Technical Report

UCAM-CL-TR-915
ISSN 1476-2986

Number 915

UNIVERSITY OF
CAMBRIDGE

Computer Laboratory

Annotating errors and disfluencies in transcriptions of speech

Andrew Caines, Diane Nicholls, Paula Buttery

December 2017

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

This document presents our guidelines for the annotation of errors and disfluencies in transcriptions of speech. There is a well-established precedent for annotating errors in written texts but the same is not true of speech transcriptions. We describe our coding scheme, discuss examples and difficult cases, and introduce new codes to deal with features characteristic of speech.

We present our guidelines for the annotation of errors and disfluencies in transcriptions of speech. This document is largely derived from Diane Nicholls' manual for error coding of learner corpora [14]. Her coding scheme was designed for written English: what we add here are amendments to deal with the characteristics of spoken language in general (specifically with English in mind).

Thus we build on the scheme set out in [14] with extra codes for features of speech transcription which relate to both production (*e.g.* disfluencies) and perception (*e.g.* inaudible words). In case there's any doubt as to the difference between spoken and written language for not only non-native speakers but also native speakers of English, consider these two examples from British English speakers in the British National Corpora:

(1) They were typical of part of what it **was like** to be homeless – having nowhere to go; having to avoid all representatives of authority; feeling tired and generally run-down; and needing to have my wits at their sharpest at a time when they had become critically undernourished.

(*Part of the furniture.* Falk, Michael. London: Bellew Pub. Ltd, 1991)

(2) S0315: I mean I said to friend once said oh –UNCLEARWORD another baby on the way ? she went no I **was like** ah oh sorry
S0255: » yeah it 's er it 's a food baby

(text S28F)

Clearly these two examples were selected to illustrate the differences between speech and writing – one could readily set out to illustrate the *similarities* and overlap between more formal genres of speech and less formal genres of writing. But they are fairly representative examples in the sense that both were found with relative ease on the first page of results for the string query, `was like`, firstly using BYU-BNC[1] [6] and secondly using CQPweb for BNC2014[2] [10].

Example (1) displays features characteristic of writing in the sense that (a) it contains a series of subordinate clauses, (b) it demarcates those clauses with punctuation, and (c) it is completely grammatical. These features arise from the time usually available to writers, and the consequent opportunity to edit and reshape texts. Subordinate clauses of course also occur in speech, but punctuation does not (prosody performs a similar function) and thus the units are not so neatly demarcated and are not always so grammatical, as we see in example (2).

[1] `http://corpus.byu.edu/bnc`

[2] `http://corpora.lancs.ac.uk/bnc2014`

In example (2) we see a conversation between two speakers featuring reported speech, discourse markers (*I mean, oh, yeah*), interjections and filled pauses (*oh, ah, er*), repetition (*it's er it's*), muffled or mumbled speech (UNCLEARWORD), ungrammatical sequences (*I said to friend once said*), and non-standard language (*she went no; I was like ah oh sorry*). These features are typical of the rapid, immediate, uneditable nature of speech, the tendency of speakers to self-monitor and self-correct, and the ability of speakers to innovate and interact to produce meaningful utterances in novel ways.

All of these speech features are of interest to us and hence we have prepared this document to codify the kind of 'error' annotation we will carry out on transcripts of speech. Note that by use of the word *error* we are meeting the terminology of the Applied Linguistics field, where *error* annotation tends to be undertaken on written essays, the division between accurate and inaccurate use of language is (slightly) clearer, and the learner of English straightforwardly wants correction on facets of writing such as spelling, subject-verb agreement, word order, and so on.

The status of errorful *speech* is less certain: *acceptability* comes to the fore and the successful delivery of meaning becomes more important than the wholly-grammatical delivery of meaning, for native speakers as much as non-native speakers. Interlocutors can repair their own or each other's utterances, or ask for clarification when they cannot, and we may therefore relax our strict notion of 'error' being misuse of language [15, 16, 7, 8]. Instead, we appeal to notions of gradient acceptability and propose that errors in speech relate more to fluency and communicative intent than to form and rules of grammar [2].

Having said that, we also have the Computational Linguistic field in mind, and the potential use of error-annotated corpora in building computer-assisted language learning (CALL) systems. In this context, errors are at least those disfluent word tokens we need to replace or remove in order that natural language processing (NLP) tools have the best chance of analysing the transcripts. Since NLP tools tend to be trained on and/or designed for grammatical written inputs, it has been shown that their performance degrades on 'as is' spoken inputs [1, 12, 13, 4].

'Cleaning up' transcripts to be more written-like is therefore one of our main concerns; on top of this we will also continue to annotate clearly ungrammatical language, with a view to helping learners of spoken English improve through automated feedback in CALL systems by providing more fluent 'native-like' versions of what they said. Also we will mark mispronunciations, where the original recording is available to us – a move made possible by provision of phonemic transcriptions in ARPAbet format [9].

