The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database

R.J.J.H. van Son[1], *Diana Binnenpoorte*[2], *Henk van den Heuvel*[2], *and Louis C.W. Pols*[1]

Email: Rob.van.Son@hum.uva.nl

[1]Institute of Phonetic Sciences (IFA) / ACLC, University of Amsterdam, the Netherlands

[2]SPEX/A2RT, Nijmegen University, the Netherlands

ABSTRACT

– Open source database
– Hand-segmented Dutch speech
– 8 speakers
– 8 speaking styles
– 50,000 words / 5 1/2 hours
– Speech preparation:
~ 3 person-weeks per speaker
– 1,000 hours of hand labeling
– Asymptotic segmentation speed:
1 word / minute *or* 4 boundaries / minute
– *Median Absolute Difference*:
6 ms between labelers
4 ms within labelers
– Substitutions, insertions, and deletions:
8% between labelers
5% within labelers
– Compiled data are available On-Line (Web, DBMS) for querying with SQL.

AIMS OF CORPUS

General phonetics research

– Hand labeled and segmented at the phoneme level
– Speech from several speakers
– Many speaking styles are covered, including "spontaneous" speech
– Overlapping "content", i.e., identical sentences uttered in several styles

Designed to be

– Reusable
– Searchable
– Accessible
– Extendable
– Correctable
– Free(GNU GPL)

CORPUS SIZE

Recorded and segmented sentences:
5 1/2 hours of segmented speech

(net time in seconds)

Sex	Age	ID	Recorded sentences	Segmented sentences	
F	20	N	3,736	2,760	(seconds)
F	28	G	4,180	3,978	
F	40	L	3,112	2,485	
F	60	E	4,181	3,245	
M	15	R	2,125	1,439	
M	40	K	2,720	1,891	
M	56	H	2,894	2,368	
M	66	O	3,781	1,696	
Total	-	-	**26,733**	**19,867**	(seconds)
			7:26'	**5:31'**	(hours)

Items: 50,000 words and 190,000 phonemes

Speaker sex	Recorded Sent	Words	Segmented Sent	Words	Syllables	Phonemes	
4 F/4 M	**6,128**	**73,067**	**4,492**	**51,782**	**74,702**	**187,544**	(excluding all pauses)

2,000 - 15,000 Words per speaking style

	Speaking Style Informal	Retold	Text	Sent.	Pseu	Word	Syll	Pronunciation	(segmented only)
# Words	**5,262**	**6,256**	**14,577**	**15,437**	**2,608**	**1,984**	**2,282**	**3,370**	(excluding all pauses)
%	*10.2*	*12.1*	*28.2*	*29.8*	*5.0*	*3.8*	*4.4*	*6.5*	
Sy	*5.5*	*5.2*	*5.7*	*5.6*	*4.6*	*3.5*	*2.4*	*3.5*	*(Syllable rate/sec)*
Ph	*13.5*	*13.1*	*1 4.4*	*14.3*	*12.2*	*9.3*	*6.7*	*6.3*	*(Phoneme rate/sec)*

LABELING EFFORT

Hand correction of automatically (HMM) pre-aligned labels by 7 naive labelers trained for this specific task

Optimum labeling speed reached after
40 hours of transcription
Top speed was **1 +/– 0.2 words per minute**

Total amount of labeling work:
50,000 words and **200,000** segment boundaries

Total time of labeling:
1000 hours (manual labeling)
6 person-months (staff-time overhead etc.)

