MODERN LINGUISTIC THEORIES

SPOKEN CORPORA

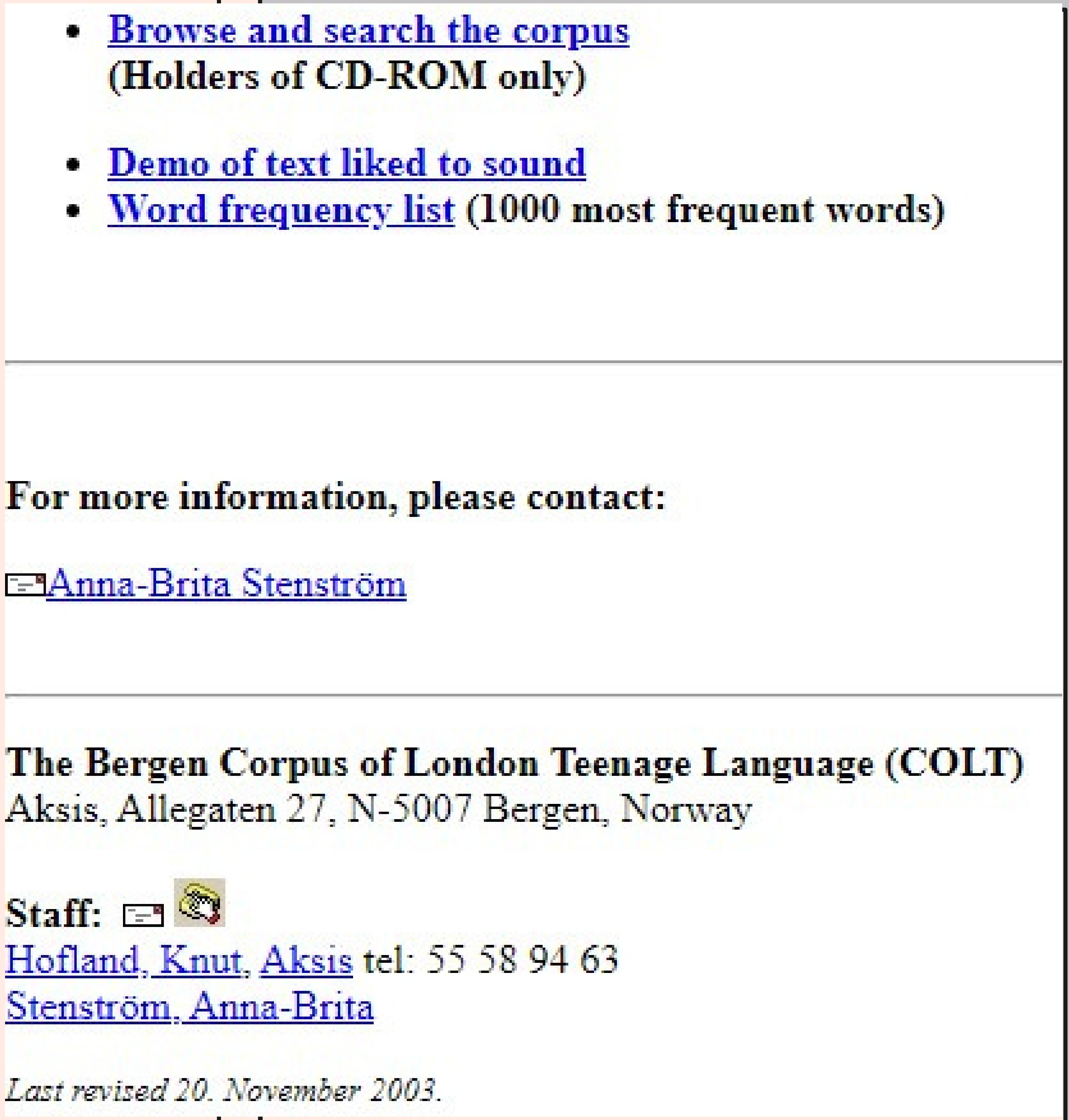
Karakina Maria Fedorovna
LMAG-101

TYPES OF SPOKEN CORPORA

Read Speech

Spontaneous Speech

THE BERGEN CORPUS OF LONDON TEENAGE LANGUAGE (COLT)



- Browse and search the corpus
  (Holders of CD-ROM only)

- Demo of text liked to sound
- Word frequency list (1000 most frequent words)

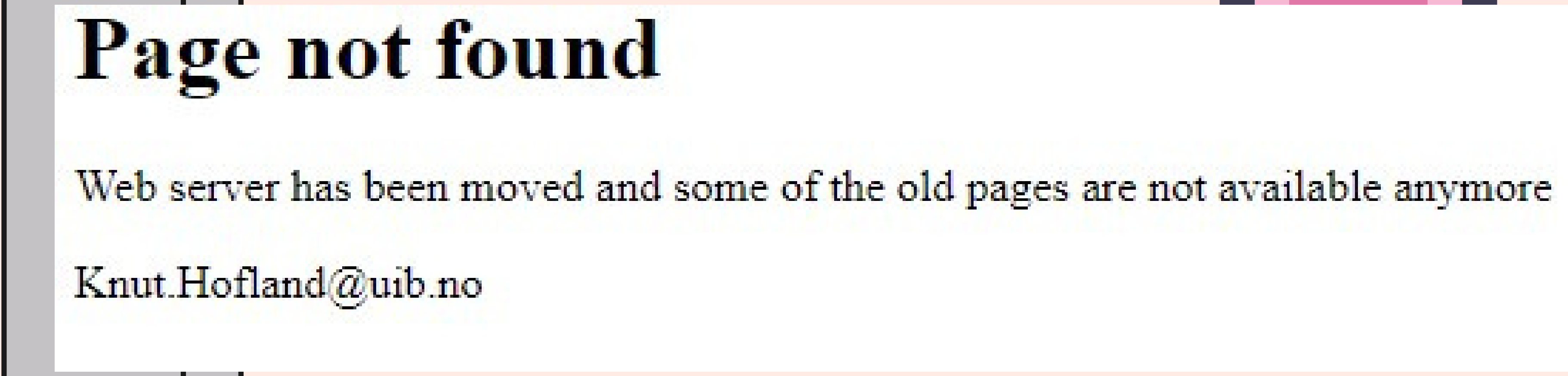
For more information, please contact:

Anna-Brita Stenström

The Bergen Corpus of London Teenage Language (COLT)
Aksis, Allegaten 27, N-5007 Bergen, Norway

