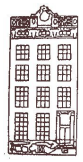
Speech recognition and synthesis



Introduction

There is no data like more data

- Speech and Language are extremely complex
- Large amounts of data are necessary to model them
- "The best application is the one with the largest corpus"
- 10-1000 hours of speech recordings needed
- 10^8 - 10^9 word text corpus needed

Introduction: Corpora for Speech and Language Technology

A language corpus is a documented collection of coherent text, speech, video, and transcriptions and annotations of these

Requirements

- Meta data (fixed)
- Normalization (fixed)
- Data (fixed)
- Transcriptions and annotations (cumulative)
- Storage, distribution, access, and software (volatile)

[Wynne(2005)]

Introduction: Corpora for S&L Technology

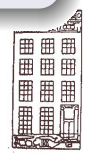
Requirements

- Meta data (fixed): Information on the items
  - Bibliographic/biographic information (author, speaker)
  - Dates
  - Origin, eg, place of publishing, recording
  - Language variant
  - Genre and style
  - Recording trail, post-processing, and formats
  - **Access criteria, Copyrights, Privacy&Ethical guidelines**
  - ...
- Normalization (fixed)
- Data (fixed)
- Transcriptions and annotations (cumulative)
- Storage, distribution, access, and software (volatile)

Introduction: Corpora for S&L Technology

Requirements

- Meta data (fixed)
- Normalization (fixed): All items must adhere to certain guidelines
  - Inclusion/selection criteria
  - Recording and text formats
  - Spelling rules, orthographic normalization
  - Storage formats (sample frequencies, file formats)
  - ...
- Data (fixed)
- Transcriptions and annotations (cumulative)
- Storage, distribution, access, and software (volatile)



Introduction: Corpora for S&L Technology

Requirements

- Meta data (fixed)
- Normalization (fixed)
- Data (fixed): Immutable text or speech records
  - Broadcast recordings
  - Speech recordings
  - Video recordings
  - Original text
  - Transliterations of speech (correctable)
  - ...
- Transcriptions and annotations (cumulative)
- Storage, distribution, access, and software (volatile)

Introduction: Corpora for S&L Technology

Requirements

- Meta data (fixed)
- Normalization (fixed)
- Data (fixed)
- Transcriptions and annotations (cumulative): Added value of interpretations and analysis
  - Orthographic transcription (transliteration) of speech
  - Paragraph and sentence boundaries
  - Phonemic transcription
  - Prosodic transcription (eg, ToBI)
  - Part-of-Speech tagging
  - Lemmatization
  - Syntactic trees (treebank)
  - ...
- Storage, distribution, access, and software (volatile)

Introduction: Corpora for S&L Technology

Requirements

- Meta data (fixed)
- Normalization (fixed)
- Data (fixed)
- Transcriptions and annotations (cumulative)
- Storage, distribution, access, and software (volatile): Practical usage
  - Digital storage, what and where
  - On-line and/or media distribution (DVD)
  - Access policies (pricing, licenses)
  - Exploration software
  - Database tables
  - DBMS
  - Updates and policy
  - ...

Language corpora

Example corpora and their sizes

- IFA Corpus: 50 thousand words ($5\frac{1}{2}$ hours) [Van Son(2003)]
- Spoken Dutch Corpus (CGN): 9 million words (800 hours) [NTU(2004)]
- British National Corpus (BNC): 100 million words [BNC(1997)]
- Twente journal corpus: 300 million words (Dutch) [Ordelman(2002)]
- Tilburg text corpus: 600 million words (Dutch, unpublished?)
- COSMAS corpus archive: 1.8 billion words (German) [IDS(2005)]
- IFA Video Dialog corpus: conversations (5 hours)
  `http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/`

Language corpora: CGN [NTU(2004)]

Contents ($\frac{2}{3}$ Dutch, $\frac{1}{3}$ Flemish)

- 500 hours (5,650,000 words) recorded in The Netherlands
- 300 hours (3,250,000 words) in Flanders
- 4250 speakers
- 15 Styles/genres
- Field recordings with Sony Minidisk
- 16/16 and 8/8 kHz/bit encoding

Language corpora: CGN Styles and Genres

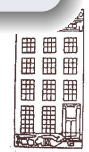
CGN: 9 million words from 800 hours of speech

Hour	kWords	Style
225	2,626	spontaneous conversations ('face-to-face')
51	565	interviews with teachers of Dutch (VNC)
92	1,209	spontaneous telephone dialogues
64	853	spontaneous telephone dialogues
11	136	simulated business negotiations
64	790	interviews/discussions/debates (broadcast)
36	360	discussions/debates/meetings (non-broad.)
44	405	lessons recorded in the classroom
21	208	live (eg sport) commentaries (broadcast)
17	186	newsreports/reportages (broadcast)
36	368	news (broadcast)
15	146	commentaries/columns/reviews (broadcast)
2	18	ceremonious speeches/sermons
16	141	lectures/seminars
104	903	read speech (read books)

Language corpora: CGN Annotations

Annotations and transcriptions

- Orthographic transcription (the full 8,900,000 words)
- Manually verified POS tagging and lemmatization (all)
- Lexicon and identification of multi word units (all)
- Automatic time alignment and phonetic transcription at the word level (all)
- Manually verified broad phonetic transcription (1,000,000 words)
- Manually verified time alignment at the word level (1,000,000 words)
- Syntactic annotation (1,000,000 words)
- Two independent prosodic annotations (250,000 words)



Use of corpora in Speech Technology: Research

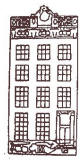
Phonetic, prosodic and syntactic research

