LANCASTER UNIVERSITY

Practical workshop

Compiling and analysing a spoken academic corpus

CL 2013, Lancaster 22nd July

Dr. Vaclav Brezina

www.lknol.com

Contents

1 Compiling and analysing a spoken academic corpus

Vaclav Brezina
CASS, Lancaster University
Email: v.brezina@lancaster.ac.uk
Web: www.lknol.com

Spoken academic English has been generally recognised as a specific genre/register with its own social dynamic and unique linguistic features (Biber 2006; Swales 2006). Swales (2006) characterises academic speech as a genre, which is much closer to informal conversation than written academic prose, a genre which contains "a considerable amount of technical lexis embedded into a loosely co-ordinated sentence structure and surrounded by heavy employment of deictic elements" (p. 23). The best way to investigate the typical characteristics of academic speech is through analysis of a corpus.

Building a corpus of academic speech, however, presents a major challenge (Biber et al. 2001; Crowdy 1993, 1994, 1995; Simpson-Vlach and Leicher 2006). Unlike samples of academic writing which are available in copious quantities online, academic speech needs to be painstakingly recorded as well as carefully transcribed and annotated.

This workshop draws on the author's experience with building a corpus of academic speech in the New Zealand context. The project has been carried out at the department of Applied Language Studies and Linguistics, University of Auckland (UoA). The result of the project is a small (160,000 tokens) single-genre corpus of spoken university English – *Academic discourse verbal interactions corpus* (*ADVICe*). *ADVICe,* which is now available to the linguistic community, consists of transcribed and morphologically annotated recordings of advisory sessions (office hours, supervisory meetings etc.) between students and lecturers at UoA.

The workshop will focus on the main issues of spoken corpus design and analysis and compare *ADVICe* with other available corpora of academic speech (*MICASE, BASE*). The workshop will be of particular interest to researchers and practitioners interested in spoken academic discourse in general as well as in particular lexico-grammatical patterns in academic speech.

Programme

9.30 – 10.00 Introduction: Academic speech and *ADVICe*
10.00 – 11.00 Corpus design: participant selection, recoding techniques, transcription
11.00 – 11.30 Tea & coffee break
11.30 – 1pm Corpus analysis (KWIC, CONTEXT, visualisations, statistics)

2 ADVICe: an overview

1. What do you think are the most important qualities of a spoken language corpus?
2. How large do you think a spoken corpus should be?

3. Can you figure out what the speakers are talking about in the transcript below?

(1) **S:** not really, no. i mean it's interesting 'cause i i i did this Indian girl yesterday, [Tx: hm] and she had almost no response, i mean it was pretty much what i would hope for, the tropical response.
T: wow.
S: it went down and it just stayed down, [Tx: yeah] you know, it didn't come up. erm but she was quite comfortable.
T: yeah, she wasn't in any pain.
S: and i had a girl today erm and she came out and she was [.] Dutch and English ancestry, and she had a really good response, you know, [Tx: yeah] went down to sort of six and then up to about ten and stayed around eight degrees so [.] and yet she was more much more uncomfortable the whole time [Tx: er] so the the pain that they feel they're feeling, doesn't seem to have a lot of relationship to how the the body's handling the cold [Tx: yeah] which is probably what you'd expect.

 For more information on the issue of context and corpus linguistics see the Widdowson-Stubbs debate (Stubbs 2001; Widdowson 1998, 2000). Cf. also Fetzer (2004) and Swales (2002, 2006).

