The Norwegian component of LINDSEI

Susan Nacey

ICAME 34

May 2013

Hedmark University College

Outline

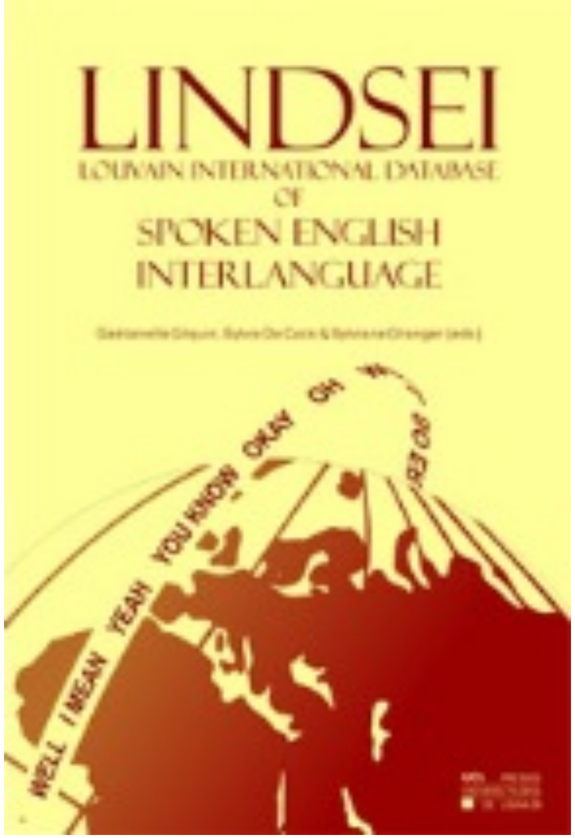
1. Introduction to LINDSEI and its Norwegian incarnation
2. Challenges in corpus compilation

The Louvain family of corpora

	Written		Spoken	
Non-native	**ICLE**		**LINDSEI**	
	v1, 2002 11 L1s 2.5m words	v2 2009 16 L1s 3.7m words	v1 2010 11 L1s 1m + words	v2 ?? 20 L1s ?? ??
Native	**LOCNESS** 324,000 words		**LOCNEC** 162,000 words	

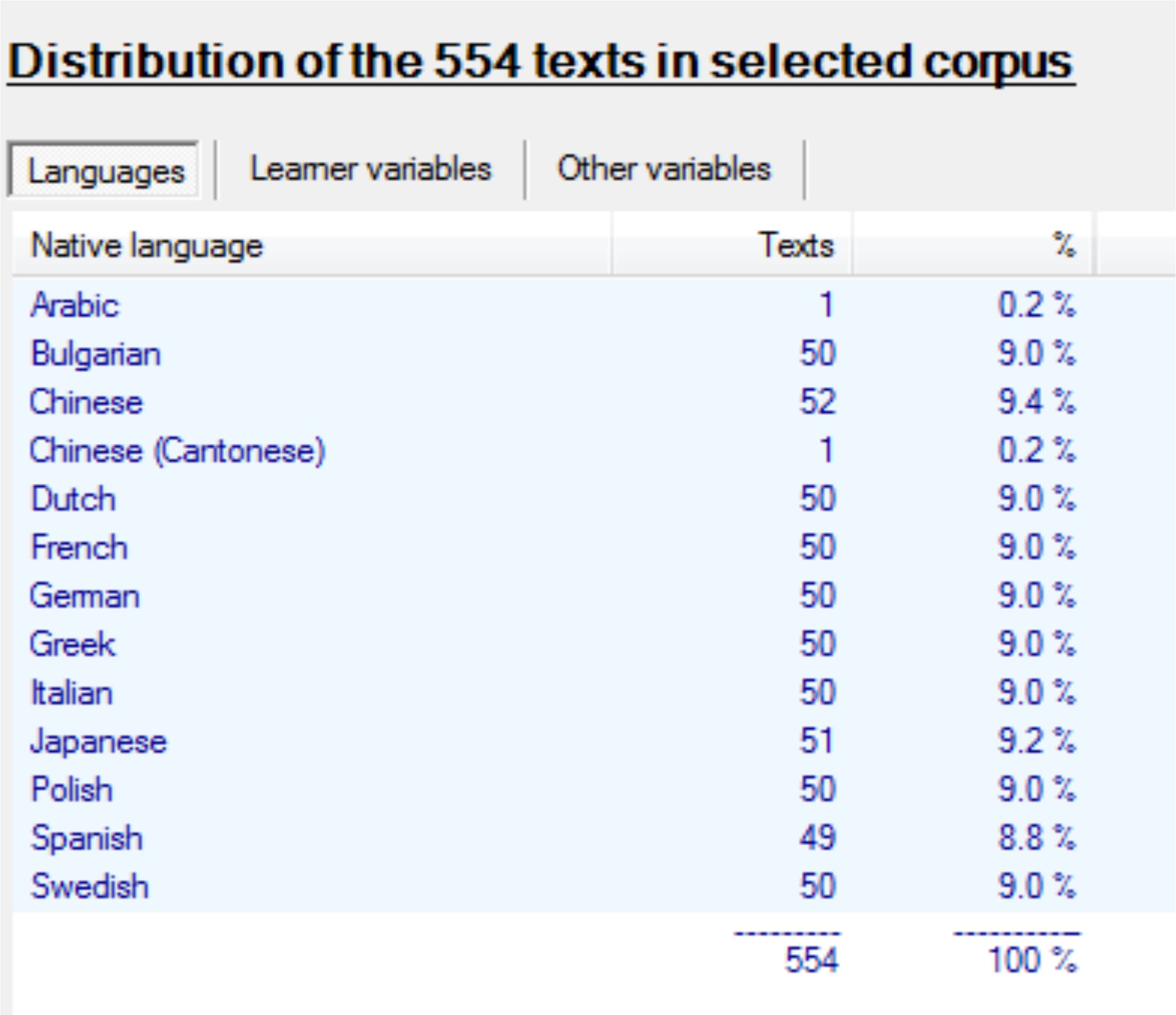
What is LINDSEI?

- 554 interviews
- 11 subcorpora
- 3 tasks
    - Set topic
    - Free discussion
    - Picture description
- 23 learner, interviewer & interview variables