Average labeling speed:
0.84 words / minute
3.3 boundaries / minute

Without pre-alignment, labeling speed approximately halves

Monetary cost of alignment:

(excluding VAT, automatic and manual alignment combined)

DFl **74,000** Total (= **33,597** €)
DFl **1.40** / word (= **0.65** €)
DFl **0.37** / boundary (= **0.17** €)

LABELING CONSISTENCY

Procedure:

– 4 Labelers participated
– 64 sentences selected for re-alignment
– *Pairwise* labeling differences only, using DTW alignment
– Ignoring differences larger than 100 ms

Between **labelers**

Median Absolute Difference		**6 ms**
	75%	**15 ms**
	95%	**46 ms**
Phoneme Substitutions		**3%**
Insertion/Deletions		**5%**

Within **labelers**

Median Absolute Difference		**4 ms**
	75%	**10 ms**
	95%	**31 ms**
Phoneme Substitutions		**2%**
Insertion/Deletions		**3%**

Unresolved boundaries (total): **3.5%** (7000)

Conclusion

Consistency is well within published standards.
Probably due to automatic pre-alignment.
There is a long "tail" of differences (DTW errors).

SPEAKERS

- **8 Speakers 4 male 4 female**
  (selected from 18 recorded speakers)
- **Ages 15 - 66 y**
- **Diverse regional background**
- **"Standard" Dutch**
  (region of origin is audible)

8 speaking "styles":

1. **Informal (I)**
   Story telling face-to-face to an "interviewer"
2. **Retold (R)**
   Retelling a previously read story without sight contact
3. **Text reading (T)**
   Reading aloud a narrative story, includes Informal story
4. **Sentence list (S)**
   A random list of all sentences of the narrative stories
5. **Pseudo-sentences (Ps)**
   Constructed by replacing all words in a sentence with randomly selected words from the text with the same POS tag
6. **Word list (W)**
   Lists of selected words from the texts
7. **Syllable list (Sy)**
   Lists of all distinct syllables from the word lists
8. **Pronunciation lists (Pr)**
   A collection of idiomatic (the Alphabet, the numbers 0-12) and "diagnostic" sequences (vowels, /hVd/ and /VCV/ lists)

SPECIAL FEATURES

- Collection of speech and data follows "best practices" (Eagles Handbook)
- 5 1/2 hours speech from 8 speakers
- Wide band recordings (audio CD)
- Intermediate data preserved for extension and correction
- Male and female speakers from matched age groupes
- Extensive meta-data on speakers
- Wide range of speaking styles with overlapping content
- Overlapping textual materials (narrative to idiomatic)
- Hand segmentation and phonemic labeling
- Free on-line access (speech&beer)
- Full SQL querying

FREE ACCESS (OPEN SOURCE)

The IFA corpus is licensed under the GNU General Public License

Freedom to:
–> Copy
–> Use
–> Modify
–> Distribute

(provided you license all derived works under the GNU GPL
see http://www.gnu.org/gpl.html)

This license covers ALL material needed to (re-)build the corpus including scripts and web site

NOTE: The IFA corpus comes without any warranty
(see license for details)

Copyrights to the IFA corpus rest with the "Dutch Language Organization" (Nederlandse Taalunie)

SPEECH CODING & FORMAT

Broadband speech: 44.1 kHz 16 bit

– Quiet, sound treated recording room
– All equipment in a separate control room
– Subject reads from a computer controlled cueing screen (sound treated CRT)
– Two-channel recordings:
  head-mounted dynamic microphone
  fixed HF condenser microphone
– Philips Audio CD-recorder
  (16 bit linear coding at 44.1 kHz stereo)
– 78 dB Standard calibrating sound source
  (white noise and pure 400 Hz tone)

Standard formats: AIFC Ogg Praat

– Paragraph sized chunks (AIFC, Ogg Vorbis)
– Sentence sized speech files (AIFC)
– Unaligned orthographic Transcriptions (ASCII)
– Alligned transcriptions in Praat Label files
  (orthographic, phonetic, syllables, POS, etc.)
– Derived data in Praat format:
  F_0, F_1-F_3, Intensity, Center of Gravity, etc.

