What we can learn from ASR errors about low-resourced languages: A case-study of Luxembourgish and Austrian

M. Adda-Decker, B. Schuppler, L. Lamel, J.-A. Morales-Cordovilla, G. Adda,

LPP/LIMSI/Graz University





Outline

1. Presentation of languages: Luxembourgish and Austrian
   speaking styles: prepared spontaneous
2. Oral language specificities w.r.t. ASR (and errors)
3. ASR system & results
4. Error analysis

Luxembourg

- small country in Western Europe (∼2500 km2) bordered by Belgium, France and Germany
- ∼500 000 inhabitants (including 350 000 natives)
- native language: Lëtzebuergesch
- Western Germanic (Moselle Franconian)
- other practised languages: German, French...

Luxembourgish as a less-ressourced language

shares many features with the world's thousands of oral languages:

- mainly oral language
- poor written standardisation
- even though official standards exist (still evolving)
- code-mixing and code-switching
- high dialectal variation
- almost not taught at school (German at the age of 6, French at 7)
- frequent use of German and French for writing

Luxembourg

Translation of the below sample:

Six months of presidency is tiresome and it is time that it goes to an end.

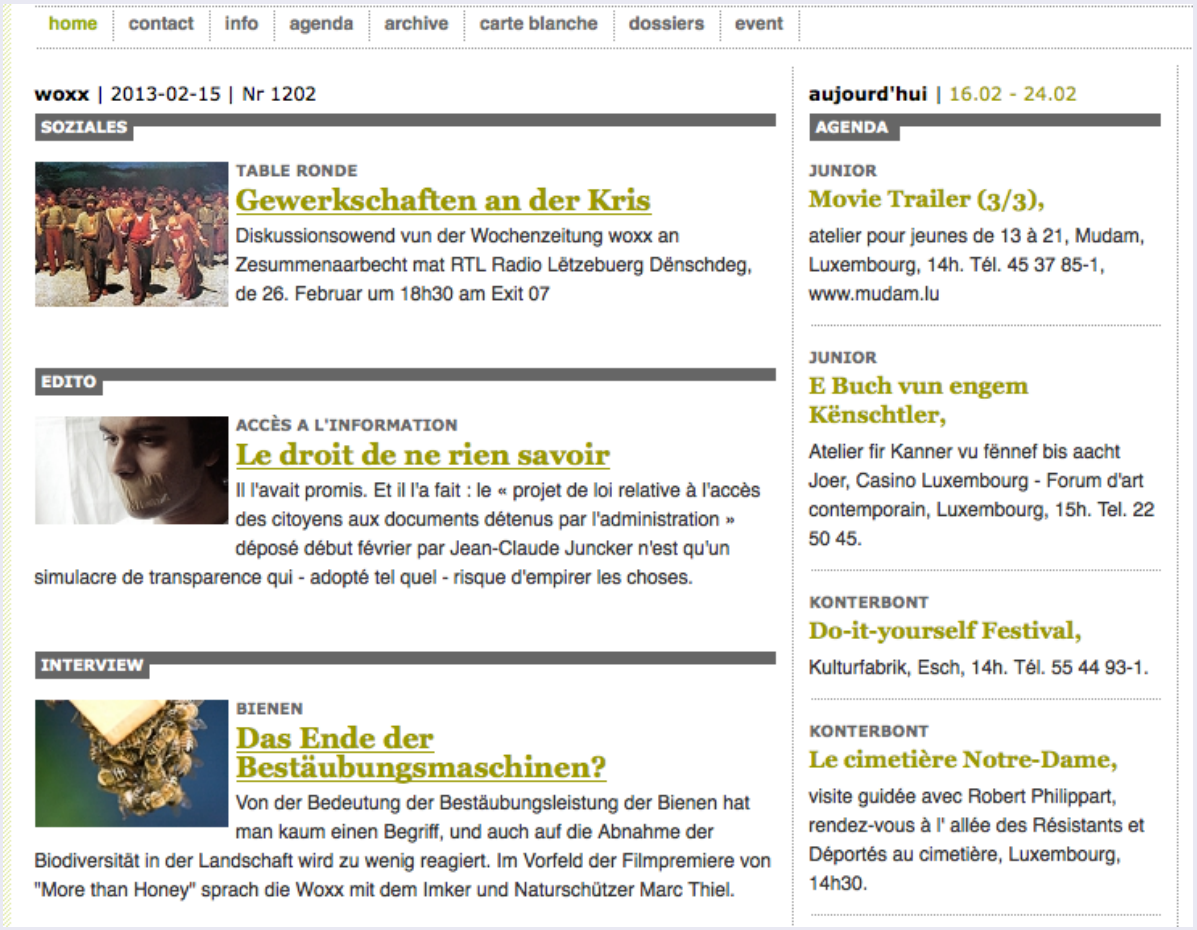
Sample of "Lëtzebuergesch" RTL TV 2005-07-08