Version 1

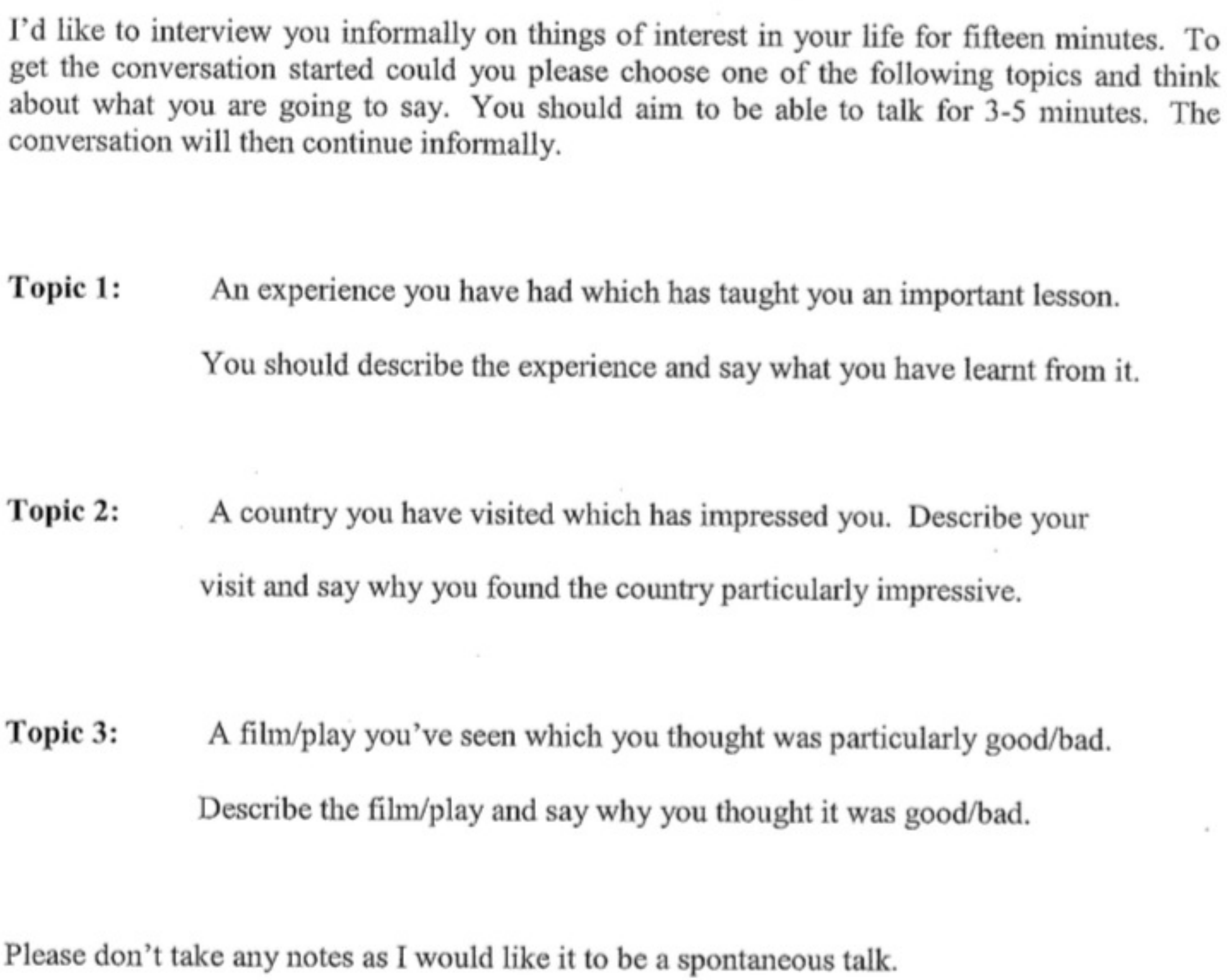
Transcription guidelines



Distribution of the 554 texts in selected corpus

Native language	Texts	%
Arabic	1	0.2 %
Bulgarian	50	9.0 %
Chinese	52	9.4 %
Chinese (Cantonese)	1	0.2 %
Dutch	50	9.0 %
French	50	9.0 %
German	50	9.0 %
Greek	50	9.0 %
Italian	50	9.0 %
Japanese	51	9.2 %
Polish	50	9.0 %
Spanish	49	8.8 %
Swedish	50	9.0 %
	---------- 554	---------- 100 %

Task 1



I'd like to interview you informally on things of interest in your life for fifteen minutes. To get the conversation started could you please choose one of the following topics and think about what you are going to say. You should aim to be able to talk for 3-5 minutes. The conversation will then continue informally.

Topic 1: An experience you have had which has taught you an important lesson.

You should describe the experience and say what you have learnt from it.

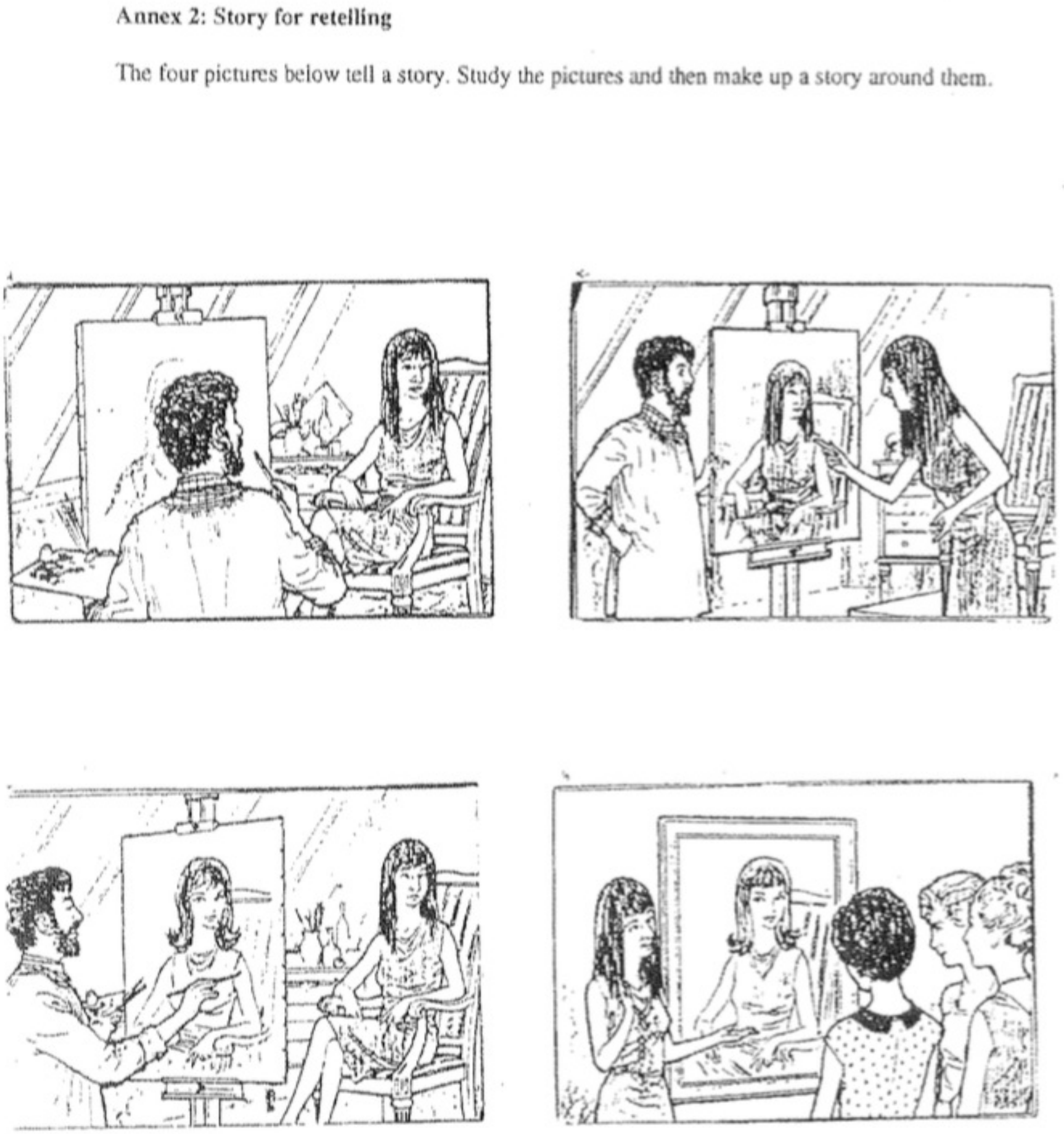
Topic 2: A country you have visited which has impressed you. Describe your visit and say why you found the country particularly impressive.