CORPUS ACCESS

Current corpus size: **~20 GB, ~100,000 files** (~ 40 CDrom's)
The corpus is dynamic
Distribution on any medium becomes a problem

INTERNET BASED ACCESS IS INEVITABLE

Static WWW browsing and downloads:

(HTTP://www.fon.hum.uva.nl/IFAcorpus)

Basic download services for *ALL* files
(**~20 GB, ~100,000 files** as of August 2001)
This includes all scripts, programs, and intermediate results

Dynamic WWW query directed access:

(HTTP://www.fon.hum.uva.nl/IFAcorpus)

Based on SQL searchs
– "Raw" speech (sentences)
– Audio fragments of any size (sentences - phonemes)
– Texts and compressed paragraph sized audio files
– Listings and Descriptive statistics of labelled data

PostgreSQL querying:

Data is stored on-line with full SQL capabillities.
Anonymous access currently disabled (available on request).
Limited SQL capabillities using the WWW front-end.

Concurrent Version System (CVS):

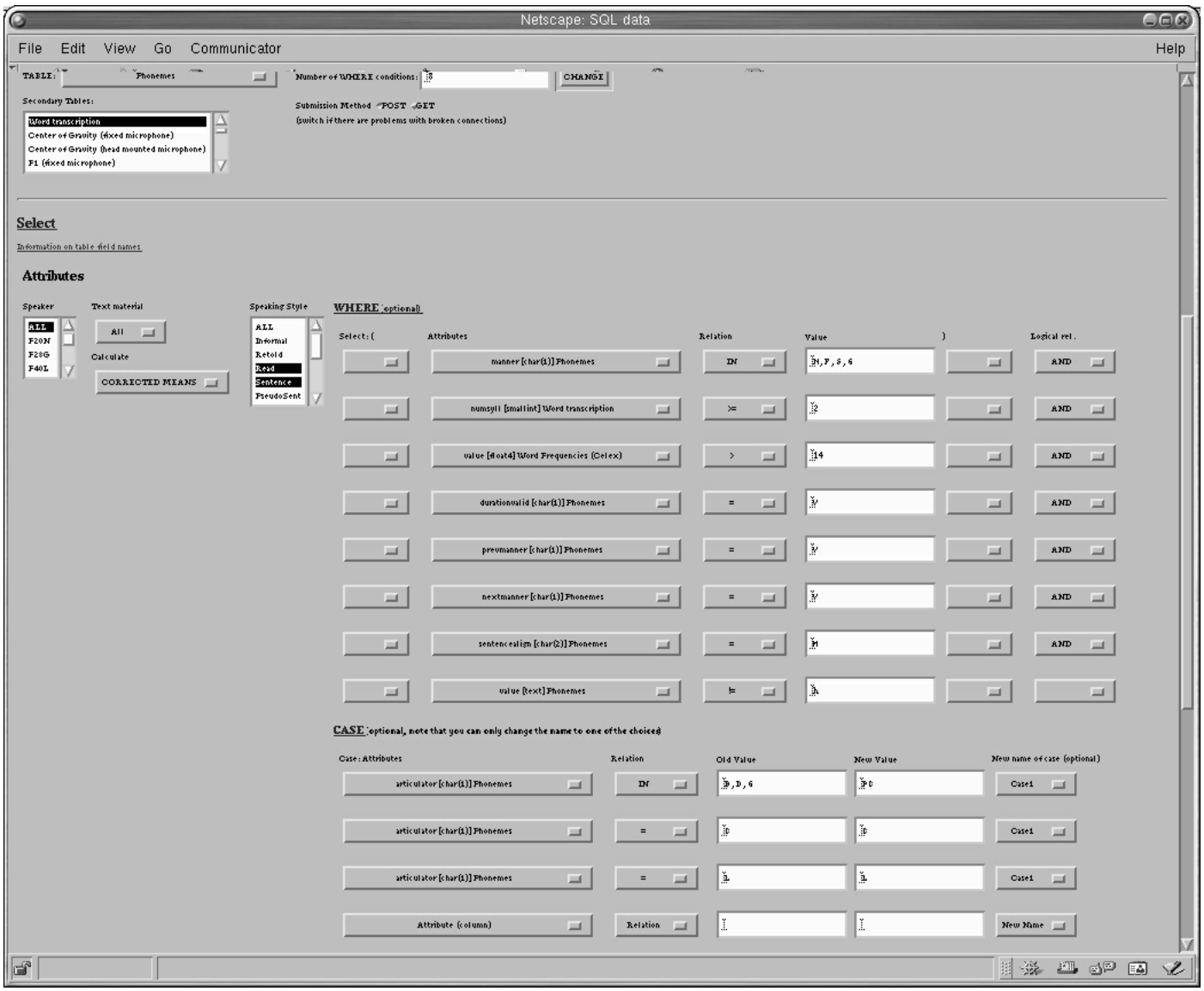
(anonymous@uvafon.hum.uva.nl:/u/cvs, password anonymous, module SLcorpus)

