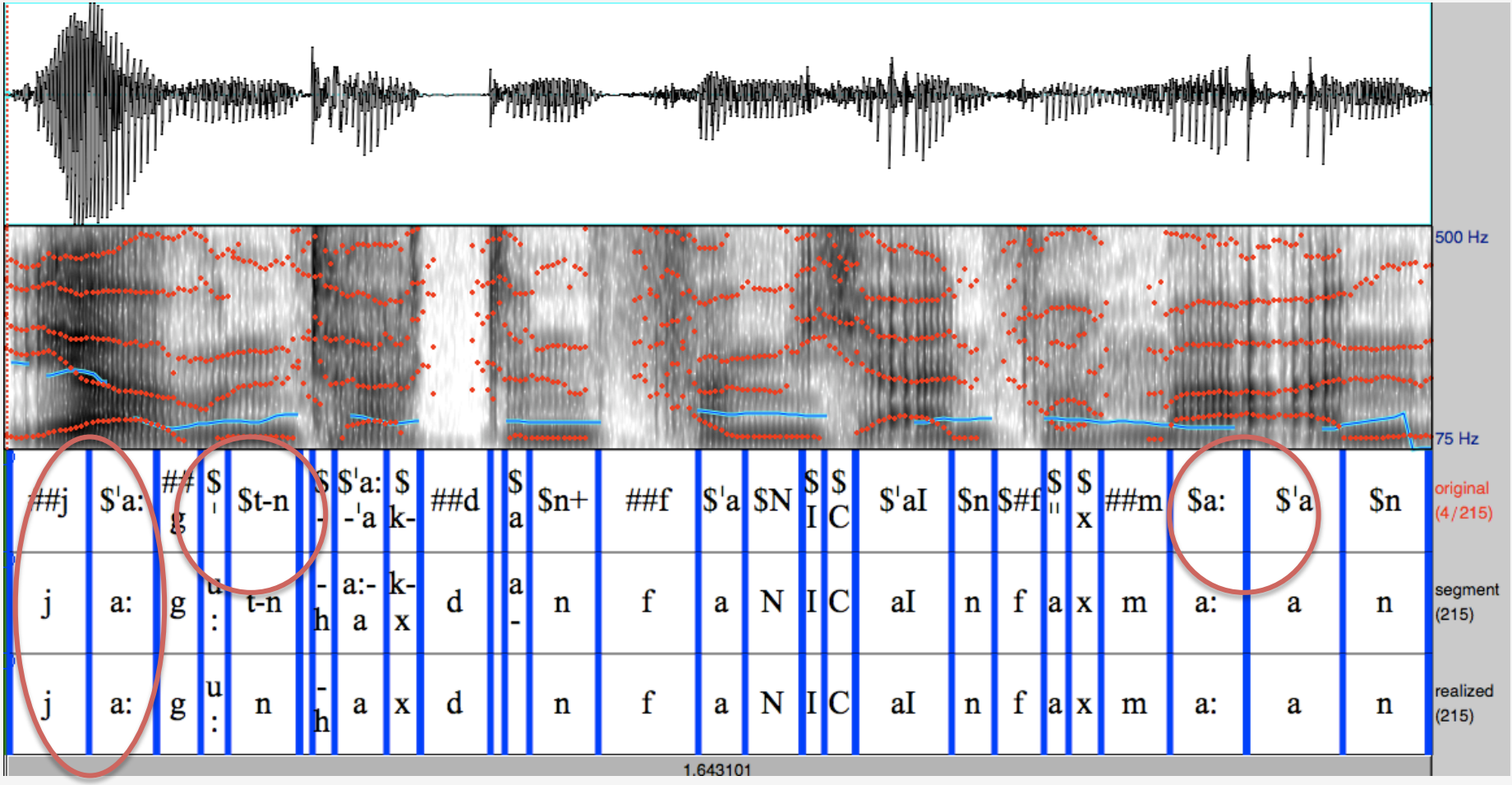
Annotation

- Has to be done in written form and saved individually
- Corpus needs to be annotated on different levels
- Requirements do not have to be clear a priori
- Can be extended optionally

Segmentation

- Word segmentation
- Sound segmentation
- Further segmentation (syllable, phrase)?
- Tools for automatic segmentation available

Spoken language example (Kiel Corpus, IPDS, 1994):



Automatic sound segmentation (forced alignment)