Staff:
Hofland, Knut, Aksis tel: 55 58 94 63
Stenström, Anna-Brita

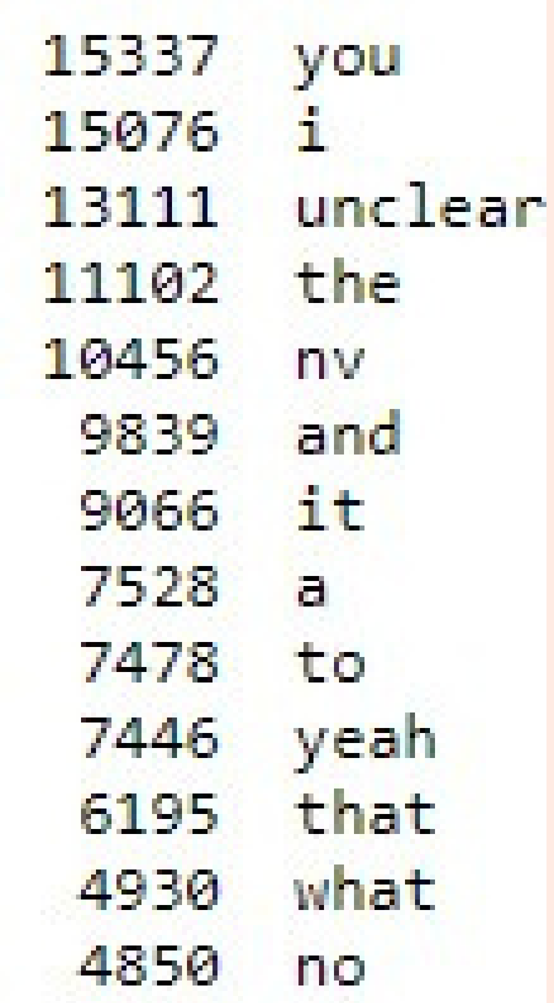
Last revised 20. November 2003.



Page not found

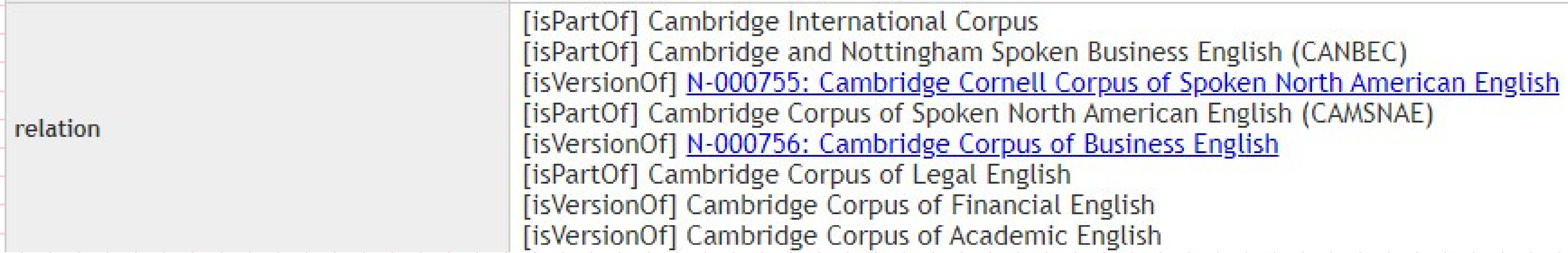
Web server has been moved and some of the old pages are not available anymore

Knut.Hofland@uib.no

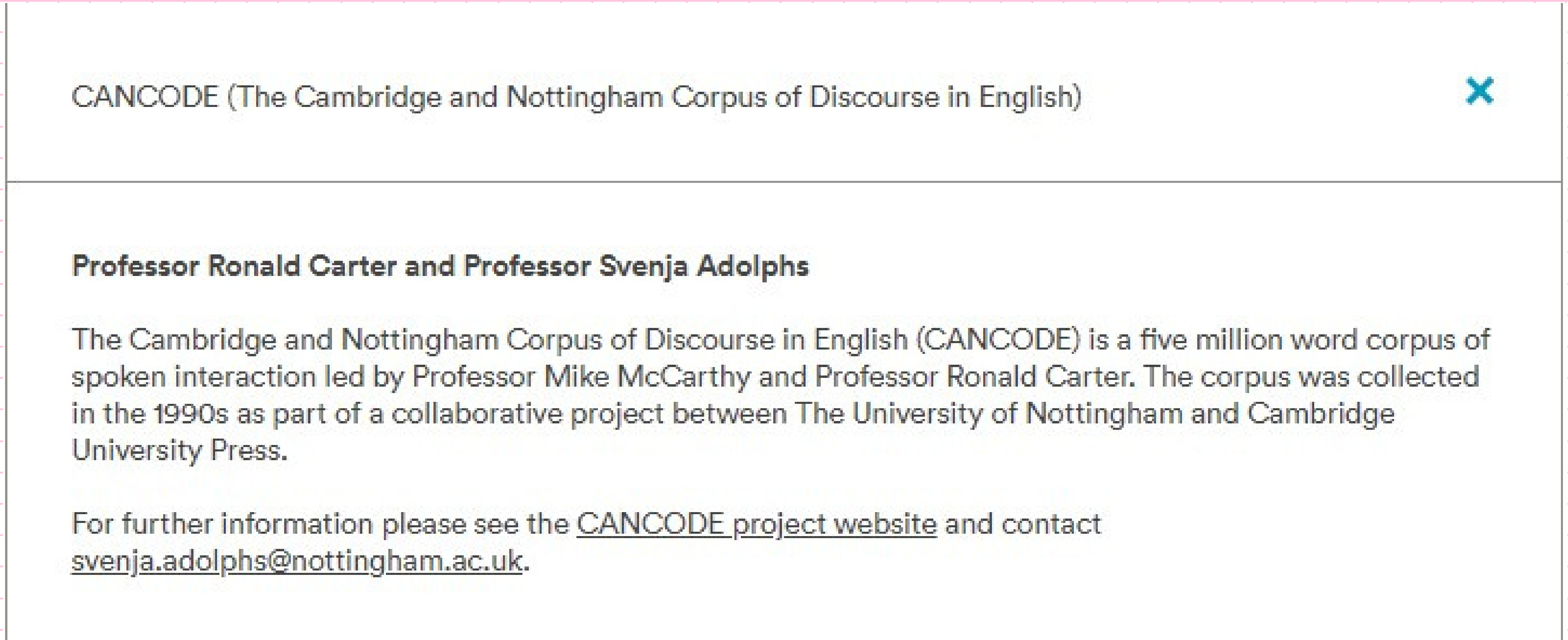


```
15337  you
15076  i
13111  unclear
11102  the
10456  nv
 9839  and
 9066  it
 7528  a
 7478  to
 7446  yeah
 6195  that
 4930  what
 4850  no
```

CAMBRIDGE AND NOTTINGHAM CORPUS OF DISCOURSE IN ENGLISH



relation	[isPartOf] Cambridge International Corpus [isPartOf] Cambridge and Nottingham Spoken Business English (CANBEC) [isVersionOf] N-000755: Cambridge Cornell Corpus of Spoken North American English [isPartOf] Cambridge Corpus of Spoken North American English (CAMSNAE) [isVersionOf] N-000756: Cambridge Corpus of Business English [isPartOf] Cambridge Corpus of Legal English [isVersionOf] Cambridge Corpus of Financial English [isVersionOf] Cambridge Corpus of Academic English