Text based materials are stored in a CVS repository.
This includes all label files, scripts, and database tables.
(**~415 MB, ~39,000 files**)

WEB INTERFACE

PLAIN HTML, ON-LINE

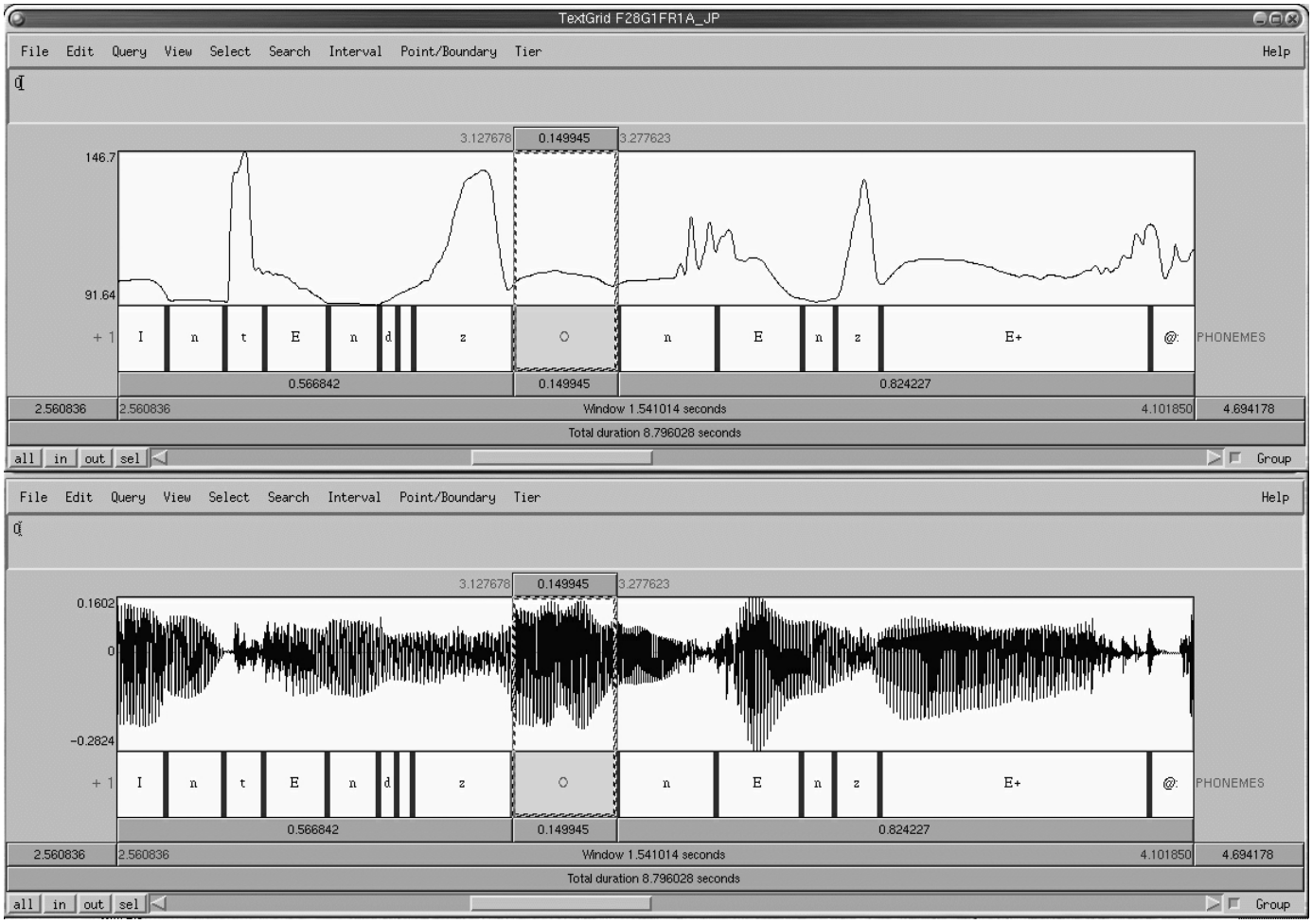
Complex SQL querying is possible on-line with plain HTML web browsers
(input for the EXAMPLE query)