Manually transcribed:	**German translation:**	**French translation:**
"sechs Méint Présidence, dat mëcht midd an elo gëtt et Zäit dass et eriwwer ass ", sot de Staatsminister.	"sechs Monate Präsidentschaft, das macht müde und jetzt ist es an der Zeit abzuschliessen ", sagte der Staatsminister.	"six mois de Présidence, cela fatigue et il est temps maintenant d'en finir", conclut le ministre d'État.

- Germanic language: many similar word forms between L and G
- import of French words (e.g. technical, administrative)
- frequent Lux. small word forms (et, de, an, um... (resp. engl. it, the, and, on)) collapse with:
  French (et (engl. and), de (engl. of)) and
  German (an (engl. at), um (engl. for))
  English (an (engl. an))

Luxembourgish written sources

Example of a web weekly newspaper (http://www.woxx.lu/)



home | contact | info | agenda | archive | carte blanche | dossiers | event

woxx | 2013-02-15 | Nr 1202

SOZIALES

TABLE RONDE

Gewerkschaften an der Kris

Diskussionsowend vun der Wochenzeitung woxx an Zesummenaarbecht mat RTL Radio Lëtzebuerg Dënschdeg, de 26. Februar um 18h30 am Exit 07

EDITO

ACCÈS A L'INFORMATION

Le droit de ne rien savoir

Il l'avait promis. Et il l'a fait : le « projet de loi relative à l'accès des citoyens aux documents détenus par l'administration » déposé début février par Jean-Claude Juncker n'est qu'un simulacre de transparence qui - adopté tel quel - risque d'empirer les choses.

INTERVIEW

BIENEN

Das Ende der Bestäubungsmaschinen?

Von der Bedeutung der Bestäubungsleistung der Bienen hat man kaum einen Begriff, und auch auf die Abnahme der Biodiversität in der Landschaft wird zu wenig reagiert. Im Vorfeld der Filmpremiere von "More than Honey" sprach die Woxx mit dem Imker und Naturschützer Marc Thiel.

aujourd'hui | 16.02 - 24.02

AGENDA

JUNIOR

Movie Trailer (3/3),

atelier pour jeunes de 13 à 21, Mudam, Luxembourg, 14h. Tél. 45 37 85-1, www.mudam.lu

JUNIOR

E Buch vun engem Kënschtler,

Atelier fir Kanner vu fënnef bis aacht Joer, Casino Luxembourg - Forum d'art contemporain, Luxembourg, 15h. Tel. 22 50 45.

KONTERBONT

Do-it-yourself Festival,

Kulturfabrik, Esch, 14h. Tél. 55 44 93-1.

KONTERBONT

Le cimetière Notre-Dame,

visite guidée avec Robert Philippart, rendez-vous à l' allée des Résistants et Déportés au cimetière, Luxembourg, 14h30.

Articles in different languages.

Sentences may include words (expressions, titles...) in different languages.