CANCODE (The Cambridge and Nottingham Corpus of Discourse in English) ×

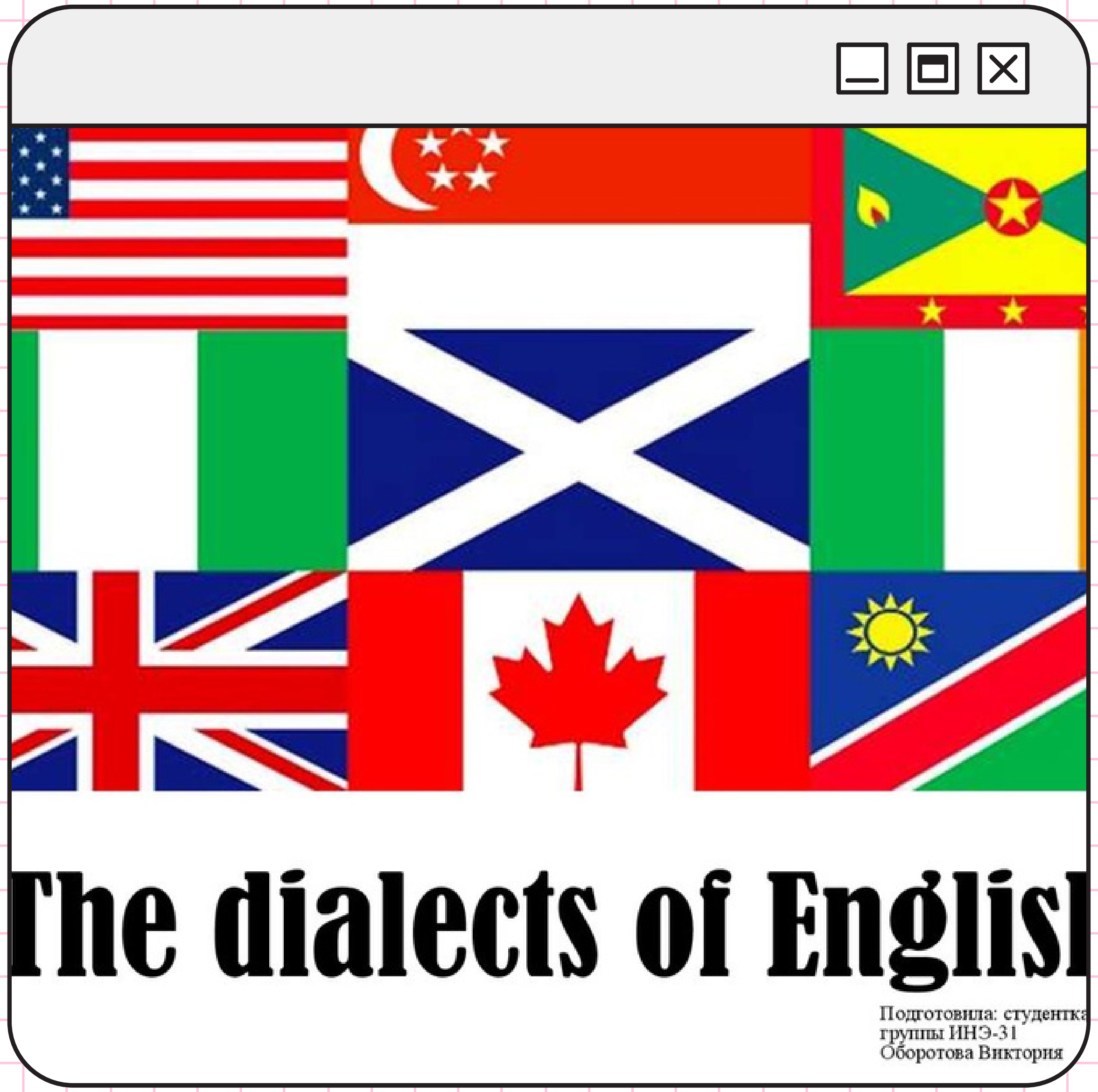
Professor Ronald Carter and Professor Svenja Adolphs

The Cambridge and Nottingham Corpus of Discourse in English (CANCODE) is a five million word corpus of spoken interaction led by Professor Mike McCarthy and Professor Ronald Carter. The corpus was collected in the 1990s as part of a collaborative project between The University of Nottingham and Cambridge University Press.

For further information please see the CANCODE project website and contact svenja.adolphs@nottingham.ac.uk.

THE SPOKEN CORPUS OF THE SURVEY OF ENGLISH DIALECTS

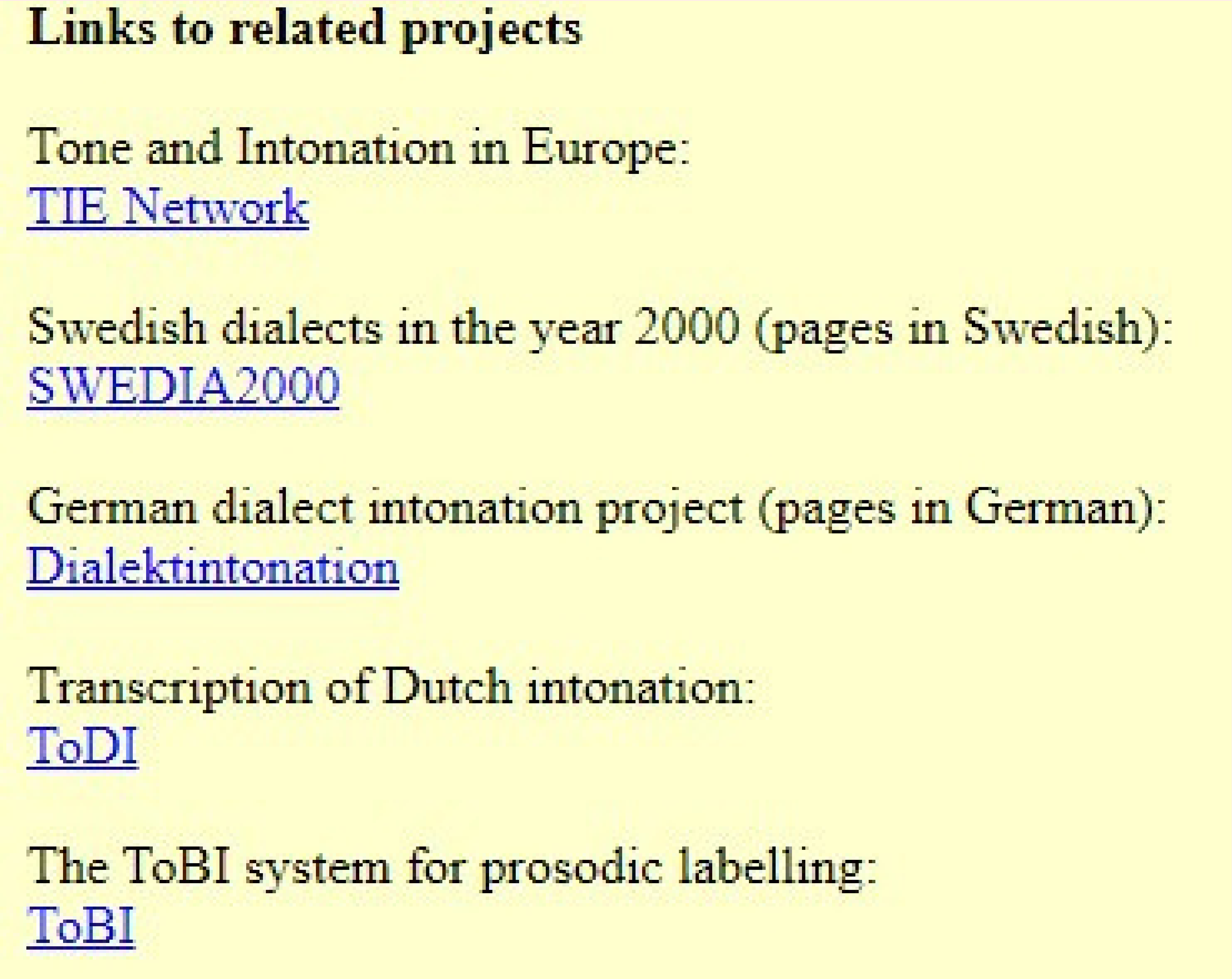
The spoken corpus consists of transcripts of 314 recordings from 289 (out of the 313) SED localities in England, totaling roughly 800,000 running words.