- Forced alignment: modification of a speech recognition module for segmentation
- Given transcription of utterance
- Alignment of transcribed data with speech data, find the single path with the highest probability
- Need for an additional training phase, especially for speaker adaption

Examples later: IFCASL, WebMAUS

Speech recognition and voice input

- Recognition of words or word sequences; necessary basis for speech understanding and dialog systems
- Training material
  - Algorithms on the basis of stochastic models need a huge amount of data to train the model's parameter
  - Sensitive to surrounding parameters
  - Facilitation regarding segmentation of already trained models

Concatenative synthesis

- Uses segments of natural speech, concatenated and resequenced to synthesize the intended utterance (typically diphones)
- Recording criteria
  - One speaker
  - Diphones in context
  - often realized in secondary stress position

Unit selection synthesis

- Dynamic selection of units at synthesis run-time
- Recording criteria
  - Requires an extensive corpus of one speaker
  - Coverage: numerous entities of all elements on the level of the smallest unit
  - Annotation/Segmentation: all required units → LNRE problem

LNRE problem (Large Number Of Rare Events)

- Spoken units are distributed unevenly on all levels
- 50 most frequent words constitute about 60% of a continuous text (in German)
- Incidence of most of the other words are vanishingly low

Corpus outline for unit selection synthesis

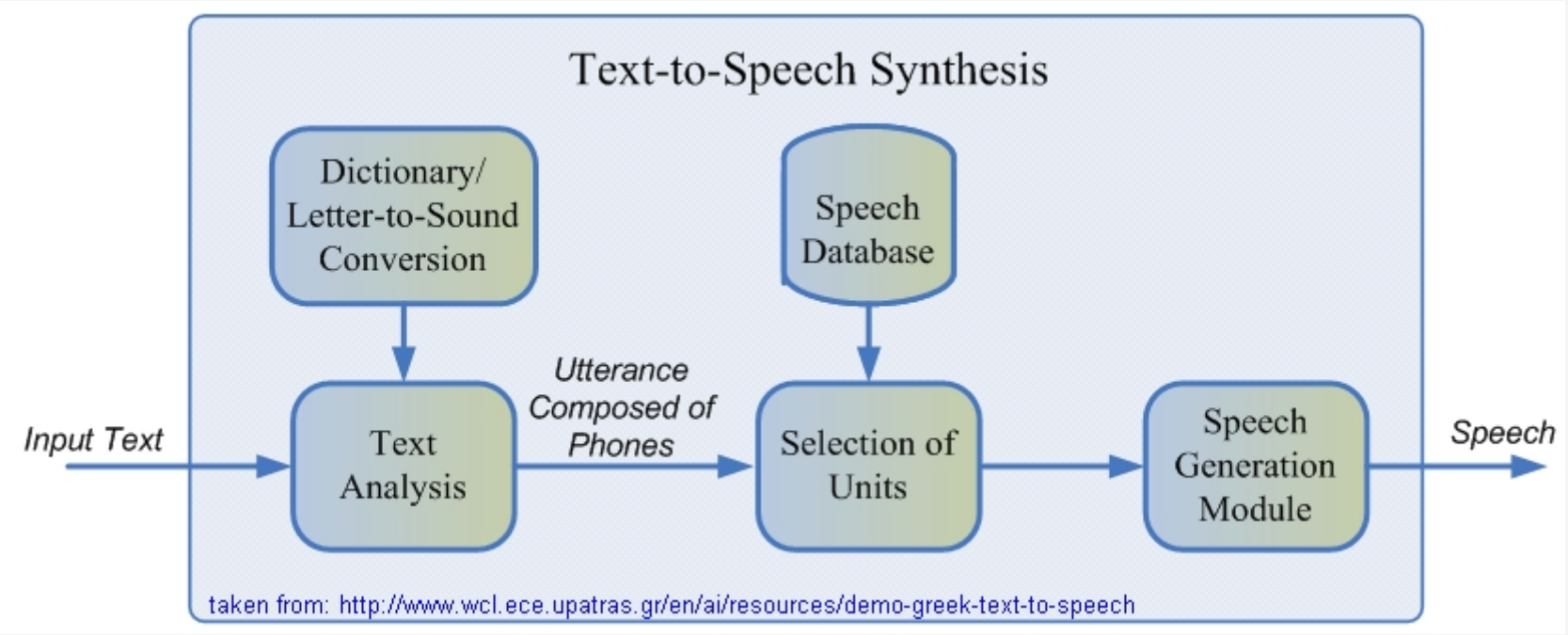
Text outline

- Different for domain-specific vs. domain independent material
- Spontaneous speech material?
- Designed sentences
  - Domain specific words/phrases
  - Text corpus
  - Frequent words