NOTE:
The web site is part of the corpus distribution

SEGMENTING AND LABELING INTERFACE

Labeling based on Waveform, Spectral Center of Gravity, and Audio feed-back



Label and segmenting work-flow was automated using the Praat scripting features

On-line manual and help were available

(RE-)USABILITY

Download selected speech fragments

Items on any linguistic level can be selected and corresponding sound fragments can be downloaded together with corresponding label-file fragments
(AIFC/AIFF, NIST, Next/Sun, WAV formats)

Research can be done directly on the query results

Currently supported:

1) ASCII listing for import into other applications (PSPP, SPSS, R, S)

2) Frequency counts (table format)

3) Mean values and Standard deviations (simplistic ANOVA)

4) Correlations (Pearson product moment)

5) Corrected Means Analysis and statistics (generalized ANOVA)

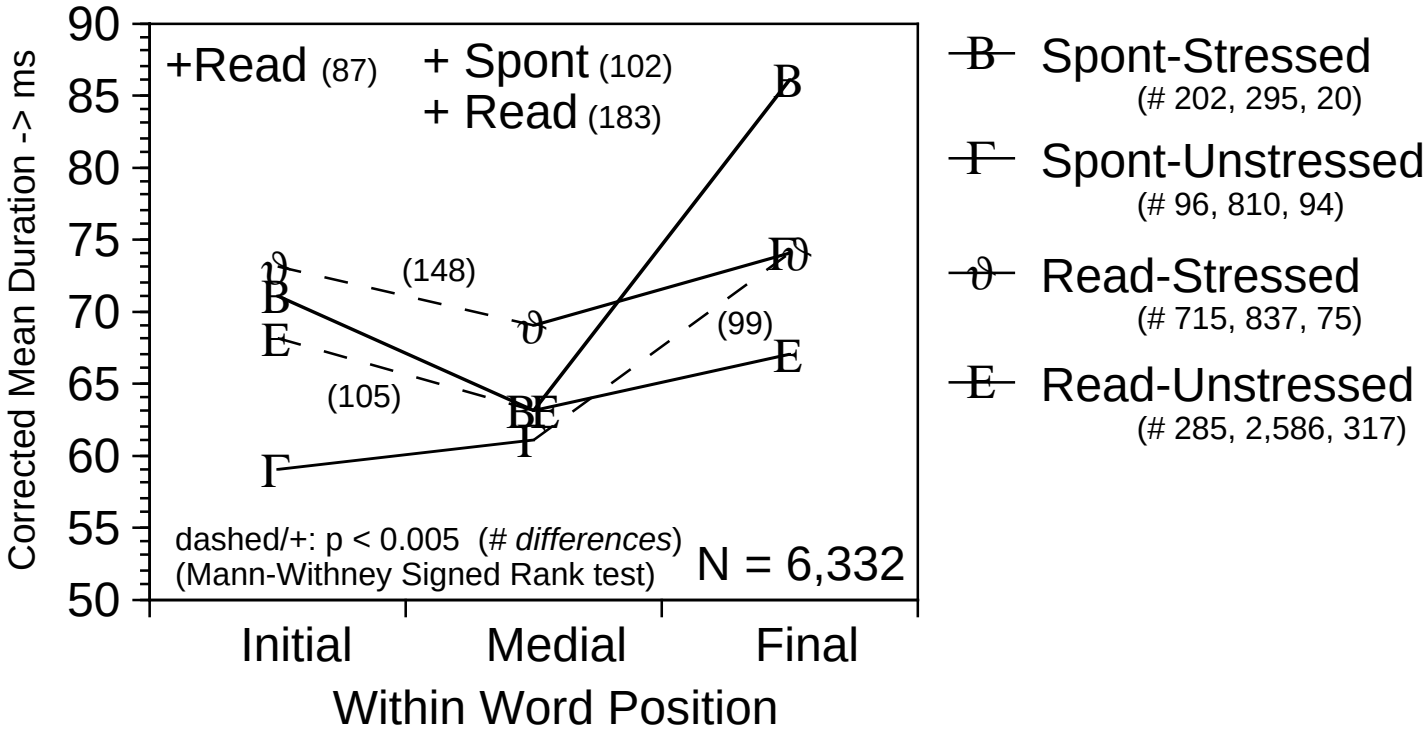
EXAMPLE

Complex analysis is possible on query results

Corrected means analysis

(generalized ANOVA)
The generalized (corrected) effect of:
Spontaneous vs Read speech, *position in the word* and *syllable stress* on *phoneme* ***duration***
Accounting for the effects of *speaker*, *specific style*, and *phoneme identity* (nuisance factors)

QUERY:
All speakers and text types
Spontaneous speech, Read texts and sentences
Intervocalic Nasals, Fricatives, Stops, and Glides (not glottal)
Polysyllabic words
Word frequency < 1/4000
Words not on sentence boundaries



EXTENDABILITY

(LIMITED ONLY BY STORAGE CAPACITY)

Content neutral data storage