The dialects of Englisl

Подготовила: студентка
группы ИНЭ-31
Оборотова Виктория

THE INTONATIONAL VARIATION IN ENGLISH CORPUS



Links to related projects

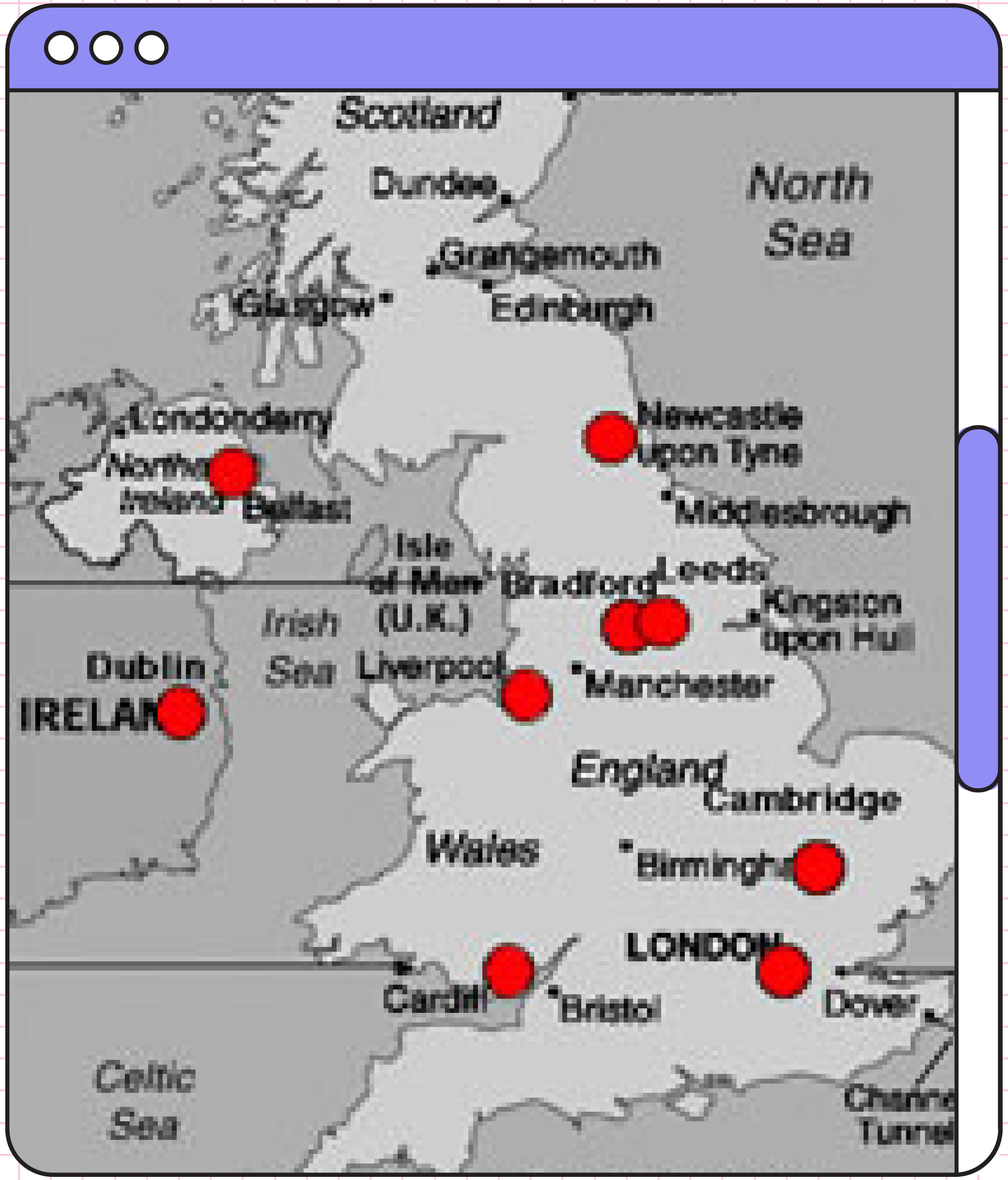
Tone and Intonation in Europe:
TIE Network

Swedish dialects in the year 2000 (pages in Swedish):
SWEDIA2000

German dialect intonation project (pages in German):
Dialektintonation

Transcription of Dutch intonation:
ToDI

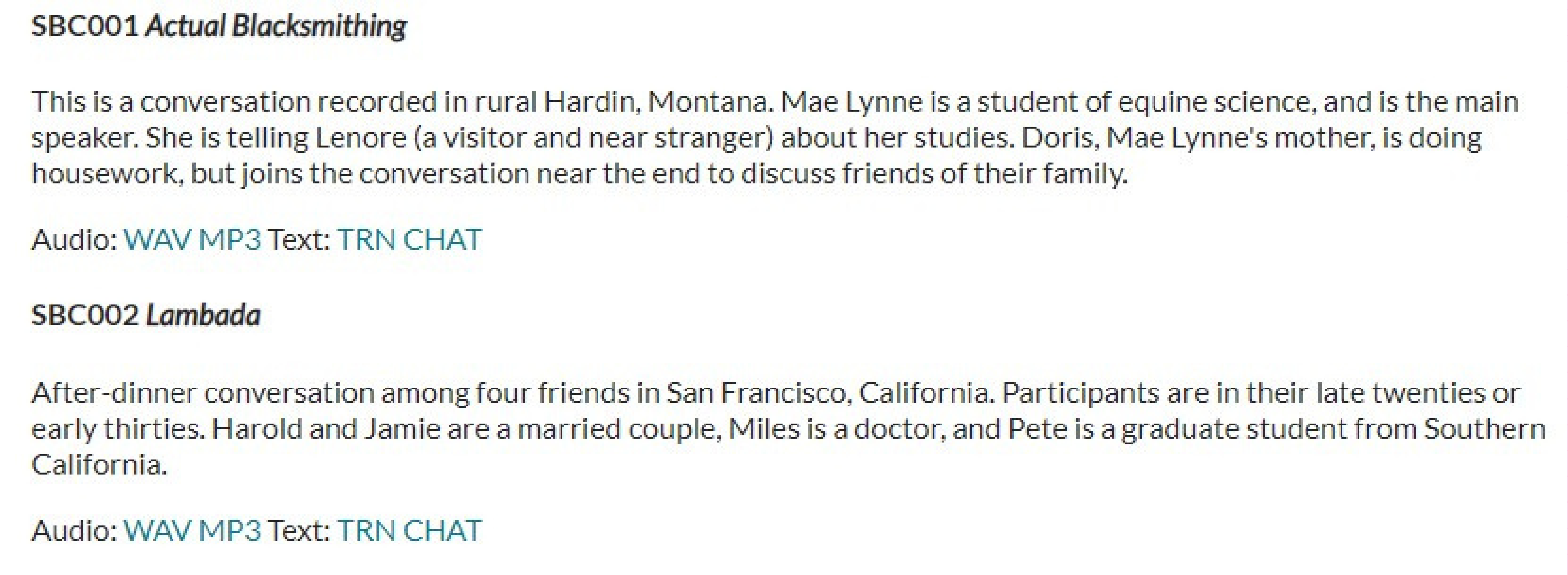
The ToBI system for prosodic labelling:
ToBI



THE LONGMAN SPOKEN AMERICAN CORPUS

Not open to the public

THE SANTA BARBARA CORPUS OF SPOKEN AMERICAN ENGLISH



SBC001 ***Actual Blacksmithing***

This is a conversation recorded in rural Hardin, Montana. Mae Lynne is a student of equine science, and is the main speaker. She is telling Lenore (a visitor and near stranger) about her studies. Doris, Mae Lynne's mother, is doing housework, but joins the conversation near the end to discuss friends of their family.