Words (compounds) may be formed using different languages

Ger-French: **Wahl-cadeauen**

L-Fr: **Laangzäit-chômeur**

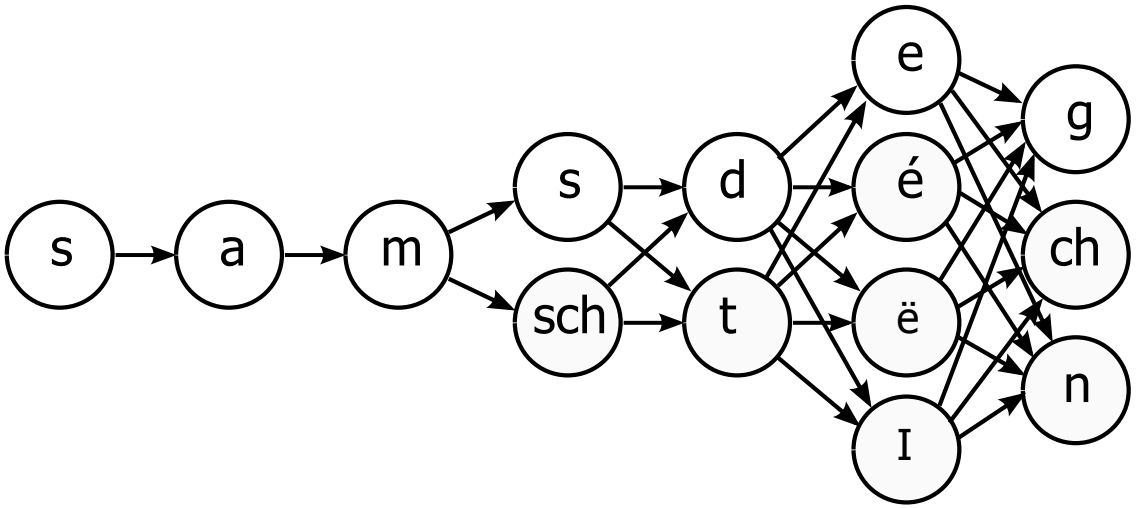
Outline

1. Presentation of languages: Luxembourgish and Austrian
   speaking styles: prepared spontaneous
2. **Luxembourgish specificities w.r.t. ASR (and errors)**
3. ASR system & results
4. Error analysis

Luxembourgish specificities w.r.t. ASR (and errors)

- large proportion of import words (French...)
- partial assimilation within the Luxembourgish system
  - Fr.: bassin $\rightarrow$ Lux.: Baséng
  - (no nasal vowels, lenition/voicing)
  - Fr./Lux.: gérance
- poorly stabilized writing conventions
- rule: write as pronounced
- (partial) fallback to French, German, English writing conventions
- multiple writings depending on the choice of writing conventions
  - Grupp (Ger), Groupe (French), Group (French adapted to L/G)
- problems with pronunciation dictionary development
- automatic processing almost requires a multilingual system

Example of writing variants: Saturday



- "standard" form: Samsdeg (Saturday)
- /s/ → /ʃ/ (regional variant)
- /d/ → /t/ (assimilation)
- unstable vowel timber of unstressed syllable
- unstable word ending

Samsdeg

Samschdeg

Samsdes

Samschden

Samsden

Samsten

Samschde

Samsdich

Samsdëg

Samsteg

Samschde

Samschten

Samsdéch

Samschteg

Outline

1. Presentation of languages: Luxembourgish and Austrian

   speaking styles: prepared spontaneous
2. Oral language specificities w.r.t. ASR (and errors)
3. **ASR system & results**
4. Error analysis

ASR system - generalities

- Phoneme inventory : 56 symbols (+ 3 silence, hesitation, breath)
  - 31 voyels (short/long; 9 diphthongs, nasals...)
  - 25 cons
- Pronunciation dictionary : 200k words
  - high rate of writing variants
    - rule: write as pronounced
    - diacritics often omitted (confusion between diphthongues)
- High ambiguity for grapheme-phoneme conversion

Spoken and written corpora

- Speech ($>$ 1000h)
  - Parliament debates (Chamber)
  - Radios (RTL, 100,7)
- Written material: multilingual, noisy
  - Chamber reports
  - RTL web site
  - Wikipedia
  - reports, books...
  - Social media

written source	tokens (k)
CHAMBER	22 110
MISC	1677
RTL2008	611
RTL2012	10 307
WIKIPEDIA	3603
BLOGS	10 243
BLOGS_COMMENTS	3106
total	51 657

Total : $>$ 50 millions of raw data

Acoustic Models

- 3-state HMM with Gaussian mixture
- Context-dependent phone models
- Gender-independent, wideband band
  - Position-dependent triphones:intra- and inter-word
  - triphone with backoff models
  - selection based on frequency in training data
- State-tying to reduce model size and increase triphone coverage
- MLLT, Pitch
- MLP features (from German)