This document is **not** a guide to speech transcription. In (2) there is an example of speaker overlap as indicated by the » double angled brackets. Here we see another complication of representing speech in writing: the fact that in conversation speakers will overlap, interrupt and co-construct their turns [5]. Transcribers may also need to mark paralinguistic features of speech recordings such as silences, laughter, coughs and sneezes, and background noise. For a comprehensive guide to speech transcription, we refer the reader to the BNC2014 manual [11].

We set out the error codes in Table 1 and give examples from the BULATS corpus provided by

Cambridge Assessment English[3]. The following are taken from [14]: F, M, R, U, D, I, AG, the word classes, CE, ID, W and X. The remaining codes have been added to deal with holistic phrase correction in the style of [3] – namely FL – and features of speech (recordings): IA, PW, PR, FP, DM, FS, RE, CO.

Not all codes have been brought over from [14], since codes such as 'P' (punctuation) or 'S' (spelling) are not relevant to speech transcripts – go to `http://ucrel.lancs.ac.uk/publications/CL2003/papers/nicholls.pdf` for the full list of error annotation codes for *written* English.

Finally, note that it is our convention to format the XML error tags in the following way –

<NS type='*X*'><i>...</i><c>...</c></NS>

- NS means 'non-standard'
- *X* is the error code
- <i>...</i> demarcates the 'incorrect' portion of text (optional; i.e. irrelevant in the case of missing word tokens)
- <c>...</c> contains the error correction (also optional; i.e. unnecessary when the word tokens should only be deleted)

Acknowledgements

This paper reports on research supported by Cambridge Assessment English, University of Cambridge. We thank Claire Dembry of Cambridge University Press for discussion, and Kate Knill of Cambridge University Engineering Department for preparation of BULATS recordings and transcriptions.

[3] Business Language Testing Service `http://www.cambridgeenglish.org/exams-and-tests/bulats`

Code	**Definition**	**BULATS example**
major types		
F	when a word is a word but the form is not the right one for the context	good service <NS type='FN'> <i>person</i><c>people</c></NS> are very helpful
M	when there's a word or phrase missing (insertion)	it's very important to help <NS type='MT'><i></i><c>in</c> </NS> job interviews
R	when the word or phrase are valid word(s) and the correct part(s)-of-speech but needs replacing (substitution)	the visitor can go to <NS type='RD'> <i>the</i><c>a</c></NS> restaurant
U	when a word or phrase is valid but superfluous or inappropriate in context (deletion)	will be the open space one <NS type='UY'><i>like</i></NS> everybody can see each other
D	when a word is wrongly derived, word derivation being the conversion of a given form from one word class to another (e.g. noun to adj: *spite, spiteful*)	there was great <NS type='DN'> <i>succeed</i><c>success</c> </NS> in twenty ten
I	when a word is incorrectly inflected, where inflection involves morpheme insertion/replacement/deletion (e.g. ask, asks, asked, asking)	I read the <NS type='IN'> <i> informations</i><c>information </c></NS>
AG_	agreement errors: number, person and gender agreement, most often between noun and verb (*he say*, he says*) but also determiner and noun (*some ship*, some ships*), pronoun co-reference (*the woman ... he*, the woman ... she*), etc	they <NS type='AGV'><i>is</i> <c>are</c></NS> far away
optional word class		
_A	pronoun; e.g. *he, she, it, I, me, who, whom*	many companies sell <NS type='FA'> <i>they</i><c>their</c></NS> products online
_C	conjunction; e.g. *and, or, but*	in two <NS type='MC'> <c>or</c></NS> in four years you will see

Table 1: Error codes for annotation of speech transcripts, selected from those in [14], with new codes introduced to annotate speech features

optional word class		
_D	determiner; e.g. *the, a, that, this*	some recruiting website such as <NS type='UD'><i>the</i><c></c></NS> jobteevee dot com
_J	adjective; e.g. *happy, sad, blue, hilarious*	the customer can return these <NS type='FJ'><i>unwanting</i><c>unwanted</c></NS> goods
_N	noun; e.g. *error, annotation, learner, corpus*	they are building new <NS type='FN'><i>home</i><c>homes</c></NS> at the same level
_Q	quantifier; e.g. *many, much, some, few*	I have <NS type='RQ'><i>much</i><c>a lot of</c></NS> technical background
_T	preposition; e.g. *in, on, to, towards*	when they stay <NS type='RT'><i>at</i><c>with</c></NS> us
_V	verb; e.g. *come, go, speak, improve*	because it <NS type='MV'><c>is</c></NS> very important
_Y	adverb; e.g. *very, partly, quickly, sometimes*	so they feel really <NS type='RY'><i>homely</i><c>at home</c></NS> in there
other errors		
CE	complex error: when the intended sense of the words cannot be established and therefore cannot be corrected (the code of last resort)	<NS type='CE'><i>And you can you can travel by airplane just in the other you you</i></NS>
FL	fluency error: when a string of words needs to be rephrased to improve its clarity, coherence or appropriateness	<NS type='FL'><i>sometimes are so</i><c>so sometimes</c></NS>
ID	idiom error; where the lexical construction of an idiomatic phrase is incorrect in some way	you can <NS type='ID'><i>see him in his face</i><c>meet him face to face</c></NS>
W	word order error	I can start <NS type='W'><i>in school the business</i><c>the business in school</c>
X	negation error	you don't make any points at the meeting you <NS type='X'><i>no</i><c>won't </c></NS> know