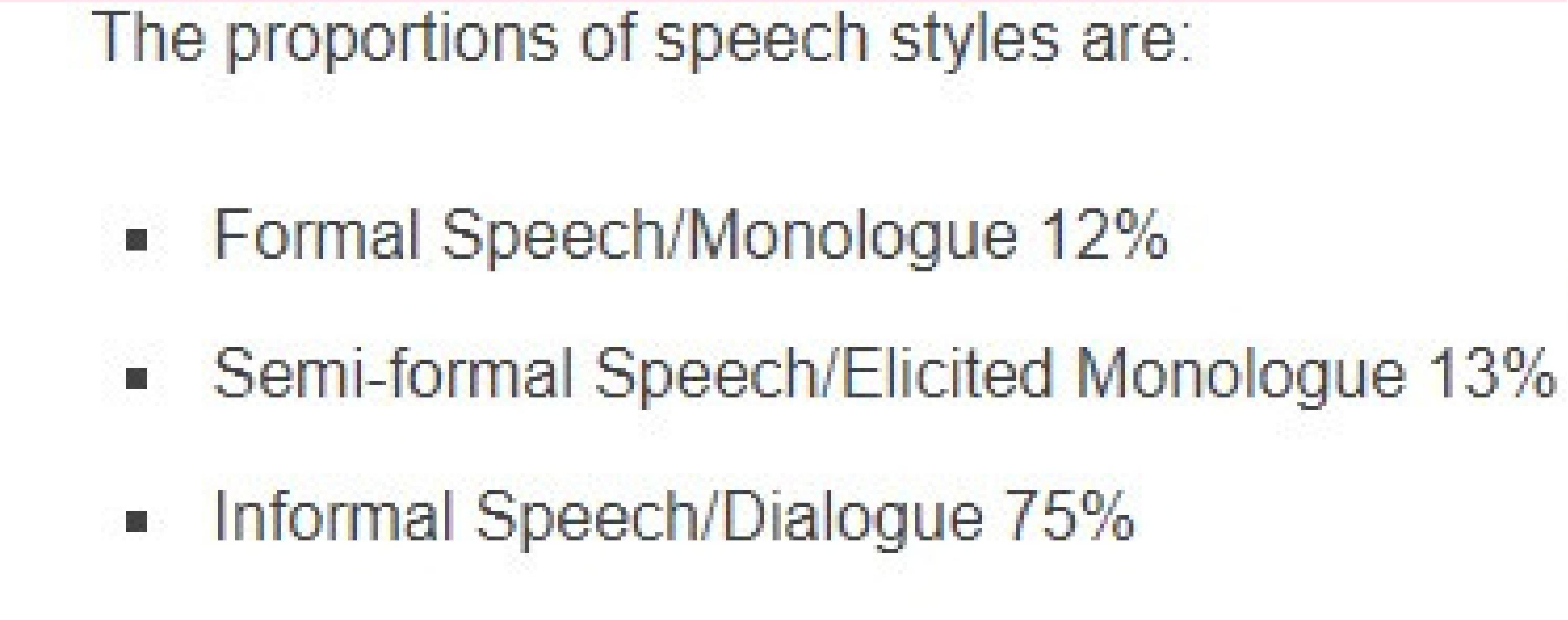
Audio: WAV MP3 Text: TRN CHAT

SBC002 ***Lambada***

After-dinner conversation among four friends in San Francisco, California. Participants are in their late twenties or early thirties. Harold and Jamie are a married couple, Miles is a doctor, and Pete is a graduate student from Southern California.