Topic 3: A film/play you've seen which you thought was particularly good/bad.

Describe the film/play and say why you thought it was good/bad.

Please don't take any notes as I would like it to be a spontaneous talk.

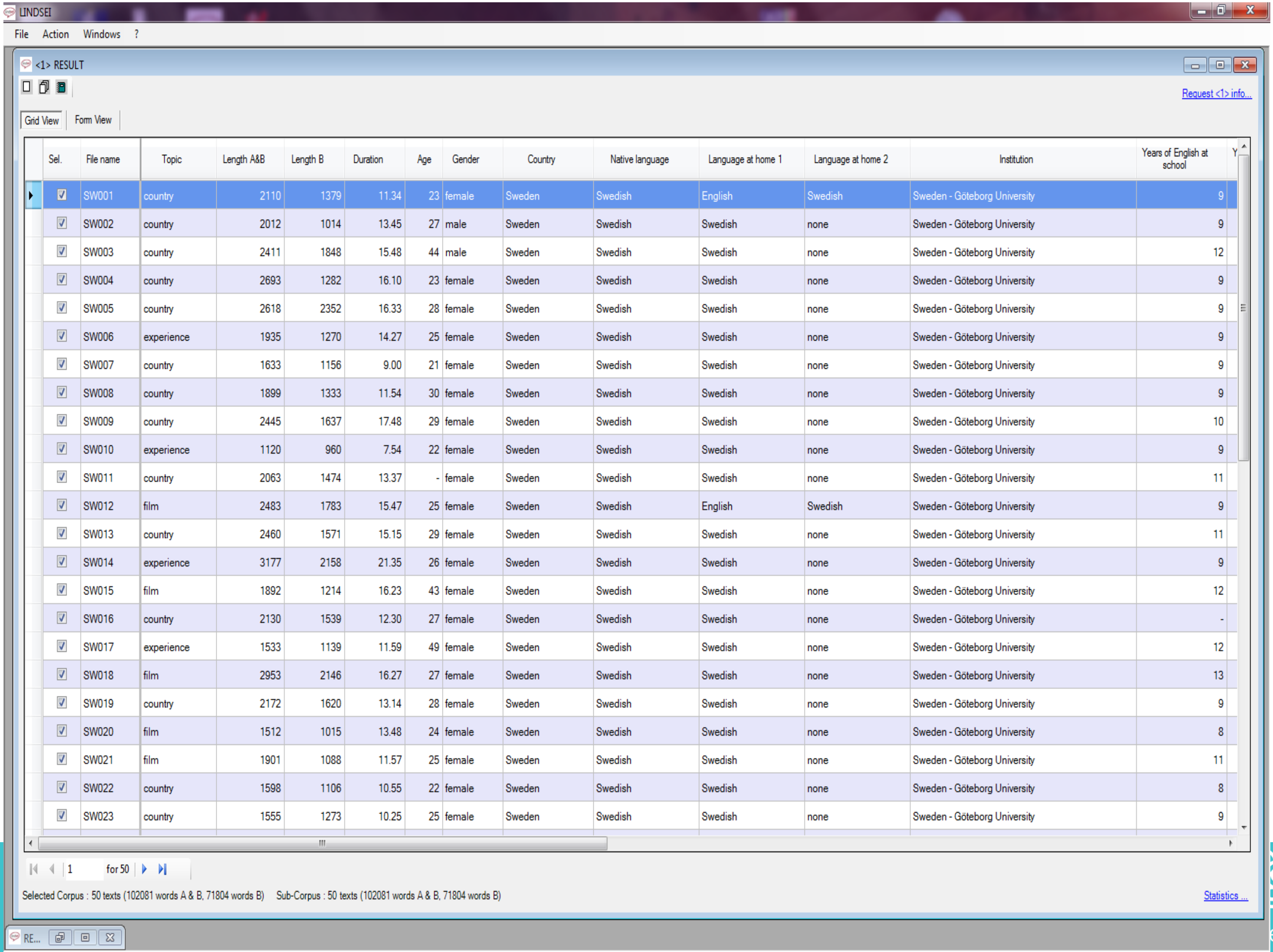
Task 3



Annex 2: Story for retelling

The four pictures below tell a story. Study the pictures and then make up a story around them.