Processing of recordings

- Word and sound annotation and segmentation
- Boundaries, phonetic context, word position (initial, medial, final) sentence mode, reduction (reduced – unreduced), prosodic labeling
- Each criterion that is included in the cost function needs to be labeled

> Corpora for different purposes



(Phonetic) Learner Corpora

- L1/L2 aspects (e.g., interferences)
- Recording criteria
  - Many speakers
  - Segments in various context
  - Minimal pairs, variation in complexity, different tasks

> Steps for corpus construction

General requirements for qualitatively good (learner) corpora:

1. Qualitatively good audio-recordings (not easy to change)
2. Qualitatively good annotation (on whatever level – most important is an annotation that suits your needs)
3. Qualitatively reliable metadata
4. Guaranteed accessibility (?)

> Steps for corpus construction

Important steps (not necessarily linear, some may think trivial?)

- Research question (theoretic, general, specific, broad, …)
- How can I answer the research question(s)
- Decision on materials (what to record)
- Decision on methods (how to record whom, and where – legal questions)
- Actual hard- and software acquisition (and testing)
- (Maybe) Trial recordings
- (Maybe) Change in Setup
- Data acquisition (recording, also metadata)
- Annotation
- Analysis
- Publication (?)

> Steps for corpus construction

Example: Learner corpus (IFCASL)

Generally: What is the research question, how do I get an answer?

What is the *language pair*? (examples: F<->G, J->G, one possible reason availability of speakers, asymmetric data may be problematic)

What are the phonological and phonetic properties of BOTH languages? What are the interferences I expect?

What kind of speech style am I interested in?

Are there existing corpora? If yes, could save time (corpus work is time consuming, expensive, …)

If there are: does it suit my goals?

If there are not: Creation of a corpus. (are you REALLY sure, you want do this?)

> Steps for corpus construction

For the selection of materials and methods, many decisions have to be made:

- All Phonemes vs. representative samples?
- Control over what has been said? (More control ~ less natural)?
- Presentation scenario (reading, repeating, interview, discussion, …)?
- Size of units that are recorded (e.g., vowels, syllables, words, sentences, stories, conversations …)?
- Where are recordings made (e.g., Sound treated room, Quiet room, Class room)?
- How are they made (e.g., single speaker, two speakers, many speakers)?

> Steps for corpus construction

Annotation/Segmentation

(TIME CONSUMING! 1 – 500 times actual speech, see later)

- What EXACTLY has been said when EXACTLY in the corpus?
- Automatic or manual (or a mixture thereof)?
- What are the units? (e.g., segments, words, sentences, …)
- Quality control (e.g., number of labelers, interrater-reliability, …)

Analysis (coming back to research questions)

Publication

Who can have access?

What (parts) can be published?

How do others know about the corpus?

-> CLARIN CENTERS across Europe can help

> Example

-> IFCASL Project (DFG and ANR):

(*Individualised **f**eedback in **c**omputer-**a**ssisted **s**poken **l**anguage learning*)

- French native speakers (L1) learning German (L2)
- German native speakers (L1) learning French (L2)

General rationale: What are the phonetic and phonological problems for learners occurring in this language pair? Were do interferences occur?

- Aims of the corpus
  - Training and test material for automated feedback system
  - Data and analysis for phonological research

> Example

Depending on language pair(s), not much data is available (e.g., French<->German):

- Often personal or anecdotal data (e.g., /h/)
- (Theoretical) contrastive comparison
- Few Corpora:
  - Mainly: written corpora
  - Most of the time: English as one member of a pair
  - Rarely corpora for Speech*pairs*
  - Examples:
    - HABLA (Hamburg Adult Bilingual LAnguage) Corpus with bilinguals (L1: French & German) (Kupisch et al. 2012)
    - IPFC-allemand (Interphonologie du Français Contemporain) with L1: German L2: French (very advanced s) (Pustka 2012)

> Example

Goal: 100 Speakers:

50 Speakers (25 female, 25 male) L1 French and L2 German recorded in Nancy

50 Speakers (25 female, 25 male) L1 German and L2 French recorded in Saarbrücken

Different Proficiency Levels within each language group:

10 High school students, 14-16 Years of age, Beginners (A1/A2)

20 Adults, Beginners (A1/A2/B1)

20 Adults, Advanced Learners (B2/C1/C2)

> Example

Questionnaire for participants to categorize proficiency levels and for Metadata (Speakers are there only once!)