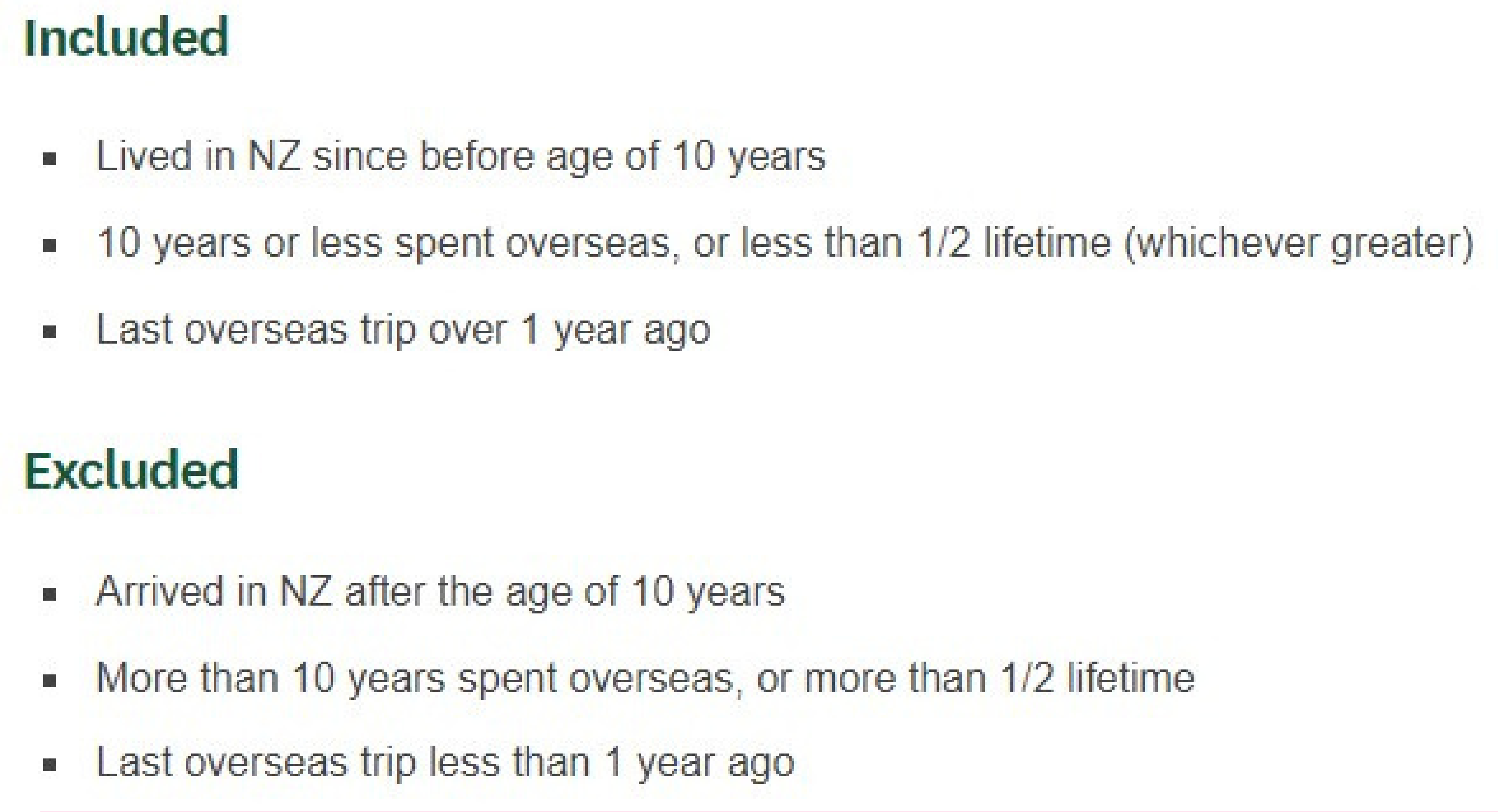
Audio: WAV MP3 Text: TRN CHAT

THE WELLINGTON CORPUS OF SPOKEN NEW ZEALAND ENGLISH



The proportions of speech styles are:

- Formal Speech/Monologue 12%
- Semi-formal Speech/Elicited Monologue 13%
- Informal Speech/Dialogue 75%

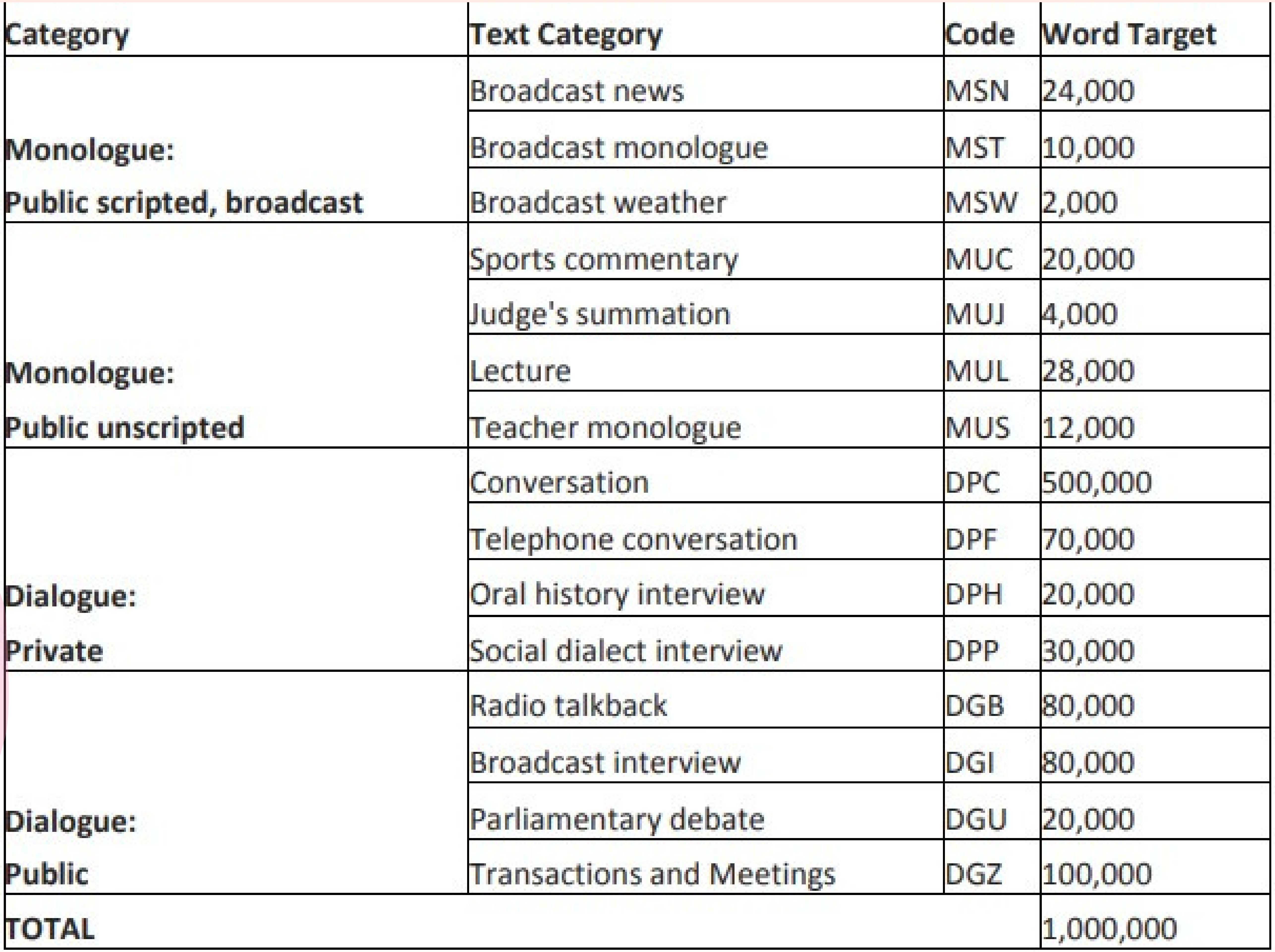


Included

- Lived in NZ since before age of 10 years
- 10 years or less spent overseas, or less than 1/2 lifetime (whichever greater)
- Last overseas trip over 1 year ago