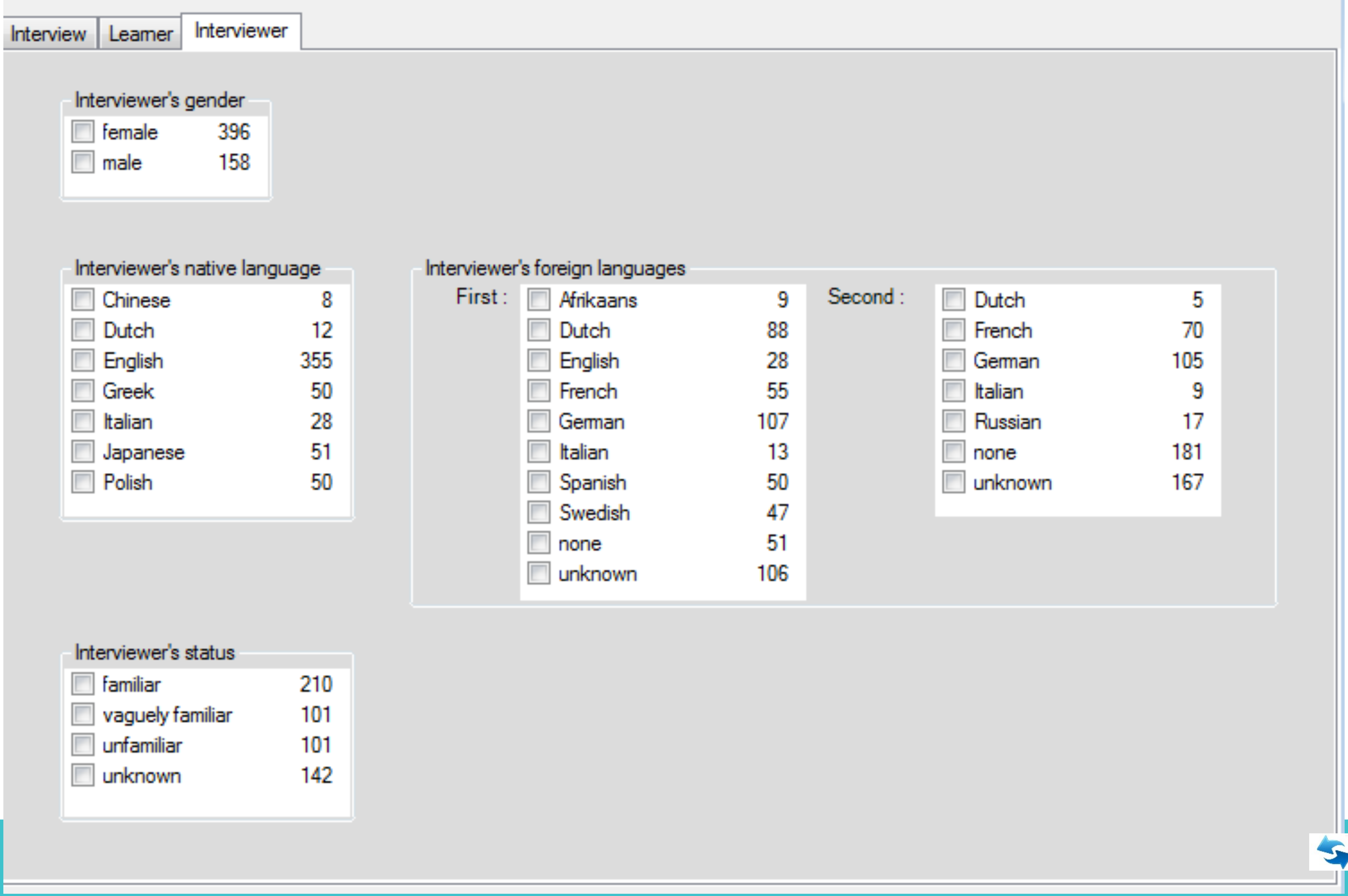
Learner and interview variables



Sel.	File name	Topic	Length A&B	Length B	Duration	Age	Gender	Country	Native language	Language at home 1	Language at home 2	Institution	Years of English at school	Y
☑	SW001	country	2110	1379	11.34	23	female	Sweden	Swedish	English	Swedish	Sweden - Göteborg University	9	
☑	SW002	country	2012	1014	13.45	27	male	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	9	
☑	SW003	country	2411	1848	15.48	44	male	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	12	
☑	SW004	country	2693	1282	16.10	23	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	9	
☑	SW005	country	2618	2352	16.33	28	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	9	
☑	SW006	experience	1935	1270	14.27	25	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	9	
☑	SW007	country	1633	1156	9.00	21	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	9	
☑	SW008	country	1899	1333	11.54	30	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	9	
☑	SW009	country	2445	1637	17.48	29	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	10	
☑	SW010	experience	1120	960	7.54	22	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	9	
☑	SW011	country	2063	1474	13.37	-	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	11	
☑	SW012	film	2483	1783	15.47	25	female	Sweden	Swedish	English	Swedish	Sweden - Göteborg University	9	
☑	SW013	country	2460	1571	15.15	29	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	11	
☑	SW014	experience	3177	2158	21.35	26	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	9	
☑	SW015	film	1892	1214	16.23	43	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	12	
☑	SW016	country	2130	1539	12.30	27	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	-	
☑	SW017	experience	1533	1139	11.59	49	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	12	
☑	SW018	film	2953	2146	16.27	27	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	13	
☑	SW019	country	2172	1620	13.14	28	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	9	
☑	SW020	film	1512	1015	13.48	24	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	8	
☑	SW021	film	1901	1088	11.57	25	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	11	
☑	SW022	country	1598	1106	10.55	22	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	8	
☑	SW023	country	1555	1273	10.25	25	female	Sweden	Swedish	Swedish	none	Sweden - Göteborg University	9	

Selected Corpus : 50 texts (102081 words A & B, 71804 words B)   Sub-Corpus : 50 texts (102081 words A & B, 71804 words B)

Interviewer variables



Interviewer's gender

□	female	396
□	male	158

Interviewer's native language

□	Chinese	8
□	Dutch	12
□	English	355
□	Greek	50
□	Italian	28
□	Japanese	51
□	Polish	50

Interviewer's foreign languages

First :

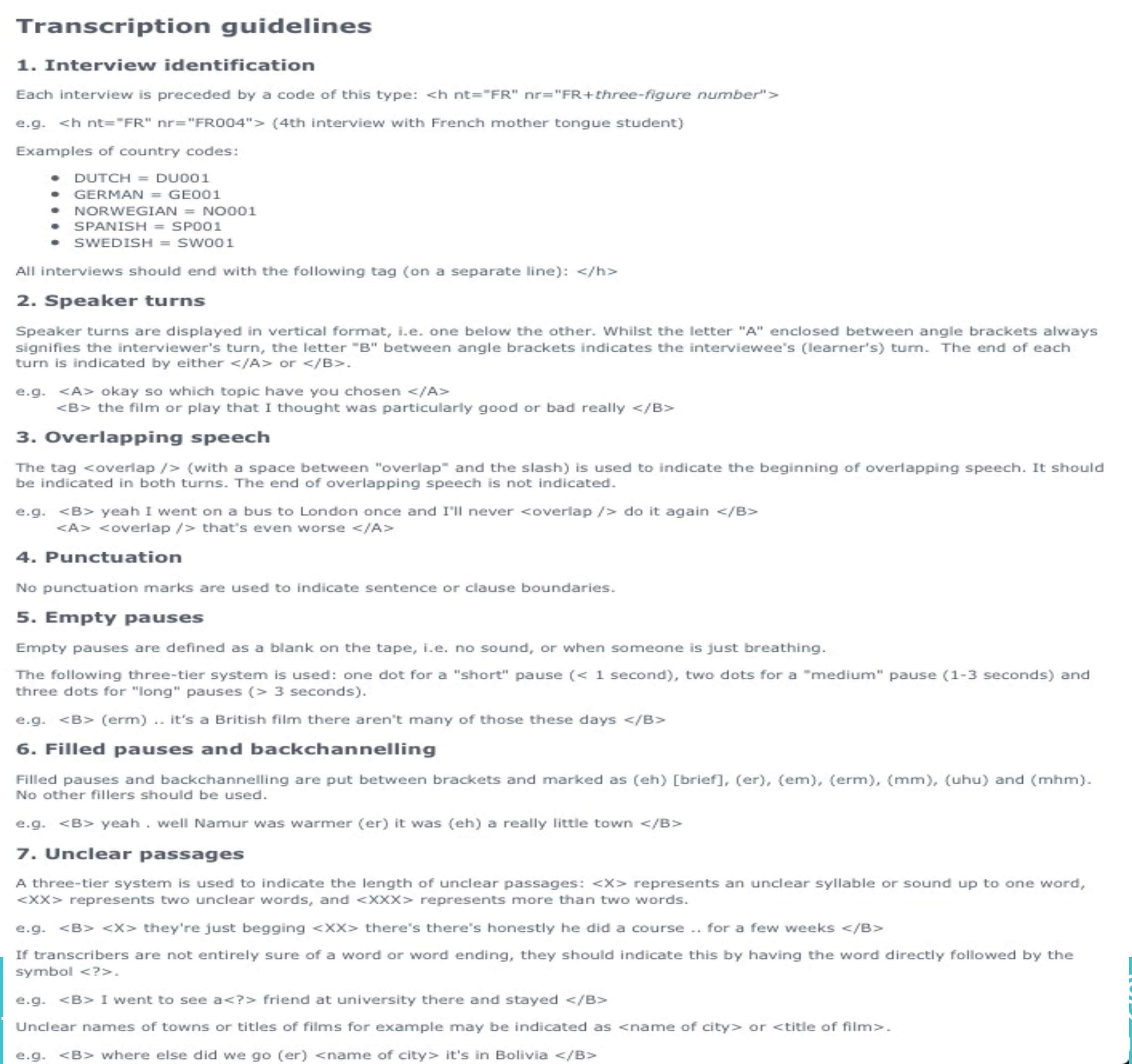
□	Afrikaans	9
□	Dutch	88
□	English	28
□	French	55
□	German	107
□	Italian	13
□	Spanish	50
□	Swedish	47
□	none	51
□	unknown	106

Second :

□	Dutch	5
□	French	70
□	German	105
□	Italian	9
□	Russian	17
□	none	181
□	unknown	167