- Linguistic biography (in L1)
  - L1 and age, highest educational degree, places of residence
  - For each L2:
    - Years of instruction (e.g., school, university, …)
    - Stays abroad (e.g., exchanges, au-pair, …)
    - Everyday use (e.g., partner, parents, tandem partners etc.)
    - Official certificates
- Self-assessment
  - Self-assessment of language skills, esp. pronunciation
  - Motivation for learning L2
  - Attitude towards language learning also with a computer

> Example

- Small remuneration for subjects
- Consent to be signed
  - Subject stays owner of the data
  - If wished data can be accessed by the owner at any time
  - Data can be used
    - in anonymous form for scientific purposes (oral and written)
    - for speech signal processing
    - for improvement of language learning software
- Access of data (audio + annotations + meta-data)
  - Data of subjects **not** for *public* use (except explicitly indicated)
  - on request for *research* purposes

> Example

Control was rated important -> Corpus of read speech (no spontaneous speech)

All speakers are recorded in in L2 *and* L1 (“double parallel”). -> Is production specific to individual, or to language (or task)?

Recordings are made in (quiet) offices -> good quality:

- Head-mounted close-talk microphone (almost invisible for speakers)
- Recording software: JCorpusRecorder (V. Colotte, 2013, Nancy)
- All Sentences are displayed on a computer screen and to be read aloud
- Some parts are a pure reading-task, another part is a repetition task
- One sentence, one audiofile
- One recording session should not take longer than 60-75 minutes (~10 minutes for questionnaire, ~50 minutes for recordings)
- First some trials – „Minicorpus“

> Example

Construction of sentences and texts to find phenomena where we expect some interferences among others:

Segmental Level

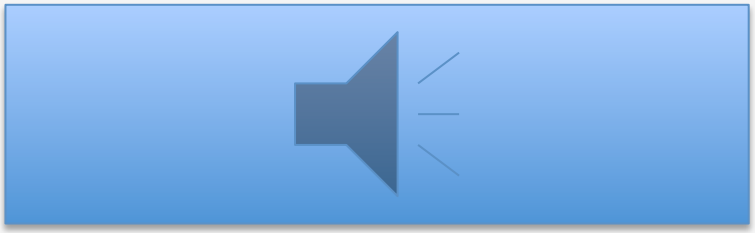
- Glottal stop [ʔ] and glottal fricative [h]
- Liaison and enchaînement consonantique (e.g., *pot-au-feu*)
- Nasal vowels [ɛ̃, ã, ɔ̃]
- (Final) devoicing of plosives and voicing patterns of fricatives
- Aspiration of unvoiced plosives [p^h, t^h, k^h]
- ich- and ach-sound [ç, x]
- Schwa: Rounded? Full vowel?
- /r/ produced as consonant [ʁ, ʀ] vs. vowel [ɐ]
- Vowel length (and quality) [iː-ɪ, eː-ɛ, ɛː-ɛ, aː-a, oː-ɔ, uː-ʊ, yː-ʏ, øː-œ]
- Consonant clusters (e.g., *Bratwürstchen*)
- reductions (e.g., lenition, assimilations)

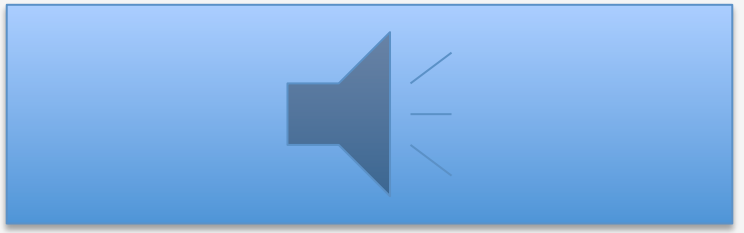
> Example

French speakers' productions of /h/ Influence of task and individual differences:

Reading task: Frankfurt liegt in Hessen.

Repeating task: Du hast eine schöne Hütte.