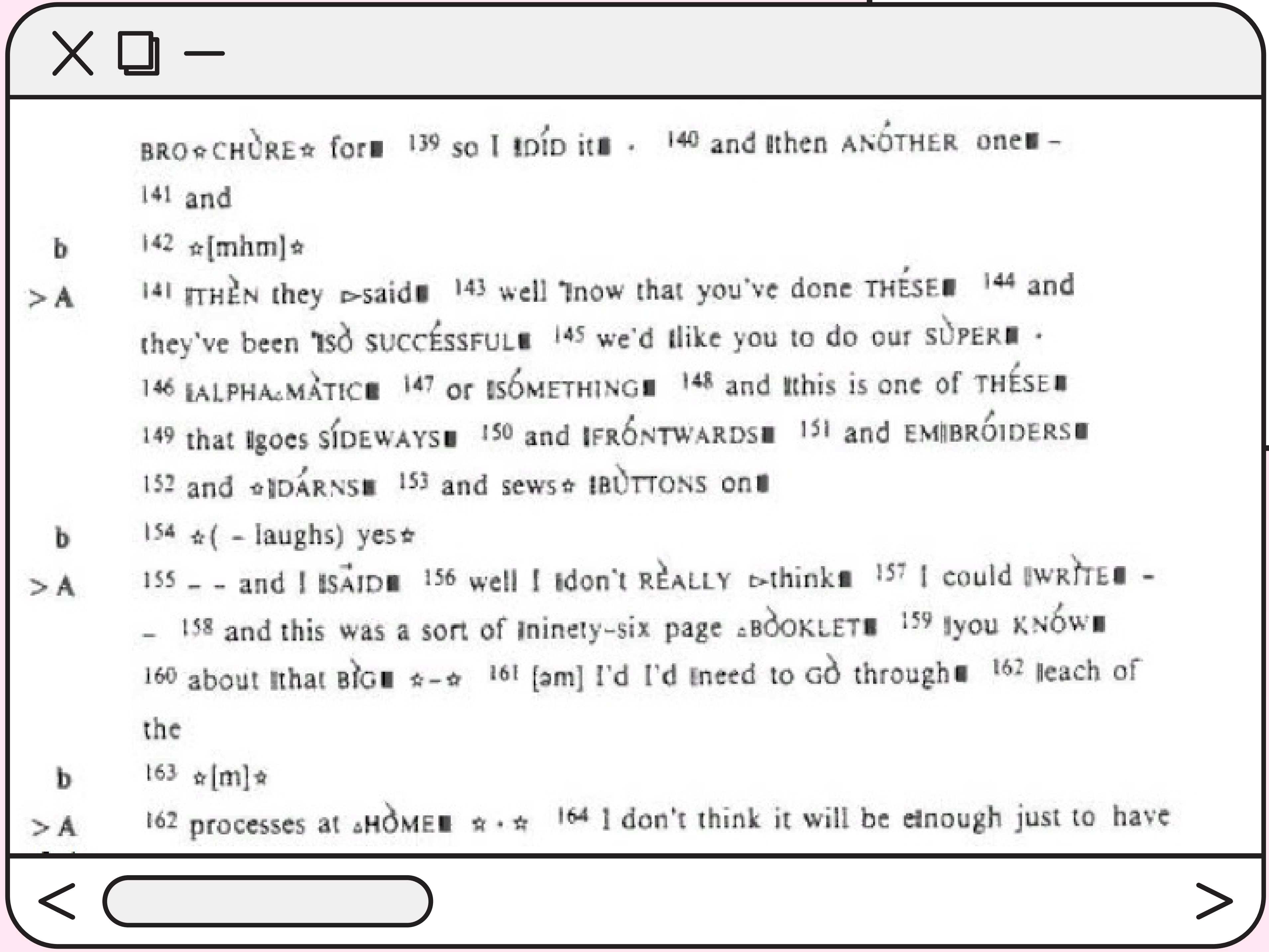
Excluded

- Arrived in NZ after the age of 10 years
- More than 10 years spent overseas, or more than 1/2 lifetime
- Last overseas trip less than 1 year ago



Category	Text Category	Code	Word Target
Monologue: Public scripted, broadcast	Broadcast news	MSN	24,000
	Broadcast monologue	MST	10,000
	Broadcast weather	MSW	2,000
Monologue: Public unscripted	Sports commentary	MUC	20,000
	Judge's summation	MUJ	4,000
	Lecture	MUL	28,000
	Teacher monologue	MUS	12,000
Dialogue: Private	Conversation	DPC	500,000
	Telephone conversation	DPF	70,000
	Oral history interview	DPH	20,000
	Social dialect interview	DPP	30,000
Dialogue: Public	Radio talkback	DGB	80,000
	Broadcast interview	DGI	80,000
	Parliamentary debate	DGU	20,000
	Transactions and Meetings	DGZ	100,000
TOTAL			1,000,000

THE LONDON-LUND CORPUS OF SPOKEN ENGLISH



BRO☆CHÙRE☆ for■ 139 so I ‖DÍD it■ · 140 and ‖then ANÓTHER one■ –
141 and

b 142 ☆[mhm]☆

> A 141 ‖THÈN they ▹said■ 143 well ˈ‖now that you've done THÉSE■ 144 and
they've been ˈ‖SÒ SUCCÉSSFUL■ 145 we'd ‖like you to do our SÙPER■ ·
146 ‖ALPHA▵MÀTIC■ 147 or ‖SÓMETHING■ 148 and ‖this is one of THÉSE■
149 that ‖goes SÍDEWAYS■ 150 and ‖FRÓNTWARDS■ 151 and EM‖BRÓIDERS■
152 and ☆‖DÁRNS■ 153 and sews☆ ‖BÙTTONS on■

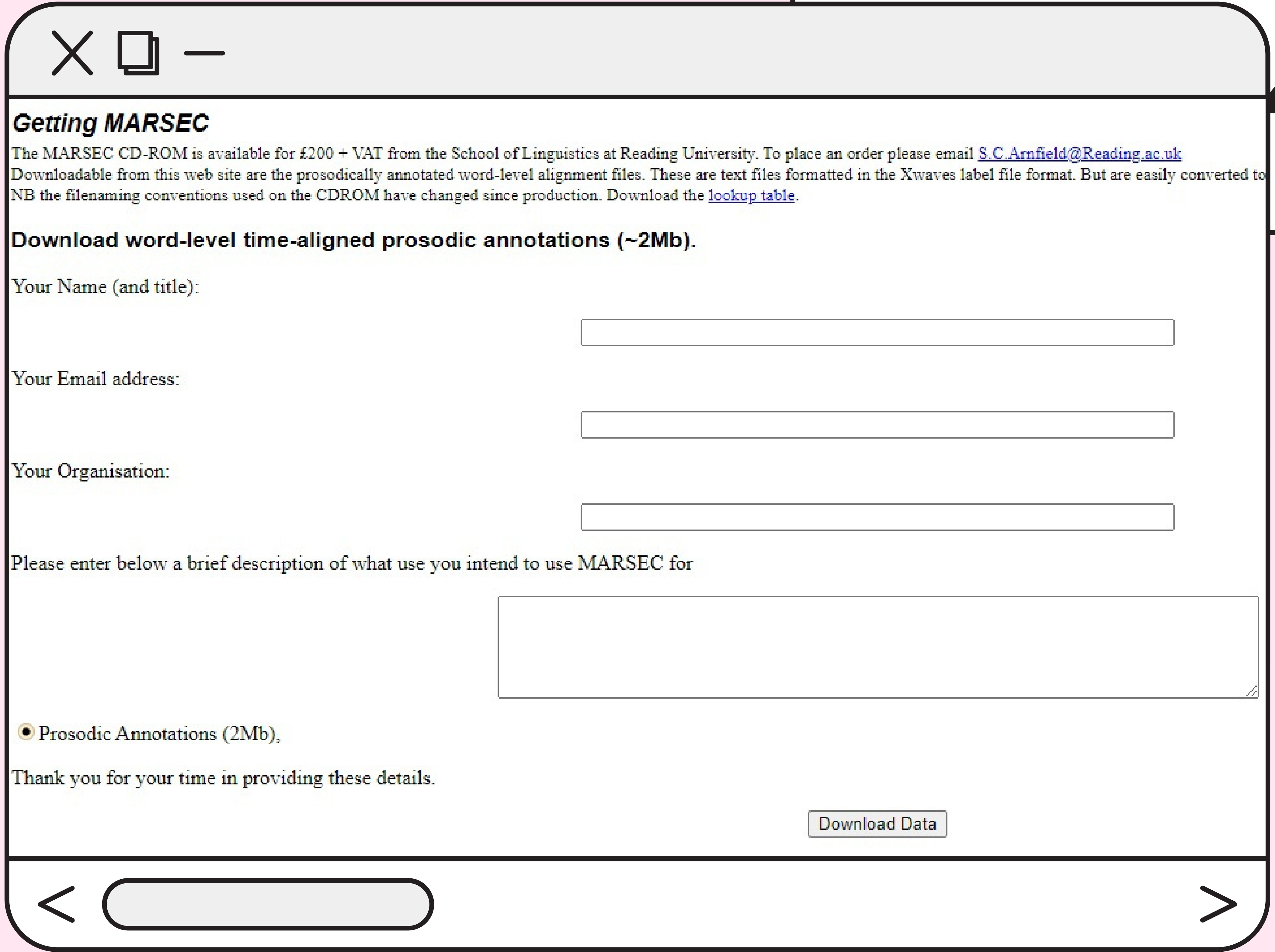
b 154 ☆(– laughs) yes☆

> A 155 – – and I ‖SAID■ 156 well I ‖don't RÈALLY ▹think■ 157 I could ‖WRÌTE■ –
– 158 and this was a sort of ‖ninety-six page ▵BÒOKLET■ 159 ‖you KNÓW■
160 about ‖that BÌG■ ☆–☆ 161 [əm] I'd I'd ‖need to GÒ through■ 162 ‖each of
the

b 163 ☆[m]☆

> A 162 processes at ▵HÒME■ ☆·☆ 164 I don't think it will be e‖nough just to have