2.1 ADVICe – basic information

Table 1 *ADVICe*

Size (tokens)	No. of speakers	Genre/speech event	Discourse mode	Variety of English	Period
approx. 160,000	49	advisory sessions	highly interactive	NZ & international	2008-2009

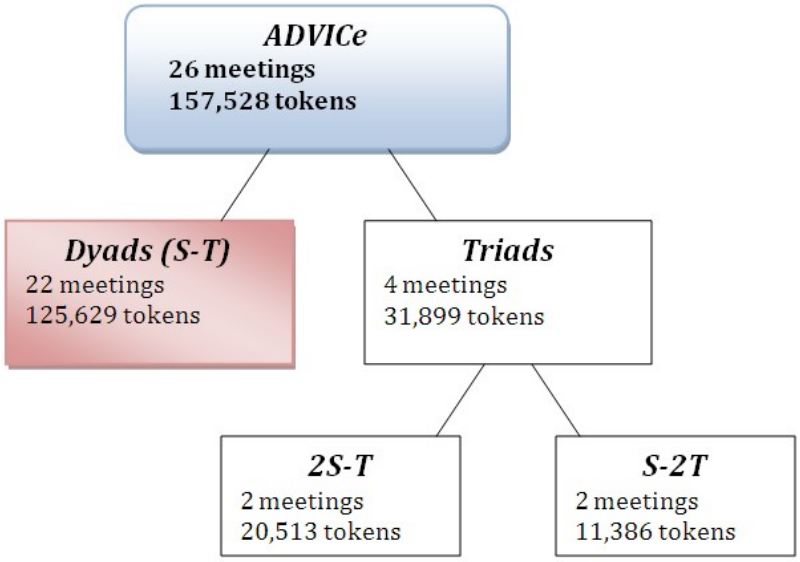

Figure 1 *ADVICe* corpus structure

2.2 NZ Academic speech

(2) **S:** hm [.] i think [.] i **definitely** refined like i **guess** my way of planning, **and** […] i mean i know i do, quite
in-depth planning, but **I guess** [Tx: **yes**] just like [the **set out**, and everything <LAUGH> **yes**, yeah]
T: [you are are (xx) aren't you, **yes yes**]
S: but I felt like my **plans** are good and that i could run with **them** and everything [Tx: **yes**] you know, that I had a good understanding of what I needed to **teach**, [Tx: yes] because of the way that I **planned**, erm […] so **yeah**, that was **definitely** good, erm […]

Brezina, V. 2012. *ADVICe* - corpus. Available at www.lknol.com/advice.html.

(3) T: and so we gotta [.] aim really **flat tacks** to get everything
S: i'm sorry aim really?
T: getting aim f- go **flat tacks** to get everything finished
S: ok [.] [Tx: yeah] flat?
T: **flat tacks** [.] it's an English New Zealand expression
S: **flat tax**?
T: yes
S: like income **tax**?
T: no **t a c k** [.][Sx: oh] **tacks**
S: or like sailing?
T: mhm?
S: like a **tack** in sailing?
T: no [**tac-**]
S: [i'm just] i love the language
T: it's [.] i don't know where the word comes from **flat tacks** is [Sx: ok] think about the te- the **tacks** [.] flattening the **tacks** (xx) **carpet tacks** down
S: oh, all the way [Tx: yeah] all right [Tx: yeah] **flat tacks** [Tx: that's right that's right] got it, all right

Brezina, V. 2012. *ADVICe* - corpus. Available at www.lknol.com/advice.html.

2.3 ADVICe and other corpora of academic speech

Table 2 Advice and other Academic speech corpora

Corpus	Region	NS and NNS speech	Speech events	Size	Comparable with ADVICe (Dyads)	General availability
MICASE	US	88% NS 12% NNS	lectures, seminars, labs, office hours, meetings, study groups etc.	1.7 M	85 k (70k)	YES
T2K-SWAL (spoken subcorpus)	US	?	class sessions, labs, study groups, office hours, service encounters	1.7 M	50 k (?)	NO
BASE	UK	Mainly NS (?)	lectures & seminars	1.6 M	0	YES
HKCSE (Academic subcorpus)	HK (NNS – Hong Kong Chinese & NS – US, UK, AUS)	15 % NS 85% NNS	lectures, seminars, workshops, supervision	200k	0	YES
ADVICe	NZ	78% NS 22% NNS	advisory sessions	160 k	160k (125k)	YES

3 Data gathering

4 Transcription

4. Use the space below to transcribe **recording 4.**

.....