Interviewer's status

□	familiar	210
□	vaguely familiar	101
□	unfamiliar	101
□	unknown	142



Transcription guidelines

1. Interview identification

Each interview is preceded by a code of this type: <h nt="FR" nr="FR+*three-figure number*">

e.g. <h nt="FR" nr="FR004"> (4th interview with French mother tongue student)

Examples of country codes:

- DUTCH = DU001
- GERMAN = GE001
- NORWEGIAN = NO001
- SPANISH = SP001
- SWEDISH = SW001

All interviews should end with the following tag (on a separate line): </h>

2. Speaker turns

Speaker turns are displayed in vertical format, i.e. one below the other. Whilst the letter "A" enclosed between angle brackets always signifies the interviewer's turn, the letter "B" between angle brackets indicates the interviewee's (learner's) turn. The end of each turn is indicated by either </A> or </B>.

e.g. <A> okay so which topic have you chosen </A>
<B> the film or play that I thought was particularly good or bad really </B>

3. Overlapping speech

The tag <overlap /> (with a space between "overlap" and the slash) is used to indicate the beginning of overlapping speech. It should be indicated in both turns. The end of overlapping speech is not indicated.

e.g. <B> yeah I went on a bus to London once and I'll never <overlap /> do it again </B>
<A> <overlap /> that's even worse </A>

4. Punctuation

No punctuation marks are used to indicate sentence or clause boundaries.

5. Empty pauses

Empty pauses are defined as a blank on the tape, i.e. no sound, or when someone is just breathing.

The following three-tier system is used: one dot for a "short" pause (< 1 second), two dots for a "medium" pause (1-3 seconds) and three dots for "long" pauses (> 3 seconds).

e.g. <B> (erm) .. it's a British film there aren't many of those these days </B>

6. Filled pauses and backchannelling

Filled pauses and backchannelling are put between brackets and marked as (eh) [brief], (er), (em), (erm), (mm), (uhu) and (mhm). No other fillers should be used.

e.g. <B> yeah . well Namur was warmer (er) it was (eh) a really little town </B>

7. Unclear passages

A three-tier system is used to indicate the length of unclear passages: <X> represents an unclear syllable or sound up to one word, <XX> represents two unclear words, and <XXX> represents more than two words.

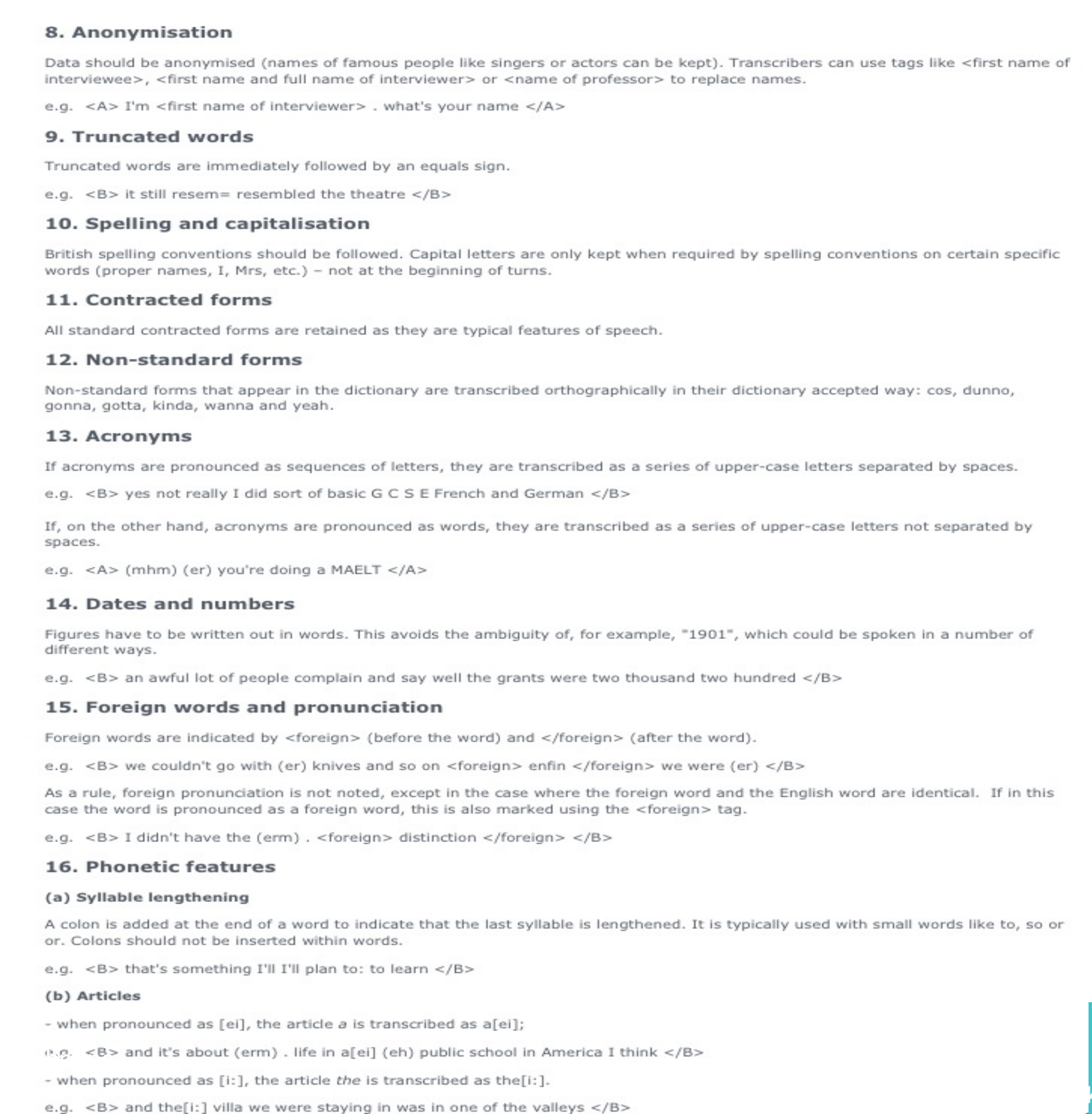
e.g. <B> <X> they're just begging <XX> there's there's honestly he did a course .. for a few weeks </B>

If transcribers are not entirely sure of a word or word ending, they should indicate this by having the word directly followed by the symbol <?>.

e.g. <B> I went to see a<?> friend at university there and stayed </B>

Unclear names of towns or titles of films for example may be indicated as <name of city> or <title of film>.

e.g. <B> where else did we go (er) <name of city> it's in Bolivia </B>



8. Anonymisation

Data should be anonymised (names of famous people like singers or actors can be kept). Transcribers can use tags like <first name of interviewee>, <first name and full name of interviewer> or <name of professor> to replace names.

e.g. <A> I'm <first name of interviewer> . what's your name </A>

9. Truncated words

Truncated words are immediately followed by an equals sign.

e.g. <B> it still resem= resembled the theatre </B>

10. Spelling and capitalisation

British spelling conventions should be followed. Capital letters are only kept when required by spelling conventions on certain specific words (proper names, I, Mrs, etc.) – not at the beginning of turns.

11. Contracted forms

All standard contracted forms are retained as they are typical features of speech.

12. Non-standard forms

Non-standard forms that appear in the dictionary are transcribed orthographically in their dictionary accepted way: cos, dunno, gonna, gotta, kinda, wanna and yeah.