Data type		Storage format
Aligned "symbolic" data (e.g., annotations)	→	**Label file tiers (XML compatible)**
Continuous data	→	**Acoustic analysis files**
Combined Relational data	→	**Relational database tables**

Automatic conversions

Combine any raw data		Database tables
Label file tiers (XML compatible)	→	**Symbolic data**
Acoustic analysis files	→	**Analysis data per item**

Possible extensions:

A) Shadowing (in progress)
B) Extensive POS information (in progress)
C) Prosodic annotation
D) Language modeling (word probabilities)

CORRECTIONS

Complete collection of all relevant files is preserved

Annotations and labels stored in an on-line CVS repository:

– Corrections can be applied reversibly
– Previous versions remain available
– Branching is possible
– Change history identifies
"*who* changed *what*, *when*, and *why*"

Public history file with contact information of contributors

All scripts and programs used are available

Special (web-based) tools for adding and correcting label files and annotations

CONCLUSIONS

- The internet is a must for distributing a 50,000 word labeled and segmented corpus

- Corpora need genuine databases with powerful querying possibilities

- Web based SQL querying of the corpus can be a valuable alternative distribution channel

- Querying and selection should be augmented by descriptive statistics

- **"Open Source" licenses simplify both the *construction* and the *use* of corpora**

- Labeling/segmentation costs were (excl. VAT)
  ~ **0.65** €/Word
  or
  ~ **0.17** €/Boundary