.....

.....

.....

.....

5. How would you transcribe hesitation sounds?
6. How would you indicate a long pause in speech?
7. How would you treat incomprehensible words in the recording?
8. Look at Table 3 which compares transcription conventions in four different corpora. Discuss the advantages & drawbacks of each solution.
9. How long do you think it takes to transcribe an hour of speech?

Table 3 Spoken language corpora: Transcription conventions

	BNC	*MICASE*	*BASE*	*ADVICe*
pauses	significant pauses timed (silence longer than was judged normal for the speaker or speakers)	, . (1-2 s.) ... (2-3s.) timed (4+s.)	[0.4] exact value	[.] very short [...] short timed (4+s.)
hesitation sound	er, erm	um	er	er, erm
backchannels	mm, mhm	hm, hm', huh, mm, mhm, uh, mkay	?	hm, mhm
reduced *because*	cos	cuz	'cause	'cause
utterance beginning	Upper case	Lower case	Lower case	Lower case
personal pronoun I	I	i	i	i
reduced forms	gonna, wanna, kinda	gonna, wanna, kinda	going to, want to, kind of	gonna, wanna, kind of

4.1 ADVICe - Guidelines for linguistic transcription

1. Transcribe all words/sounds in the recording including fillers (you know, erm, hm etc.), laughter <LAUGH> , etc.
2. Follow the **TRANSCRIPTION AND MARK-UP CONVENTIONS** (see below).
3. Produce a rough transcript first and then attend to details.
4. If the recording includes a new term/word you are not sure about, try to google it.
5. If you are not sure how to transcribe something, highlight that word/passage and ask about it (you can always email me at v.brezina@auckland.ac.nz)
6. Be consistent.

A. TRANSCRIPTION AND MARK-UP CONVENTIONS	
MEANING/DESCRIPTION	**TRANSCRIPTION SYMBOLS**
Speaker: Student (S:), Teacher (T:)	**S:** at the beginning of each turn or interruption/backchannel.
Two or more speakers, in unison (used mostly for laughter)	**ST**
PAUSES	
Comma indicates a brief (1-2 second) pause after a clause.	**,**
Period indicates a brief pause accompanied by an utterance final (falling) intonation contour; not used in a syntactic sense to indicate complete sentences.	**.**
Question mark indicates intonation contour typical of a question.	**?**
short pause (up to 1 second)	[.]
longer pause	[...] up to 3 seconds [5s], [15s] measured pause (approximate time)
BACKCHANNEL CUES and FAILED INTERRUPTIONS	
BACKCHANNEL	T: the main utterance [Sx: Backchannel] the main utterance continues S: the main utterance [Tx: Backchannel] the main utterance continues
OVERLAP	
	S: [simultaneous utterance] T: [simultaneous utterance] the utterance continues
LAUGHTER	
All laughter is marked. Speaker ID not marked if current speaker laughs.	<LAUGH>, <S LAUGH> <ST LAUGH>, etc.
CONTEXTUAL EVENTS	
Various contextual (non-speech) events are noted, usually only when they affect comprehension of the surrounding discourse.	<WRITING ON BOARD>
READING PASSAGES	
Used when part of an utterance is read verbatim.	<READING>.....</READING>
UNCERTAIN or UNINTELLIGIBLE SPEECH	
Two x's in parentheses indicate one or more words that are completely unintelligible. Words surrounded by parentheses indicate the transcription is uncertain.	(xx)
NAMES-ANONYMISING	
When participants' names occur in a recording, they are changed to <NAME> in the transcript. Names of non-present people referred to in the recording are also changed, unless these people are public persons (authors of books, famous scientists, politicians).	