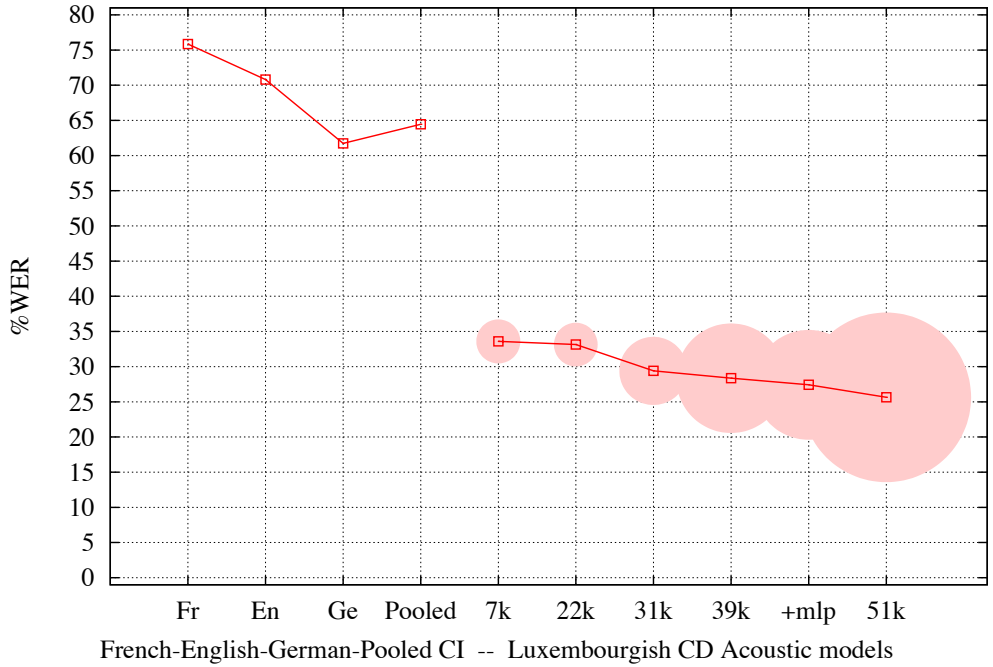
Acoustic Features

- PLP features
  - Classical 39 cepstral parameters
  - 13 Cep + 13ΔCep + 13 ΔΔCep
  - 30ms window with 10ms offset
- MLP features
  - Wide temporal context (250 ms)
  - TRAP-DCT (TD) features (give similar performance to time-warped linear predictive TRAP (wLP-TRAP)
  - Concatenation with PLP features and pitch
  - 81 component feature vector