13. Acronyms

If acronyms are pronounced as sequences of letters, they are transcribed as a series of upper-case letters separated by spaces.

e.g. <B> yes not really I did sort of basic G C S E French and German </B>

If, on the other hand, acronyms are pronounced as words, they are transcribed as a series of upper-case letters not separated by spaces.

e.g. <A> (mhm) (er) you're doing a MAELT </A>

14. Dates and numbers

Figures have to be written out in words. This avoids the ambiguity of, for example, "1901", which could be spoken in a number of different ways.

e.g. <B> an awful lot of people complain and say well the grants were two thousand two hundred </B>

15. Foreign words and pronunciation

Foreign words are indicated by <foreign> (before the word) and </foreign> (after the word).

e.g. <B> we couldn't go with (er) knives and so on <foreign> enfin </foreign> we were (er) </B>

As a rule, foreign pronunciation is not noted, except in the case where the foreign word and the English word are identical. If in this case the word is pronounced as a foreign word, this is also marked using the <foreign> tag.

e.g. <B> I didn't have the (erm) . <foreign> distinction </foreign> </B>

16. Phonetic features

(a) Syllable lengthening

A colon is added at the end of a word to indicate that the last syllable is lengthened. It is typically used with small words like to, so or or. Colons should not be inserted within words.

e.g. <B> that's something I'll I'll plan to: to learn </B>

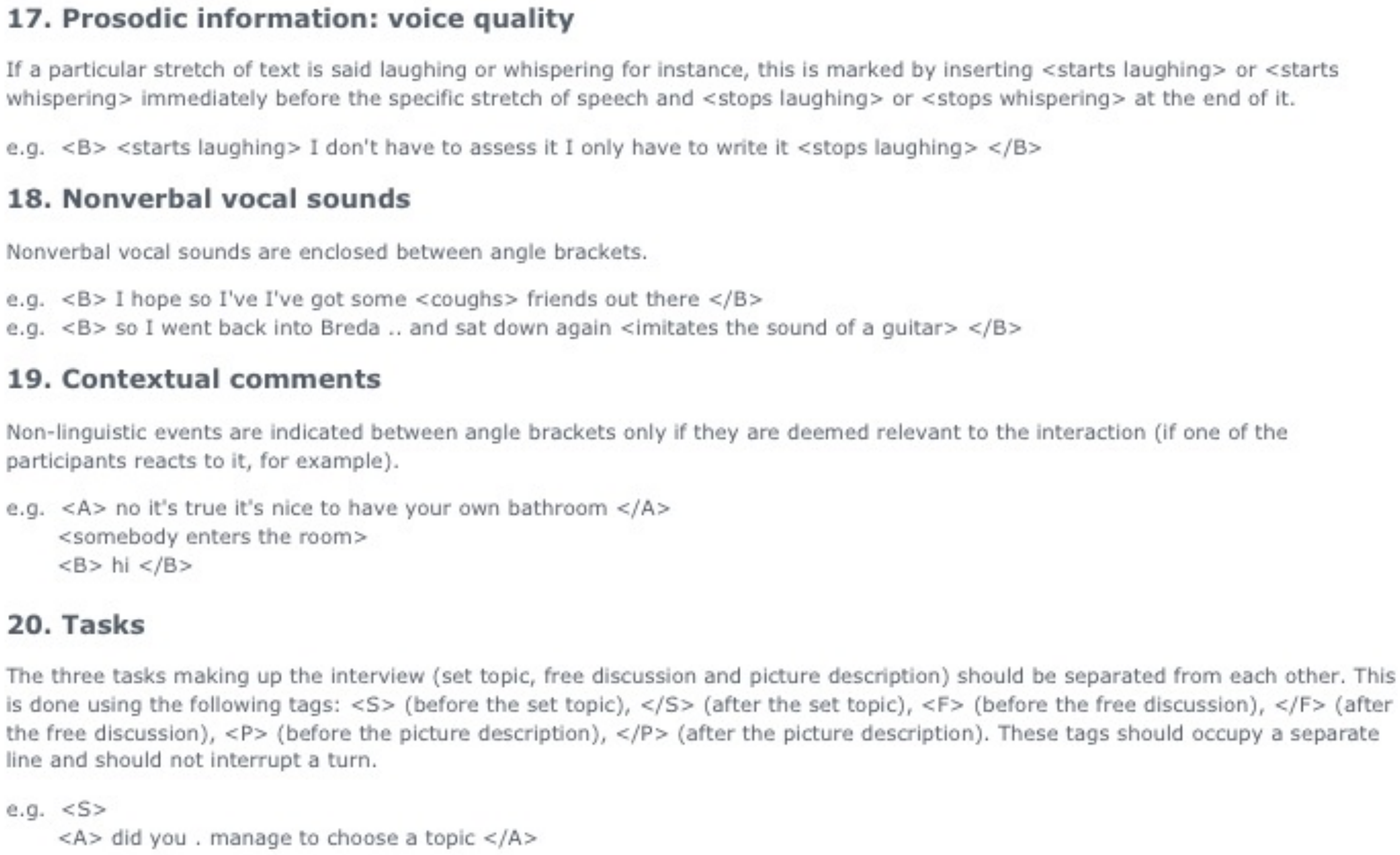
(b) Articles

- when pronounced as [ei], the article *a* is transcribed as a[ei];

e.g. <B> and it's about (erm) . life in a[ei] (eh) public school in America I think </B>

- when pronounced as [i:], the article *the* is transcribed as the[i:].

e.g. <B> and the[i:] villa we were staying in was in one of the valleys </B>



17. Prosodic information: voice quality

If a particular stretch of text is said laughing or whispering for instance, this is marked by inserting <starts laughing> or <starts whispering> immediately before the specific stretch of speech and <stops laughing> or <stops whispering> at the end of it.

e.g. <B> <starts laughing> I don't have to assess it I only have to write it <stops laughing> </B>

18. Nonverbal vocal sounds

Nonverbal vocal sounds are enclosed between angle brackets.

e.g. <B> I hope so I've I've got some <coughs> friends out there </B>
e.g. <B> so I went back into Breda .. and sat down again <imitates the sound of a guitar> </B>

19. Contextual comments

Non-linguistic events are indicated between angle brackets only if they are deemed relevant to the interaction (if one of the participants reacts to it, for example).

e.g. <A> no it's true it's nice to have your own bathroom </A>
<somebody enters the room>
<B> hi </B>

20. Tasks

The three tasks making up the interview (set topic, free discussion and picture description) should be separated from each other. This is done using the following tags: <S> (before the set topic), </S> (after the set topic), <F> (before the free discussion), </F> (after the free discussion), <P> (before the picture description), </P> (after the picture description). These tags should occupy a separate line and should not interrupt a turn.

e.g. <S>
<A> did you . manage to choose a topic </A>

Phonemic transcription?

Articles

-when pronounced as [ei], the article ‘a’ is transcribed as ‘a[ei]’;

-when pronounced as [i:] the article ‘the’ is transcribed as ‘the[i:]’.