B. SPELLING CONVENTIONS		
GENERAL	Standard orthography is used for most words, even though they may not be fully	

	pronounced, may be pronounced with a foreign accent, etc.	
CAPITALIZATION	Only proper nouns (names, departments, course titles, organizations, etc.) are capitalized (in addition to acronyms; see below). Neither the beginnings of turns nor the pronoun 'i' are capitalized.	
FILLED PAUSES, BACKCHANNEL CUES, EXCLAMATIONS, etc.	All hesitation and filler words, backchannel cues, and transcribable exclamations are spelled out, as shown on the right.	Hesitation/Filler Words/Backchannels: er, erm, hm, mhm, ok
		Yes/No Responses: yes: yeah, yep, ok, hm, mhm, aha, nope
		Exclamations/Doubt/Misc.: ah, oh, ooh, oops
		End of sentence tag: eh
CONTRACTIONS and LEXICALIZED REDUCED FORMS	All standard contractions of is, am, are, had, have,	i'd, i've, i'm, i'll, she's, she'll, he's, they've, etc.
NON STANDARD FORMS		could of been
ACRONYMS, ABBREVIATIONS, LETTERS AS VARIABLES	Acronyms are written in all caps. Three commonly abbreviated titles are left as abbreviations, but without periods. An acronym pronounced as a word is run together as one word. When an acronym is spelled out, it appears in all caps with hyphens between each letter (except PhD).	
NUMBERS	All numbers are fully spelled out as words. Standard hyphenation rules apply, with some additional guidelines: page numbers, course numbers, and room numbers are all hyphenated.	nineteen ten nineteen twenty-nine page one-fifty-seven Poli Sci one-sixty room thirty-twelve
REPETITIONS and REPAIRS	All repetitions of a word, partial word or phrase are transcribed.	it's it's a kind of
	Truncated or cut-off words have a hyphen at the end of the last audible sound/letter.	it's a transcrip-

Adapted from: *MICASE*: Transcription and Spelling Conventions
http://micase.elicorpora.info/micase-statistics-and-transcription-conventions/micase-transcription-and-mark-up-convent


Go to workshop webpage and work with the recordings provided.
http://www.lknol.com/Workshop.html

10. Transcribe RECORDING 5.

.....

.....

.....

.....

.....

.....

.....

.....

.....

11. Transcribe RECORDING 6.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

12. Compare your transcripts with those on p. 16.

5 Tagging

5.1 Before tagging…

- MS Word text to PLAIN text
- MS Word Macro – search & replace:

“” → ""
‘ → '
double space →single space

- CLAWS & capitalisation and punctuation.

hm i think [.] i definitely refined like i guess my way of planning, and […] i mean i know i do, quite in-depth planning

Pauses, *I* lower case

hm_ITJ [_PUL ._PUN]_SENT -----_PUN i_ZZ0 think_VVB [_PUL ._PUN]_SENT -----_PUN i_ZZ0 definitely_AV0 refined_VVN like_PRP i_ZZ0 guess_VVB my_DPS way_NN1 of_PRF planning_NN1 ,_PUN and_CJC [_PUL]_PUR i_ZZ0 mean_VVB i_ZZ0 know_VVB i_ZZ0 do_VDB ,_PUN quite_AV0 in-depth_AJ0 planning_SENT

No pauses, *I* lower case

hm_ITJ i_ZZ0 think_VVB i_ZZ0 definitely_AV0 refined_VVN like_PRP i_ZZ0 guess_VVB my_DPS way_NN1 of_PRF planning_NN1 ,_PUN and_CJC i_ZZ0 mean_VVB i_ZZ0 know_VVB i_ZZ0 do_VDB ,_PUN quite_AV0 in-depth_AJ0 planning_SENT

No pauses, *I* capitalised

hm_ITJ I_PNP think_VVB I_PNP definitely_AV0 refined_AJ0 like_CJS I_PNP guess_VVB my_DPS way_NN1 of_PRF planning_NN1 ,_PUN and_CJC I_PNP mean_VVB I_PNP know_VVB I_PNP do_VDB ,_PUN quite_AV0 in-depth_AJ0 planning_SENT

5.2 Tagging in practice

13. Look at the CLAWS5 Tagset below. Which of the word classes do you think are the most difficult for the automating tagger to assign correctly?