- Phoneme durations
- Stress and Accent placement
- Intonation and expressive speech (emotions)
- Part-of-Speech tagging
- Prosodic and syntactic boundaries
- Phoneme assimilation (eg, word boundaries)
- Pronunciation variation
- Morphological decomposition
- Visual speech

Use of corpora in Speech Technology: TTS Modeling

Text to Speech synthesis

- Produce accentuation and boundaries from text
- Produce phoneme durations from text
- Grapheme-to-phoneme conversion (lexicon)
- Chunk words into groups (punctuation)
- Decompose words into components (compound words)



Use of corpora in Speech Technology: ASR Modeling

Automatic Speech Recognition

- Hidden Markov Model training
- Speech templates for template based recognition
- Language model (smoothed N-grams)
- Pronunciation variation
- Treebank training (syntactic probabilities)

Annotation, Segmentation, and labeling: Orthography

Manual Orthographic transcription (transliteration) is used to automatically generate

- Tokens (words) ⇒ Word alignment
- Phonemic transcription ⇒ Phone alignment
- POS tags

All annotations and segmentation can be manually verified (at great cost)

Annotation, Segmentation, and labeling: POS tagging

POS tags are used to automatically generate

- Lexical stress
- Syntactic structure
- Lemmas
- Prosodic structure (ToBI) $\Rightarrow$ **currently only by hand**

All annotations and segmentation can be manually verified (at great cost)

Phonetic symbols

Speech corpora needed an unambiguous digital encoding of IPA symbols (now there is UNICODE)

- Language specific encodings
  - 1 character ASCII encodings + diacritics (SAMPA)
  - 2 character ASCII encodings (SWITCHBOARD)
- Complete IPA encodings
  - 2 character ASCII encoding (eg, Worldbet [Hieronymus(1994)])
  - Control encodings (LaTeX Tipa, Praat)
- Currently, control encodings are impractical for manual labeling
- Note that mapping sounds to the IPA is *not* trivial

Phonetic symbols: CGN's SAMPA vs Worldbet encoding

Vowels IPA	CGN	Wbet	Example	Word
ɪ	I	'I '	lIp	lip
ɛ	E	'E '	lEx	leg
ɑ	A	'A '	lAt	lat
ɔ	O	'> '	bOm	bom
ʏ	Y	'ux'	pYt	put
i	i	'i '	lip	liep
y	y	'y '	byr	buur
e	e	'e '	lex	leeg
ø	2	'7 '	d2k	deuk
a	a	'a '	lat	laat
o	o	'o '	bom	boom
u	u	'u '	buk	boek
ə	@	'& '	x@-lE+k	gelijk
ɛi	E+	'Ei'	wE+s	wijs
øy	9+	'8y'	h9+s	huis
ɔu	O+	'Ou'	kO+t	koud

Assignment: Week 3 Manipulating prosody

Change intonation and duration

- Open sentence in praat (eg, assignment 1/2)
- Create a Word tier (Help → Praat Intro → Intro 7. Annotation)
- Add the (aligned) words to the tier
- Copy to a Phoneme tier
- Then add (split into) the phonemes
- Create a manipulation (Help → Praat Intro → Intro 8. Manipulation)
- Move the stress(-es) to a different word(s)
- What are the contributions of intonation, duration, or intensity?
- Hand in your report as a PDF

Further Reading I

BNC.
Britisch National Corpus.
Corpus, 1997.
URL http://www.natcorp.ox.ac.uk/.

P. Boersma.
Praat, a system for doing phonetics by computer.
Glot International, 5:341–345, 2001.
URL http://www.Praat.org/.

P. Boersma and D. Weenink.
Praat 4.2: doing phonetics by computer.
Computer program: http://www.Praat.org/, 2004.
URL http://www.Praat.org/.

FSF.
GNU General Public License.
Web, June 1991.
URL http://www.gnu.org/licenses/gpl.html.

James L. Hieronymus.
Ascii phonetic symbols for the world's languages: Worldbet.
Web, 1994.
URL http://www.ling.ohio-state.edu/~edwards/worldbet.pdf.

Further Reading II

IDS.
COSMAS.
Corpus, 2005.
URL http://corpora.ids-mannheim.de/~cosmas/.

NTU.
Spoken Dutch Corpus (CGN).
Corpus, 2004.
URL http://www.tst.inl.nl/cgn.htm.
Metadata (MPI) - http://www.mpi.nl/world/ISLE/overview/Overview_CGN.html Contents - http://www.elis.ugent.be/cgn/ Descriptions and references - http://lands.let.ru.nl/cgn/ehome.htm.

Roeland Ordelman.
Twente Nieuws Corpus (TwNC).
Corpus, 2002.
URL http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html.

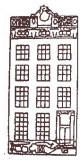
R.J.J.H. Van Son.
IFA corpus 1.0.
Corpus, 2003.
URL http://www.fon.hum.uva.nl/Service/IFAcorpus.

D. Weenink, G. Chen, Z. Chen, S. de Konink, D. Vierkant, E. van Hagen, , and R.J.J.H. van Son.
Learning tone distinctions for Mandarin Chinese.
In *Proceedings of INTERSPEECH 2007*, pages 950–953, Antwerp, Belgium, August 2007.
URL http://www.fon.hum.uva.nl/rob/Publications/p950I07_WeeninkEtAl2007.pdf.

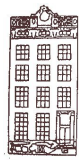
Further Reading III

M Wynne, editor.

Developing Linguistic Corpora: a Guide to Good Practice.
Oxford: Oxbow Books, 2005.
URL `http://ahds.ac.uk/linguistic-corpora`.
Accessed 2007-09-04.



Appendix A



Copyright License