> Example

Prosodic Level

- Word stress
- Contrastive focus
- Pitch Range

Orthography-Pronunciation Problems

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

Internationalisms, proper names

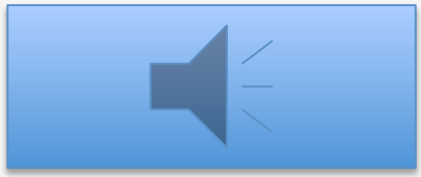
- Is French. "énergie“ read as ([enɛɐ'giː]) by German learners?
- Is German "Berlin” read as ([bɛr'lɛ̃]) by French learners?

5. Mistakes in reading (influences from other L2s)
   - German "Licht" [lɪçt] (Engl. "light") read as [laɪtʃ]
6. Minimal pairs, e.g., German. „Paar-Bar“, or French „port-bord“
7. Every phoneme of every language should appear at least once

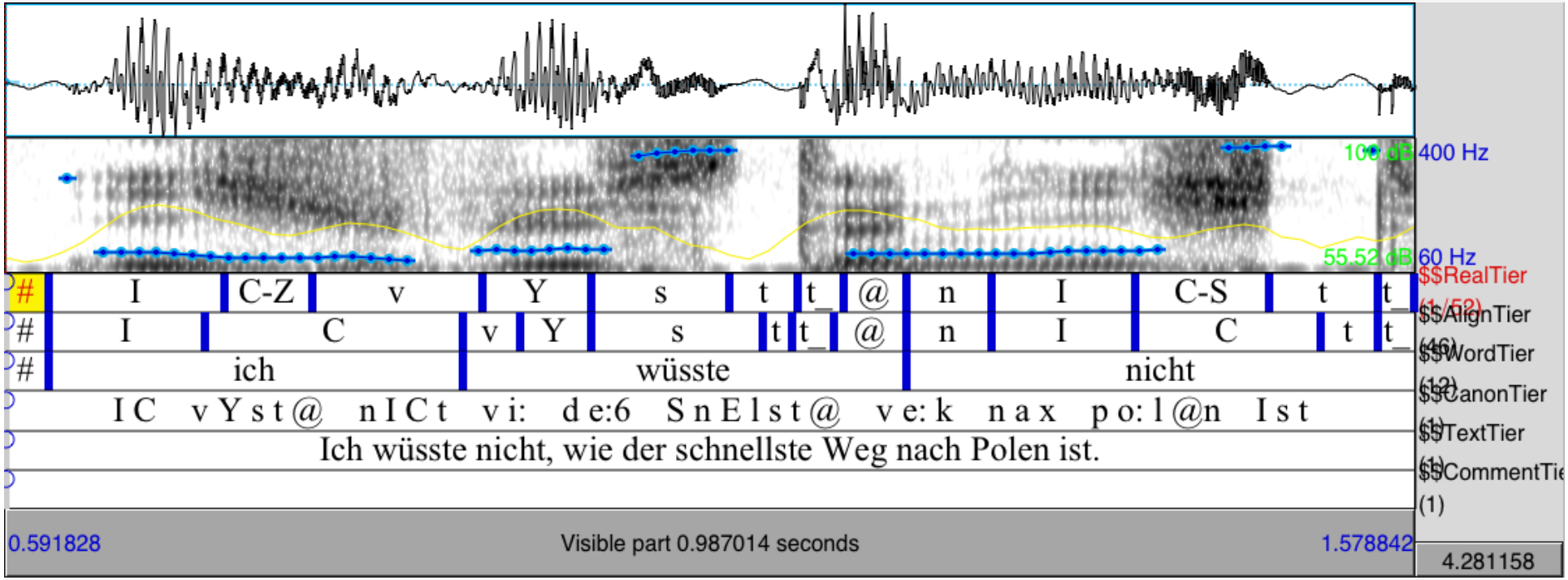
> Example

4 Conditions for each of the languages, L2 recorded first

1. ***Reading task***
   Sentences appearing on the screen have to be read aloud.
2. ***Repetition task*** *(Only reading in L1)*
   Sentences are presented, read by a native speaker, and have to be repeated. (Control for mistakes that are purely based on orthography)
3. ***Focus-Task***
   Questions are played, speakers have to read the answer to these questions (Word in focus is also written in CAPITALS)
4. ***Reading task***
   Stories to be read: “The three little pigs”
   Text about jobs in the environmental sector