14. Look at the short *ADVICe* transcript below (transcript 7). Try to assign word-class membership to each of the lexical units.

(7) **T:** sorry it started as a quick one and i thought he was talking [Sx: ok] to me and it would be a bit longer. right.

S: ok so i'm kind of half way through my stuff. i'm still transcribing.

15. Were there any difficult/ambiguous cases in task 14? Discuss.

UCREL CLAWS5 Tagset

adjectives	**AJ0**	adjective (unmarked) (e.g. GOOD, OLD)
	AJC	comparative adjective (e.g. BETTER, OLDER)
	AJS	superlative adjective (e.g. BEST, OLDEST)
articles	**AT0**	article (e.g. THE, A, AN)
adverbs	**AV0**	adverb (unmarked) (e.g. OFTEN, WELL, LONGER, FURTHEST)
	AVP	adverb particle (e.g. UP, OFF, OUT)
	AVQ	wh-adverb (e.g. WHEN, HOW, WHY)
conjunctions	**CJC**	coordinating conjunction (e.g. AND, OR)
	CJS	subordinating conjunction (e.g. ALTHOUGH, WHEN)
	CJT	the conjunction THAT
numeral (card.)	**CRD**	cardinal numeral (e.g. 3, FIFTY-FIVE, 6609) (excl ONE)
determiners	**DPS**	possessive determiner form (e.g. YOUR, THEIR)
	DT0	general determiner (e.g. THESE, SOME)
	DTQ	wh-determiner (e.g. WHOSE, WHICH)
there	**EX0**	existential THERE
interjections	**ITJ**	interjection or other isolate (e.g. OH, YES, MHM)
nouns	**NN0**	noun (neutral for number) (e.g. AIRCRAFT, DATA)
	NN1	singular noun (e.g. PENCIL, GOOSE)
	NN2	plural noun (e.g. PENCILS, GEESE)
	NP0	proper noun (e.g. LONDON, MICHAEL, MARS)
numeral (ord.)	**ORD**	ordinal (e.g. SIXTH, 77TH, LAST)
pronouns	**PNI**	indefinite pronoun (e.g. NONE, EVERYTHING)
	PNP	personal pronoun (e.g. YOU, THEM, OURS)
	PNQ	wh-pronoun (e.g. WHO, WHOEVER)
	PNX	reflexive pronoun (e.g. ITSELF, OURSELVES)
possessive 's	**POS**	the possessive (or genitive morpheme) 'S or '
prepositions	**PRF**	the preposition OF
	PRP	preposition (except for OF) (e.g. FOR, ABOVE, TO)
punctuation	**PUL**	punctuation - left bracket (i.e. (or [ )
	PUN	punctuation - general mark (i.e. . ! , : ; - ? ...)
	PUQ	punctuation - quotation mark (i.e. ` ' ")
	PUR	punctuation - right bracket (i.e.) or])
to-inf	**TO0**	infinitive marker TO
verbs	**VBB**	the "base forms" of the verb "BE" (except the infinitive), i.e. AM, ARE
	VBD	past form of the verb "BE", i.e. WAS, WERE
	VBG	-ing form of the verb "BE", i.e. BEING
	VBI	infinitive of the verb "BE"
	VBN	past participle of the verb "BE", i.e. BEEN
	VBZ	-s form of the verb "BE", i.e. IS, 'S
	VDB	base form of the verb "DO" (except the infinitive), i.e.
	VDD	past form of the verb "DO", i.e. DID