Lost data

1) Example 1 genre
2) Example 2 southern
3) Example 3 three
4) Example 4 Viking

Lost data

NO002

<A> (em) . do you have well you have hobbies **I see** . music </A>

<B>

<A>

<B>

the p

thing

<A>

NO0

<B>

side

<A>

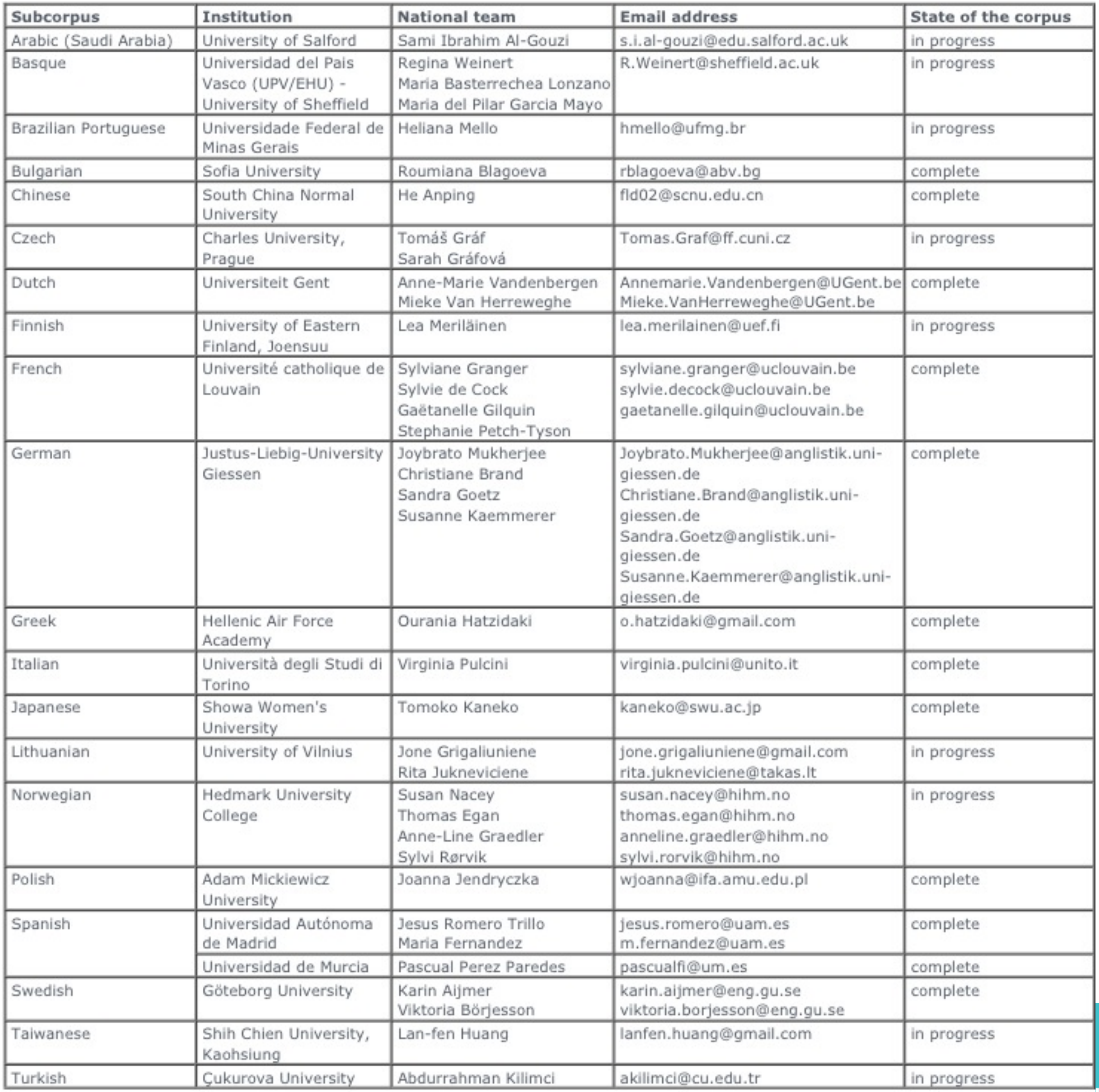
<B>

as oh kill <overlap /> **as many persons** as you can . on the way over </B>

<A> <overlap /> yeah . (mhm)</A>

LINDSEI, version 2

- Include audio files
- Link audio files to transcripts
- Additional subcorpora



Subcorpus	Institution	National team	Email address	State of the corpus
Arabic (Saudi Arabia)	University of Salford	Sami Ibrahim Al-Gouzi	s.i.al-gouzi@edu.salford.ac.uk	in progress
Basque	Universidad del Pais Vasco (UPV/EHU) - University of Sheffield	Regina Weinert Maria Basterrechea Lonzano Maria del Pilar Garcia Mayo	R.Weinert@sheffield.ac.uk	in progress
Brazilian Portuguese	Universidade Federal de Minas Gerais	Heliana Mello	hmello@ufmg.br	in progress
Bulgarian	Sofia University	Roumiana Blagoeva	rblagoeva@abv.bg	complete
Chinese	South China Normal University	He Anping	fld02@scnu.edu.cn	complete
Czech	Charles University, Prague	Tomáš Gráf Sarah Gráfová	Tomas.Graf@ff.cuni.cz	in progress
Dutch	Universiteit Gent	Anne-Marie Vandenbergen Mieke Van Herreweghe	Annemarie.Vandenbergen@UGent.be Mieke.VanHerreweghe@UGent.be	complete
Finnish	University of Eastern Finland, Joensuu	Lea Meriläinen	lea.merilainen@uef.fi	in progress
French	Université catholique de Louvain	Sylviane Granger Sylvie de Cock Gaëtanelle Gilquin Stephanie Petch-Tyson	sylviane.granger@uclouvain.be sylvie.decock@uclouvain.be gaetanelle.gilquin@uclouvain.be	complete
German	Justus-Liebig-University Giessen	Joybrato Mukherjee Christiane Brand Sandra Goetz Susanne Kaemmerer	Joybrato.Mukherjee@anglistik.uni-giessen.de Christiane.Brand@anglistik.uni-giessen.de Sandra.Goetz@anglistik.uni-giessen.de Susanne.Kaemmerer@anglistik.uni-giessen.de	complete
Greek	Hellenic Air Force Academy	Ourania Hatzidaki	o.hatzidaki@gmail.com	complete
Italian	Università degli Studi di Torino	Virginia Pulcini	virginia.pulcini@unito.it	complete
Japanese	Showa Women's University	Tomoko Kaneko	kaneko@swu.ac.jp	complete
Lithuanian	University of Vilnius	Jone Grigaliuniene Rita Jukneviciene	jone.grigaliuniene@gmail.com rita.jukneviciene@takas.lt	in progress
Norwegian	Hedmark University College	Susan Nacey Thomas Egan Anne-Line Graedler Sylvi Rørvik	susan.nacey@hihm.no thomas.egan@hihm.no anneline.graedler@hihm.no sylvi.rorvik@hihm.no	in progress
Polish	Adam Mickiewicz University	Joanna Jendryczka	wjoanna@ifa.amu.edu.pl	complete
Spanish	Universidad Autónoma de Madrid	Jesus Romero Trillo Maria Fernandez	jesus.romero@uam.es m.fernandez@uam.es	complete
	Universidad de Murcia	Pascual Perez Paredes	pascualfi@um.es	complete
Swedish	Göteborg University	Karin Aijmer Viktoria Börjesson	karin.aijmer@eng.gu.se viktoria.borjesson@eng.gu.se	complete
Taiwanese	Shih Chien University, Kaohsiung	Lan-fen Huang	lanfen.huang@gmail.com	in progress
Turkish	Çukurova University	Abdurrahman Kilimci	akilimci@cu.edu.tr	in progress

The Norwegian component of LINDSEI

- Started: Fall 2009
  - Team
  - NSD
- Recording: February 2010 - November2012
- Transcription: Spring 2013-February 2013
- Next step: Anonymization

- Funding?