speech features		
IA	inaudible word; where for reasons of speaker production or recording factors what was said cannot be perceived (in BULATS transcriptions the inaudible word token is denoted by ‘%unclear%’)	together with percentile <NS type='IA'>%unclear%</NS>
PW	partial word; where the speaker starts and interrupts a word token – note that the speaker may have produced enough of the word that its identity is unambiguous, in which case the full word can be transcribed, or if not the partial word may be transcribed orthographically or phonemically using ARPAbet symbols	by a lot of <NS type='PW'><i>people</i></NS>; <NS type='PW'><i>withou</i></NS> using emails
PR	pronunciation error; where the speaker mis-pronounces a word token (ARPAbet format)	<NS type='PR'><i>K AE R OW K EY </i><c>K EH R IY OW K IY</c></NS> (‘karaoke’)
FP	filled pause; tokens such as *er* and *um* with which the speaker hesitates before continuing to speak (in BULATS transcriptions filled pauses are denoted by ‘%hesitation%’)	<NS type='FP'><i>%hesitation%</i></NS>
DM	discourse marker; words and phrases such as *yeah, well, I mean, you know* which serve to maintain discourse coherence but can also be omitted without altering meaning (note that these examples are **not** always discourse markers)	<NS type='DM'><i>Well</i></NS> I think you will go to the bank
FS	false start; when the speaker begins a multi-word phrase but interrupts part way through, resuming their utterance with a self-correction (‘reparandum’)	<NS type='FS'><i>not no</i><c>with no regard</c></NS> for his age
RE	repetition; when the speaker reuses the same token or phrase two or more times consecutively without self-correction	<NS type='RE'><i>if</i><c>if</c></NS> he is very motivated
CO	cut-off; where the end of the recording has interrupted what the speaker was saying – an extra-linguistic property but one which affects machine reading of transcripts all the same	you could almost have your personal assistant there <NS type='CO'><i>asking them for anything that you</i></NS>

Bibliography

[1] Andrew Caines and Paula Buttery. The effect of disfluencies and learner errors on the parsing of spoken learner language. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, 2014.

[2] Andrew Caines, Paula Buttery, and Michael McCarthy. Darker shades of grey: experiments in multi-dimensional error analysis. Presentation at the IVACS Symposium, Universitat Pompeu Fabra, 2016.

[3] Andrew Caines, Emma Flint, and Paula Buttery. Collecting fluency corrections for spoken learner English. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 2017.

[4] Andrew Caines, Michael McCarthy, and Paula Buttery. Parsing transcripts of speech. In *Proceedings of the First Workshop on Speech-Centric Natural Language Processing*, 2017.

[5] Ronald Carter and Michael McCarthy. Spoken grammar: where are we and where are we going? *Applied Linguistics*, 38:1–20, 2017.

[6] Mark Davies. BYU-BNC (based on the British National Corpus from Oxford University Press), 2004-.

[7] Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and Nick J. Enfield. Universal principles in the repair of communication problems. *PLOS ONE*, 10:1–15, 2015.

[8] Alice Foucart, Elisa Ruiz-Tada, and Albert Costa. How do you know I was about to say "book"? Anticipation processes affect speech processing and lexical recognition. *Language, Cognition and Neuroscience*, 30:768–780, 2015.

[9] Aldebaro Klautau. *ARPABET and the TIMIT alphabet*, 2001.

[10] Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22:319–344, 2017.

[11] Robbie Love, Abi Hawtin, and Andrew Hardie. *The British National Corpus 2014: User manual and reference guide*, 2017.

[12] Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery. Incremental dependency parsing and disfluency detection in spoken learner English. In *Proceedings of the 18th International Conference on Text, Speech and Dialogue (TSD)*. Berlin: Springer-Verlag, 2015.

[13] Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery. Automated speech-unit delimitation in spoken learner English. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016.

[14] Diane Nicholls. The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of the Corpus Linguistics 2003 conference; UCREL technical paper number 16*. Lancaster University, 2003.

[15] Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735, 1974.

[16] Emanuel Schegloff. When ‘others’ initiate repair. *Applied Linguistics*, 21(2):205–243, 2000.