	VDG	-ing form of the verb "DO", i.e. DOING
	VDI	infinitive of the verb "DO"
	VDN	past participle of the verb "DO", i.e. DONE
	VDZ	-s form of the verb "DO", i.e. DOES
	VHB	base form of the verb "HAVE" (except the infinitive), i.e. HAVE
	VHD	past tense form of the verb "HAVE", i.e. HAD, 'D
	VHG	-ing form of the verb "HAVE", i.e. HAVING
	VHI	infinitive of the verb "HAVE"
	VHN	past participle of the verb "HAVE", i.e. HAD
	VHZ	-s form of the verb "HAVE", i.e. HAS, 'S
	VM0	modal auxiliary verb (e.g. CAN, COULD, WILL, 'LL)
	VVB	base form of lexical verb (except the infinitive)(e.g. TAKE, LIVE)
	VVD	past tense form of lexical verb (e.g. TOOK, LIVED)
	VVG	-ing form of lexical verb (e.g. TAKING, LIVING)
	VVI	infinitive of lexical verb
	VVN	past participle form of lex. verb (e.g. TAKEN, LIVED)
	VVZ	-s form of lexical verb (e.g. TAKES, LIVES)
negatives	**XX0**	the negative NOT or N'T
letters	**ZZ0**	alphabetical symbol (e.g. A, B, c, d)

16. Look at the excerpt below which was automatically assigned CLAWS5 tags. Do you think this has been done successfully? Compare with your answer from task 14.

T: sorry_AJ0 it_PNP started_VVD as_PRP a_AT0 quick_AJ0 one_PNI and_CJC I_PNP thought_VVD he_PNP was_VBD talking_VVG [Sx: ok_AV0] to_PRP me_PNP and_CJC it_PNP would_VM0 be_VBI a_AV0 bit_AV0 longer_AV0 ._PUN right_AV0 ._PUN

S: ok_AV0 so_CJS I_PNP 'm_VBB kind_AV0 of_AV0 half_AV0 way_AV0 through_PRP my_DPS stuff_NN1 ._PUN I_PNP 'm_VBB still_AV0 transcribing_VVG.

For more information about the accuracy of CLAWS tagging see Burnard (2007); Vine (2011).

6 Corpus analysis

6.1 Basic analysis

Use *ADVICe* online Basic search (http://www.lknol.com/ADVICE/Corpsearch.html**) to answer the following questions:**

17. What is the most common verb in academic speech? Test your hypotheses.
18. Can you find examples of taboo words in academic speech? Discuss.
19. Which of the following phrases is more frequent in academic speech: *I think* or *I don't think.* Look at the examples and discuss the reason for higher frequency of occurrence of one of the forms.

Use *ADVICe* Comparison tool (http://www.lknol.com/ADVICE/Compare.html**) to answer the following question:**

20. Compare academic speech of students with academic speech of lecturers:

*Look at the use of personal pronouns.
*Look at the use of directives such as *you should, you ought to, you must, you'd better.*

6.2 Visualisations

Use MANY EYES visualisation tool to:

21. Create a word cloud of the most frequent words in *ADVICe.*
22. Create a word cloud of the most frequent epistemic markers in *ADVICe.*
23. Create and explore a word tree based on *ADVICe.*

7 Further activities & discussion

Choose ONE of the three questions (24, 25 or 26) below. Use *ADVICe* web tools and/or MANY EYES visualisations to answer the question.

24. Table 4 below shows a comparison between *ADVICe* and *BNC*-demographic. What are the typical words in academic speech? (They are marked + in the "overused in ADVICe "column.) Use examples from *ADVICe* to create a teaching material that would help students with some of the frequent words they may encounter in spoken academic discourse.

25. Table 2 shows 20 most frequent tri-grams in academic speech. Choose three or four interesting trigrams from the list. Use *ADVICe* basic search and the word tree tool to provide examples of larger syntactic structures in which the trigrams occur. Think about ways in which these can be used in EAP practice.

26. Many university students who are non-native speakers of English have problems with expressing their opinions both clearly and politely. Use examples from *ADVICe* to create a teaching material that would help students with the appropriate ways of expressing their positions when talking to their lecturers.