ASR experiments

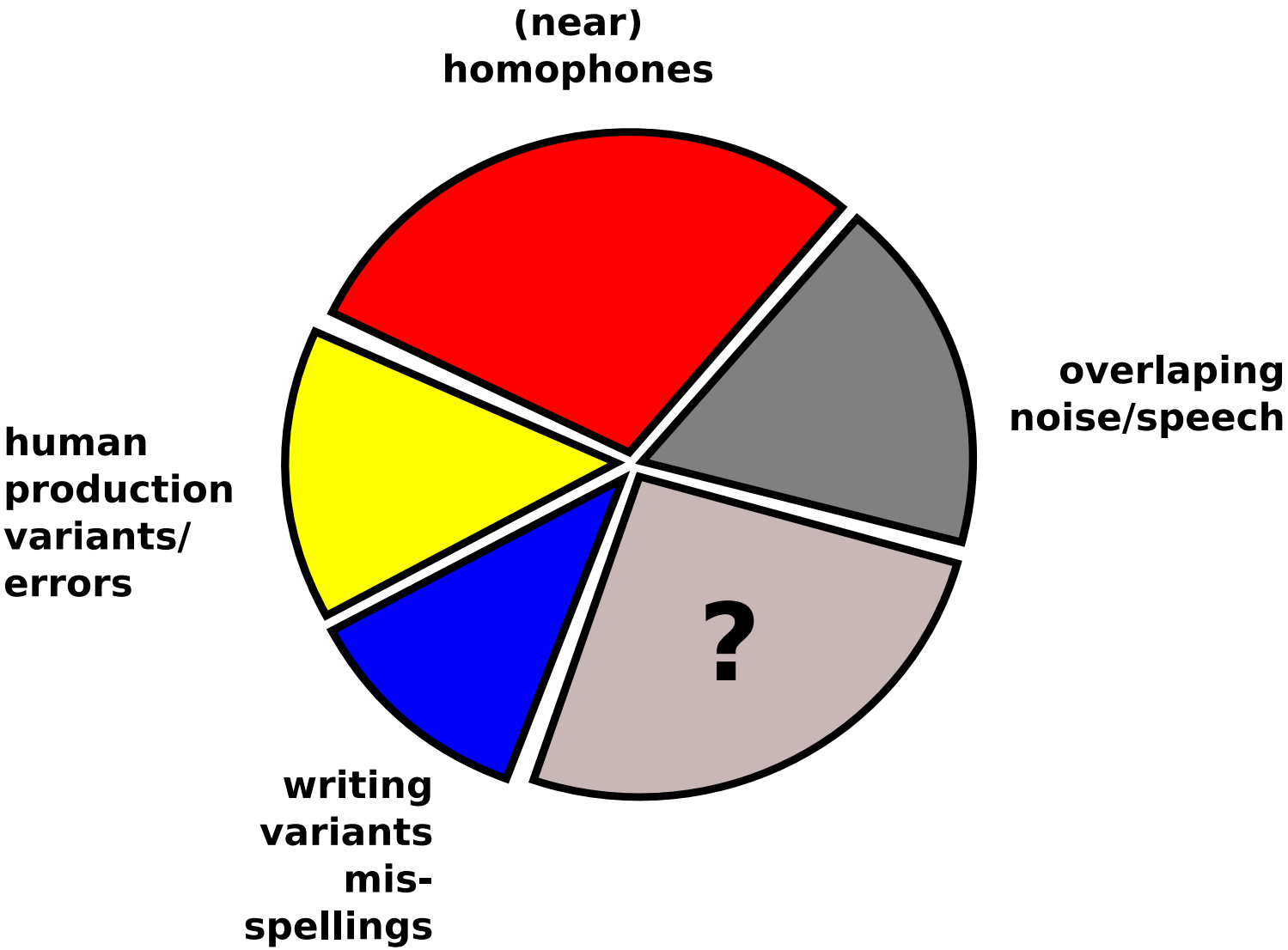
- 10 hours of various radio shows (journalistic)
- 1st set of experiments: acoustic models from major Western languages
- unsupervised transcription to train Luxbg. models
- 2nd set of experiments: Luxbg. models with progressively increasing number of contexts (training data)

ASR results

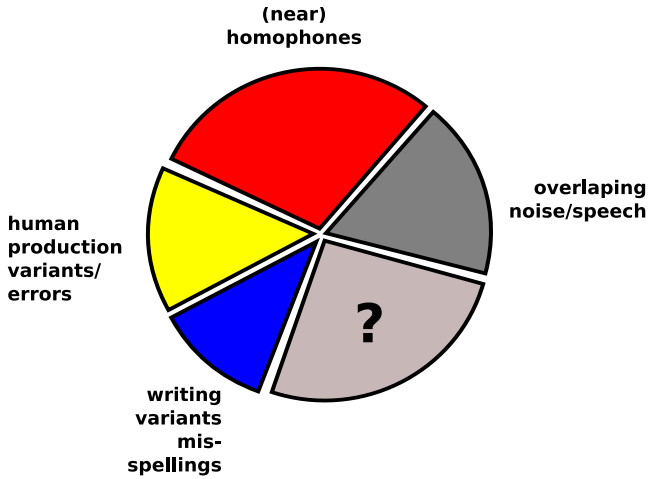


- left: acoustic models from related major languages (CI models)
- right: Luxembourgish models (unsupervised training, CD models
- increasing number of contexts (7k $\rightarrow$ 51k)
- increasing volume of training data (approx. 100, 500, 1000 hours)

Typical ASR error types



Luxembourgish ASR errors



- strong increase of the rate of writing variants and misspellings:

REF $\rightarrow$ HYP

vir $\rightarrow$ fir (Eng.: for)

schonn $\rightarrow$ schon (Eng.: already)

Trakter $\rightarrow$ Tracteur

- word-initial and word-final schwa deletions:

esou $\rightarrow$ sou (Eng.: so)

gëschter $\rightarrow$ gëscht (Eng.: yesterday)

...

Luxembourgish ASR errors

Error study provides:

- insights in Luxbg. phonological processes (schwa, lenition, reduction, mobile-n, voicing assimilation)
- ongoing changes in phonemic inventory (merging of alveolar/palatal fricatives /S/, /ç/)?)