The Norwegian team

Sample

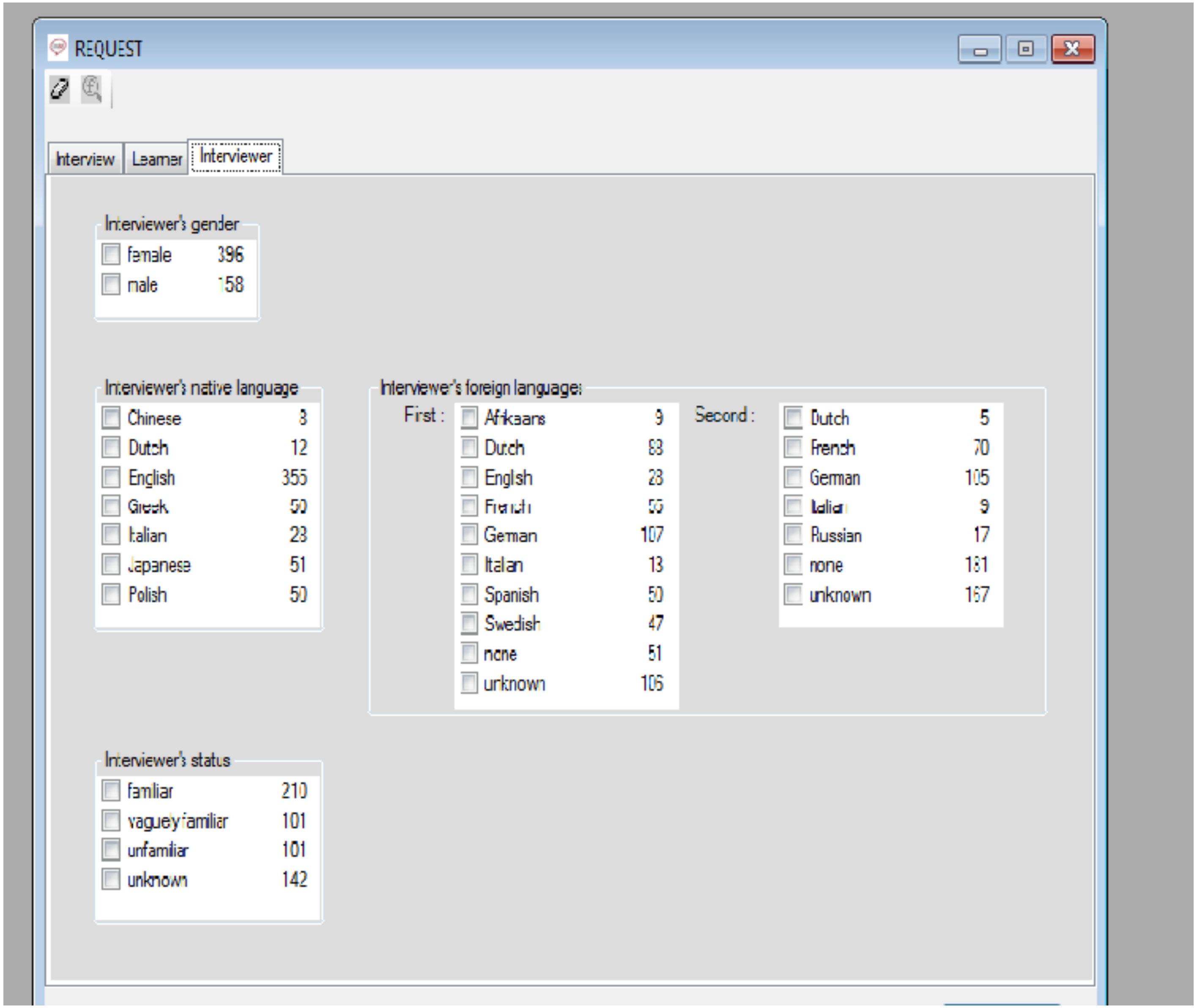
- NO005

Comparable corpora?

- Corpus compilation
- Transcription

Corpus compilation

- Interviewer
  - Identity
  - Knowledge of the learner’s L1
- Learner
  - Variety of L1



REQUEST

Interview | Learner | Interviewer

Interviewer's gender

□ female	396
□ male	158

Interviewer's native language

□ Chinese	8
□ Dutch	12
□ English	355
□ Greek	50
□ Italian	28
□ Japanese	51
□ Polish	50

Interviewer's foreign language:

First :

□ Afrikaans	9
□ Dutch	88
□ English	28
□ French	55
□ German	107
□ Italan	13
□ Spanish	50
□ Swedish	47
□ none	51
□ urknown	106

Second :

□ Dutch	5
□ French	70
□ German	105
□ Italian	9
□ Russian	17
□ none	181
□ unknown	167

Interviewer's status

□ familiar	210
□ vaguely familiar	101
□ unfamiliar	101
□ unknown	142

Variety of L1

NO007

<B> I had some teachers when I went to <foreign> ungdomsskulen </foreign> (eh)

NO026

<B> I . my .. I took vocational school </B>
<A> <overlap /> (mhm) </A>
<B> <overlap /> after . finishing (eh) <foreign> ungdomsskolen </foreign> . and

Transcription

- Empty pauses
- Filled pauses
- Non-existent words
- Foreign words and pronunciation
  - Learner's L1
  - Learner's L3/L4, etc.
  - Interviewer's pronunciation
- Syllable lengthening

Empty pauses

Empty pauses are defined as a blank on the tape, i.e. no sound, or when someone is just breathing.

The following three tier system is used: one dot for a 'short' pause (< 1 second), two dots for a 'medium' pause (1-3 seconds) and three dots for 'long' pauses (> 3 seconds).

Filled pauses

Filled pauses and backchannelling are marked as **(eh) [brief], (er), (em), (erm), (mm), (uhu) and (mhm)**. No other fillers should be used.

e.g.: <B> yeah . well Namur was warmer (er) it was (eh) a really little town </B>

- NO021
- NO040
- NO043
- NO043

- Miscellaneous sounds

NO031

NO002

NO036

Non-existent words

- teached NO048
- equipments NO037
- NO021

<B> yeah there's a particular scene in there when . I think it's his bachelor party or hers or something when there's .. probably hers because there's a lot of guys walking around with swim feets on </B>

- a coincident NO047

• interpreted NO006	interpretated
• renovating NO037	reinvenovating
• tactics NO036	tacticals

Foreign pronunciation

As a rule, foreign pronunciation is not noted, except in the case where the foreign word and the English word are identical. If in this case the word is pronounced as a foreign word, this is also marked using the <foreign> tag.

- Oslo
  - NO045
- Lillehammer
  - NO033
  - NO027
  - NO050
- Hamar
  - NO016
  - NO040
  - NO048
- Etc.

Trondheim
NO036

Valdres
NO026

Gjøvik
NO041

Foreign words and pronunciation

L3

- NO001
- NO041

Foreign terms

- Learner
- NO004
- NO012

- Interviewer

Syllable lengthening

- this big husky security guard
  NO049

Phonetic features
(a) Syllable lengthening

A colon is added at the end of a word to indicate that the last syllable is lengthened. It is typically used with small words like *to*, *so* or *or*. Colons should not be inserted within words.