Tip: Search for key words such as *I think, opinion, I believe, maybe, possibly, perhaps* etc.
Use word cloud visualisation based on the epistemic markers dataset. Use the word-tree tool.

Table 4 Comparison of *BNC Demographic* and *ADVICe*

Word form	NF: *ADVICe*	% *ADVICe*	NF: *BNC Demographic*	% *BNC Demographic*	Overused in *ADVICe*	LL-score
he	333	0.20	6431	1.28	-	1952.43
she	301	0.18	4454	0.89	-	1142.68
n't	1238	0.74	9254	1.84	-	1125.08
yeah	4619	2.78	7532	1.50	+	1019.55
oh	548	0.33	4793	0.95	-	734.50
erm	1865	1.12	2477	0.49	+	672.67
so	2033	1.22	2867	0.57	+	645.74
of	2437	1.47	3791	0.76	+	611.46
said	142	0.09	2207	0.44	-	586.11
no	588	0.35	4391	0.87	-	532.99
him	40	0.02	1353	0.27	-	522.83
er	389	0.23	3267	0.65	-	473.79
got	450	0.27	3425	0.68	-	428.77
her	60	0.04	1259	0.25	-	399.46
well	581	0.35	3826	0.76	-	369.78
they	1033	0.62	5702	1.14	-	364.85
that	4676	2.81	10177	2.03	+	327.79
literature	114	0.07	0	0.00	+	317.10
the	5873	3.53	13345	2.66	+	315.90
research	140	0.08	13	0.00	+	307.89
kind_of	145	0.09	20	0.00	+	292.89
is	2270	1.36	4347	0.87	+	292.28
go	145	0.09	1509	0.30	-	284.23
part	183	0.11	57	0.01	+	278.53
data	103	0.06	4	0.00	+	254.65
need	354	0.21	307	0.06	+	247.40
because	640	0.38	827	0.16	+	243.77
in_terms_of	92	0.06	2	0.00	+	237.69
says	42	0.03	763	0.15	-	223.72
his	55	0.03	833	0.17	-	217.28
sort_of	343	0.21	326	0.06	+	213.67

Table 5 Trigrams in *ADVICe*

Trigram	Frequency
yeah yeah yeah	190
a lot of	134
i don't know	112
you need to	105
a little bit	83
i mean i	68
i think i	68
i don't think	62
and i think	61
i think you	57
you want to	55
one of the	55
be able to	54
you know what	54
i think it's	54
part of the	51
i think that	51
yeah i think	51
some of the	50
i think that's	49

8 Sample transcripts

Transcript 5 – School of Biology [311words]

T: [.] obviously the first thing is that this is our very first meeting, [Sx: hm] so we haven't even got to a point where we need to tick off objectives [Sx: yeah] for the first year, so that's no stresses about it. [...] and i guess the obvious thing is, [Sx: hm] you know <LAUGH> you've come from a background that isn't [.] hard science
S: that's right.
T: how have you found stepping up to [.] this challenge?
S: erm, yeah, no, it's it's been good, erm, you know i i spent i had the luxury of you know [.] starting before Christmas, i had the sort of Christmas period to think about erm how i wanted to do things and then think about sort of the equipment. and erm so ah i think that was quite useful, and and i did a lot of practicing on myself to see what seemed to work and what didn't seem to work. erm and really i haven't, you know modified the sort of methodology much at all, it seems to be going fine. erm i mean the big issue is really how people would respond [Tx: sure] to having their hand [Tx: yeah] in the water and how hard it [Tx: yeah] would be to find people to participate.
T: yeah. ok w- w- w- w- we'll come to sort of some of the details. i mean i think the main [.] one of [.] sort of informal things i'm really keen to know is (whether) you feel comfortable in the environment [Sx: oh] and and interactions with other people who may have had a stronger science background. [Sx: yeah] and how do y- how do you find just being in the lab there (xx)?
S: yeah no, i i really enjoy it. (xx) they're they're a good bunch in there.