What we can learn from ASR errors about Austrian German

Barbara Schuppler and
Juan Andrés Morales-Cordovilla
SPSC Laboratory,
Graz University of Technology

MAIN GOAL

Use ASR as a tool to get an insight into a not well documented (variety of a) language

- Incorporate findings into pronunciation modeling and create an Austrian German dictionary

ASR system

Front-End:

- Feature size: 39 (MFCC+D+DD with CMN).
- Frame shift and length: 10 and 32 ms.

Back-End: recognizer based on **HTK**

- 36 monophones to generate triphone HMMs
  (6 states, 8 Gauss/state).
- Lexicon size: 3794 words, one pronunciation per word
- Language model: bigram trained with the transcriptions of the test set!

J. A. Morales-Cordovilla, H. Pessentheiner, M. Hagmüller, P. Mowlaee, F. Pernkopf and G. Kubin. "A German distant speech recognizer based on 3D beamforming and harmonic missing data mask". *AIA-DAGA*. 2013.

(written) Standard Austrian German

-- Lexical differences > Austrian German Dictionary
-- Grammatical differences: tenses and cases

(read) Standard Austrian German

-- Plosives: **B**ar**b**ara vs. **P**etra;
(same for dental and velar plosives)

-- No /s/ - /z/ distinction

-- Mögli**ch** vs. lusti**g**

-- . . .

> Acoustic Models!

(read) Standard Austrian German

Read Austrian poses no problem for ASR system trained on German data?

EXPERIMENT 1

Read Austrian poses no problem for ASR system trained on German data?

>> ASR system trained on 5000 read sentences from German speakers of mixed regional (BAS corpus):

>> 95.7% WAcc on sentences read by Austrians.

EXPERIMENT 1

Read Austrian poses no problem for ASR system trained on German data?

>> ASR system trained on 5000 read sentences from German speakers of mixed regional (BAS corpus):

>> 95.7% WAcc on sentences read by Austrians.

How about spontaneous Austrian German?

GRASS: Graz corpus of Read and Spontaneous Speech

-- 38 Speakers, m & f
-- Speakers of standard Austrian

– **Conversational speech**:
1h free conversation between good friends or family members > tot. 20h

– Read and spontaneous commands
– Read text (phonetically balanced)

EXPERIMENT 2

Is an ASR system trained on German speech suitable for recognizing spontaneous Austrian German ?

>> 28.4% Acc. on 5140 “CLEAN“ utterances from 10 speakers of conversational Austrian German

EXPERIMENT 3

Train set

5873 utterances of the 10 test speakers, of which

> 656 utt. of read speech and commands

> 5217 ‘CLEAN’ utterances from conversational speech

EXPERIMENT 3

Train set

5873 utterances of the 10 test speakers, of which

> 656 utt. of read speech and commands

> 5217 ‘CLEAN’ utterances from conversational speech

Test set

7694 utterances different from train set, 10 male and female GRASS speakers

> 2554 read speech and spoken commands

> 5140 from conversations

RESULTS

Read Speech (#= 1785):
EXP 1: 93.7 WAcc (12 D, 80 S, 21 I)
EXP 3: 96.1 WAcc (8 D, 45 S, 17 I)

Spontaneous Commands (#= 226):
EXP 1: 82.3 WAcc (2 D, 22 S, 16 I)
EXP 3: 88.8 WAcc (1 D, 14 S, 10 I)

Conversational Speech (#=5140):
EXP 2: 29.5 WAcc
EXP 3: 56.3 WAcc

What can we learn from ASR errors?

Spotting of Austrian words

> In total 723 words from 347 different word types

➢ PROBLEM: no spelling conventions!

What can we learn from ASR errors?

From Substitutions: Specific for Austrian German

- Stressed /a/ to /o/ and stressed /i/ to /a/
  sind > dann
  ja > wo

- Monophthongations: /aI/ to /a/ and /au/ to /a/
  nein > ja; ein > an
  haut > grad
  glaub ich > hab ich

What can we learn from ASR errors?

From Deletions

1. Some also typical for spontaneous German:
   -- 20 most deleted words are all function words:
   ich 'I' 81.8% deleted;
   -- Reflecting consonant deletion:
   kannst du 'you can' > kann 'he can'
   einmal 'once' > mal 'times'

2. Multi-word expressions:
   das ist mir dann > dass man
   fahrst du schon > geforscht
   'do you leave already' 'research done'

Thank you for your attention!