MARSEC: THE MACHINE READABLE SPOKEN ENGLISH CORPUS



Getting MARSEC

The MARSEC CD-ROM is available for £200 + VAT from the School of Linguistics at Reading University. To place an order please email S.C.Arnfield@Reading.ac.uk
Downloadable from this web site are the prosodically annotated word-level alignment files. These are text files formatted in the Xwaves label file format. But are easily converted to
NB the filenaming conventions used on the CDROM have changed since production. Download the lookup table.

Download word-level time-aligned prosodic annotations (~2Mb).

Your Name (and title):

Your Email address:

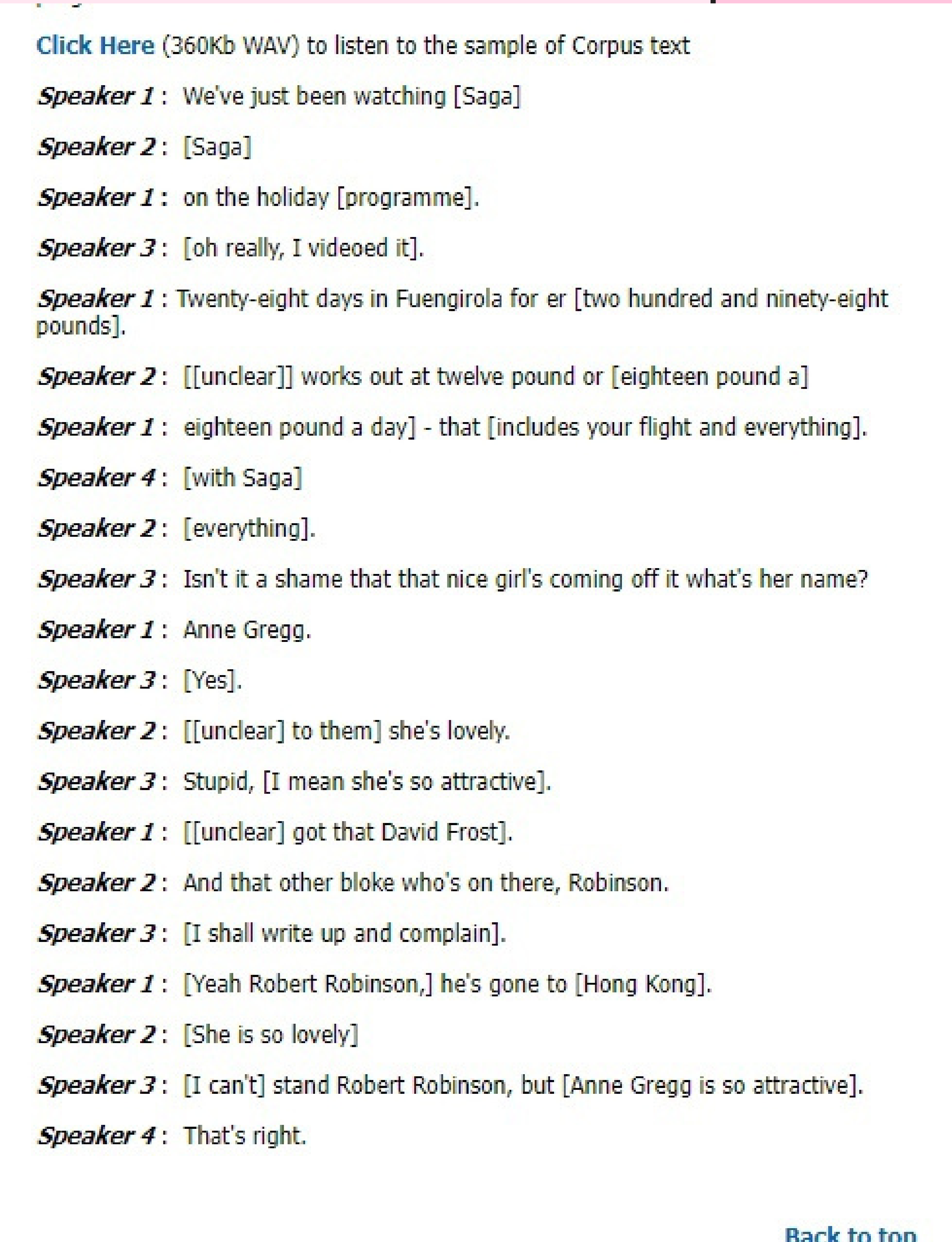
Your Organisation:

Please enter below a brief description of what use you intend to use MARSEC for

● Prosodic Annotations (2Mb),

Thank you for your time in providing these details.

THE LONGMAN BRITISH SPOKEN CORPUS



Click Here (360Kb WAV) to listen to the sample of Corpus text

Speaker 1 : We've just been watching [Saga]

Speaker 2 : [Saga]

Speaker 1 : on the holiday [programme].

Speaker 3 : [oh really, I videoed it].

Speaker 1 : Twenty-eight days in Fuengirola for er [two hundred and ninety-eight pounds].

Speaker 2 : [[unclear]] works out at twelve pound or [eighteen pound a]

Speaker 1 : eighteen pound a day] - that [includes your flight and everything].

Speaker 4 : [with Saga]

Speaker 2 : [everything].

Speaker 3 : Isn't it a shame that that nice girl's coming off it what's her name?

Speaker 1 : Anne Gregg.

Speaker 3 : [Yes].

Speaker 2 : [[unclear] to them] she's lovely.

Speaker 3 : Stupid, [I mean she's so attractive].

Speaker 1 : [[unclear] got that David Frost].

Speaker 2 : And that other bloke who's on there, Robinson.

Speaker 3 : [I shall write up and complain].

Speaker 1 : [Yeah Robert Robinson,] he's gone to [Hong Kong].

Speaker 2 : [She is so lovely]

Speaker 3 : [I can't] stand Robert Robinson, but [Anne Gregg is so attractive].

Speaker 4 : That's right.

Back to top

Примечание учителю:
Вызовите любого ученика для ответа. Дайте ему хотя бы минуту, чтобы он мог поделиться своими мнениями.

THANK YOU FOR ATTENTION!

MODERN LINGUISTIC THEORIES

SPOKEN CORPORA

Karakina Maria Fedorovna
LMAG-101