Brezina, V. 2012. *ADVICe* - corpus. Available at www.lknol.com/advice.html.

Transcript 6 – Department of Geography [254 words]

T: [...]ok cool right see you <NAME> cheers. sorry it started as a quick one and i thought he was talking [Sx: ok] to me and it would be a bit longer. hi.
S: hi ok so i'm kind of half way through my stuff. i'm still transcribing.
T: right
S: but i've got most of it done so that's all that there
T: cool
S: erm i did like a little like kind of plan for my report but it's really brief about you know what i'm going to say in my results and stuff
T: ok
S: so you can like get that
T: yeah. it's great looks pretty detailed from here
S: [well not really i'm just kind of answer-]
T: [just that it's just that] it's two pages worth
S: well that's not really it it's just like erm answering my [Tx: yeah] kind of questions
T: well tha- that's great i mean you obviously picked up on the the idea that that that research questions can be a very useful starting point for a structure and you've clearly done that here, because the the structure is not [.] very specific actually other than we suggest there should be a theoretic- evident theoretical review and there's the evidence that, reporting of of findings and analysis and how you and and in- an introduction and a conclusion and apart from that [Sx: yeah] and the rest of it is
S: whatever you want
T: the the the the the superstructure around it. erm

Brezina, V. 2012. *ADVICe* - corpus. Available at www.lknol.com/advice.html.

References

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, Douglas, Randi Reppen, Victoria Clark, and Jenia Walter. 2001. " Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus." In *Corpus linguistics in North America: Selections from the 1999 symposium*, edited by Rita C. Simpson and John M. Swales, 48-57. Ann Arbor: University of Michigan Press.

Burnard, Lou. 2007. "Reference Guide for the British National Corpus (XML Edition)." In. Oxford: Research Technologies Service at Oxford University Computing Services http://www.natcorp.ox.ac.uk/XMLedition/URG/.

Crowdy, Steve. 1993. "Spoken corpus design." *Literary and linguistic computing* no. 8 (4):259-265. doi: 10.1093/llc/8.4.259.

———. 1994. "Spoken corpus transcription." *Literary and linguistic computing* no. 9 (1):25-28. doi: 10.1093/llc/9.1.25.

———. 1995. "The BNC spoken corpus." In *Spoken English on Computer*, edited by Geoffrey Leech, Greg Myers and Jenny Thomas, 224-234. Harlow: Longman.

Fetzer, Anita. 2004. *Recontextualizing context: Grammaticality meets appropriateness*. Amsterdam: John Benjamins.

Simpson-Vlach, Rita C., and Sheryl Leicher. 2006. *The MICASE Handbook: A Resource for Users of The Michigan Corpus of Academic Spoken English*. Ann Arbor: University of Michigan Press.

Stubbs, Michael. 2001. "Texts, corpora, and problems of interpretation: A response to Widdowson." *Applied Linguistics* no. 22 (2):149-172.

Swales, John M. 2002. "Integrated and fragmented worlds: EAP materials and corpus linguistics." In *Academic discourse*, edited by John Flowerdew, 150–164. Harlow: Longman.

———. 2006. "Corpus linguistics and English for academic purposes." In *Information technology in languages for specific purposes*, 19-33. New York: Springer.

Vine, Elaine W. 2011. "High frequency multifunctional words: accuracy of word-class tagging." *Te Reo* no. 54:71-82.

Widdowson, Henry G. 1998. "Context, community, and authentic language." *TESOL Quarterly* no. 32 (4):705-716.

———. 2000. "On the limitations of linguistics applied." *Applied Linguistics* no. 21 (1):3-25. doi: 10.1093/applin/21.1.3.

If you have any comments, questions or suggestions, I would be very happy to hear from you. Please get in touch at **v.brezina@lancaster.ac.uk**

Notes