Levels of annotation



> Example

	Symbol	IPA	Sample word	Transcription
Consonants				
Plosives	p | p_	p	Pein	p p_ aI n
	b | b_	b	Bein	b b_ aI n
	t | t_	t	Teich	t t_ aI C
	d | d_	d	Deich	d d_ aI C
	k | k_	k	Kunst	k k_ U n s t t_
	g | g_	g	Gunst	g g_ U n s t t_
	Q | Q_	ʔ	Verein	f E6 Q Q_ aI n
Fricatives	f	f	fast	f a s t t_
	v	v	was	v a s
	s	s	Tasse	t t_ a s @
	z	z	Hase	h a: z @
	S	ʃ	waschen	v a S =n
	Z	ʒ	Genie	Z e n i:
	C	ç	sicher	z I C 6
	x	x	Buch	b b_ u: x
	h	h	Hand	h a n t t_

> Example

	Symbol	IPA	Sample word	Transcription
Affricates	pf | pf_	pf	Pfahl	pf pf_ a: l
	ts | ts_	ts	Zahl	ts ts_ a: l
	tS | tS_	tʃ	deutsch	d d_ OY tS tS_
	dZ | dZ_	dʒ	Dschungel	dZ dZ_ U N =l
Glides/Liquids	j	j	Jahr	j a:6
	l	l	Leim	l aI m
	r	ʁ	Reim	r aI m
Nasal consonants	m	m	mein	m ai n
	n	n	nein	n aI n
	N	ŋ	Ding	d d_ I N
Syllabic consonants	=n	n̩	waschen	v a S =n
	=m	m̩	großem	g g_ r o: s =m
	=l	l̩	segelt	z e: g g_ =l t t_

Vowels

Tense vowels (long)	i:	iː	Lied	l i: t t_
	e:	eː	Beet	b b_ e: t t_
	E:	ɛː	spät	S p p_ E: t t_
	a:	aː	Tat	t t_ a: t t_
	o:	oː	rot	r o: t t_
	u:	uː	Blut	b b_ l u: t t_
	y:	yː	süß	z y: s
	2:	øː	blöd	b b_ l 2: t t_
Tense vowels (short)	e	e	Revanche	r e v a~ S
	i	i	Kontinent	k k_ O n t t_ i n E n t t_
	o	o	Oase	o Q Q_ a: z @
Lax vowels	I	ɪ	Sitz	z I ts ts_
	E	ɛ	Bett	b b_ e t t_
	a	a	Satz	z a ts ts_
	O	ɔ	Gott	g g_ O t t_
	U	ʊ	Schutz	S U ts ts_
	Y	ʏ	hübsch	h Y p p_ S
	9	œ	Götter	g g_ 9 t t_ 6

> Example

Table 3: Symbols for annotating segment- and word-level deviations

Phenomenon	Symbol	Example(s)	Usage notes
substitution	*old-new*	C-S (1SR04_FGBA2_501)	
insertion	*-new*	-d (1SR04_FGWA2_507)	
deletion	*old-*	r- (1SR07_FGMC1_505)	RealTier interval with deleted segment should not be removed, but given duration < 1 ms
devoiced	_0	Z_0 (1SH24_GFBA2_001) g_0, g__0 (1SR04_GFWA2_003)	RealTier only
voiced	_V	s_V (1SH05_FGMA2_502)	RealTier only
filled pause/hesitation	&	&ist (1SH05_FGWA2_507) &l' (1SH01_GFBA2_001)	WordTier only
truncated words	/	schnel/ (1SH05_FGMA2_506)	WordTier only
breathing	§	1SR07_FGMC1_505	RealTier only
pause	#	1SR04_FGBA2_501 1SR04_GFWA2_003	
word-medial pause	-#	1SR02_FGMA2_502	RealTier only (no change to the word on WordTier)
non-vocalic noise	!!	1SR02_FGMA2_502	Taps on microphone, mouse clicks, etc.; RealTier only
vocalic noise	!	1SH05_FGMA2_504	Lip smacks, coughs, etc.; RealTier only
uncertain label	_?	-Q_?, -Q__? (1SR07_FGMC1_505)	Details (e.g. possible alternative labels) may be included on CommentTier
uncertain boundary	%	%aU (1SR07_FGMC1_505)	The symbol is used at the left of the interval following the uncertain boundary

> Example

Now, it is your turn to work with automatically generated annotations (download from the course website)

> Resources / URLs

https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/links-en/korpora_links

http://titus.uni-frankfurt.de/indexd.htm?/texte/texte2.htm

http://www.ifcasl.org/

http://www.korpora.org/Limas/

http://clarin.phonetik.uni-muenchen.de/BASWebServices/#/services