UNNC-CAWSE

University of Nottingham Ningbo China

Corpus of Academic Written and Spoken Corpus

TRANSCRIPTION CONVENTIONS

For the Spoken Subcorpus [v.1]

December 2017

Yu-Hua Chen & Qianqian Zhou

UNNC

Ningbo, China

Overview

Type (subtype in alphabetical order)	**Definition**	**Example**
1. Interactional		
1.1 Speaker Turns	Speaker turns are indicated by individual speaker IDs followed by a colon ':' and utterances. Each speaker is labeled with their unique ID in the format of t/s/r plus four digits of Arabic numbers: 't' for teacher, 's' for student, 'r' for researcher and 'x' for unknown speaker.	t0001: *right okay thank you that's the end of the test* s0001: *okay* t0001: *you're free to go*
1.2 Latching	The lack of a pause between different speakers is marked by '=' in the utterances: one at the end of the first speaker and the other at the beginning of another where latching occurs.	s0001: *so er: I think from you two=* s0002: *=OK*
1.3 Overlaps	The tagset and are used for overlaps, and two pairs are used where the overlap occurs: one with the first speaker, and the other with the second speaker.	s0005: *oh oh* *oh* s0002: *you know* *this one is much sweeter than this one*
2. Verbal		
2.1 Acronyms	2.1.1 If an acronym is pronounced as a sequence of letters, it is transcribed as a sequence of capital letters separated by spaces.	*I'll now ask you some general questions about U N N C life OK*
	2.1.2 If an acronym is uttered as a word, it is transcribed as a sequence of capital letters without any spaces.	*I'm doing a TESOL*
2.2 Capitalisation	Capital letters are NOT used at the beginning of sentences. They are only retained as required in spelling conventions such as	*so er: I think from you two*

	proper nouns (e.g. 'New York', 'Ningbo'), first personal pronoun 'I', or 'Mr', 'Mrs', 'Dr', etc.	
2.3 Code Switching	Utterances in Chinese are marked up by the tagset <cs n= "zh"> and </cs>. The language code is specified after the attribute "n", and in this case the code for Chinese is "zh". Translations into English are provided wherever possible in curly brackets {} after the utterance of code switching for those who may not understand Chinese. If it is neither English nor Chinese, the language code can be found in the Library of Congress (US): https://www.loc.gov/standards/iso639-2/php/code_list.php	<cs n="zh">*那个地壳是不是{is that earth crust}*</cs> *or maybe on some fa- some face* <cs n= "zh">表情怎么说啊*{how to say 'facial expression'}*</cs>
2.4 Contractions	Standard spelling of all contractions are retained.	*I'm* *she's* *you'd* *we've*
2.5 Fillers/Filled Pauses	All filled pauses are standardized in orthography and marked as one of the following: *ah, en, er, erm, huh, mhm, mm, oh, uh,* or *uhu*. No other fillers are used.	s0001: *mm: when I get up early*
2.6 Lengthening	Lengthened sounds are represented by the symbol of a colon ':'.	s0008: (2.3) hh (1.4*) just er*: *ok animals human* s0001: *oh*: *oh*:
2.7 Lexicalised Reduced Forms	When a lexicalised reduced form is uttered, the original shortened form (e.g. *gonna*) is rendered as opposed to a full standard form.	*cos* *kinda* *gonna* *gotta* *wanna*
2.8 Numbers & Dates	Numbers and dates are written out in full words in the same way of how they are uttered.	*nineteen ninety nine* (rather than 1999)
2.9 Orthography & Hyphenation	2.9.1 Both hyphenation and spelling follows the rules of British English, and the online Oxford Dictionary is consulted: (https://en.oxforddictionaries.com/). For example, for the suffix of *-ise/-ize,* only the *-ise* variant is used rather than the *-ize* variant although both morphemes are used in British English. Other common instances include suffixes such as *-our* (e.g. *colour), -re* (e.g. *centre*).	
	2.9.2 In the case of more than one spelling variations available for a word in the Oxford Dictionary, the first variation is chosen unless it involves British/American spelling. For	*OK all right we'll now move on to part two*

	example, the spellings of *OK* and *all right* are standardised (as opposed to, for example, *okay* or *alright*).	
2.10 Punctuation	No punctuation is used for sentence or clause boundaries.	*okay right now I'm gonna move on to part two within part two I want you to give a short speech on a topic that I will give you*
2.11 Repetition	All repetitions of utterances are transcribed.	*sometimes er*: *er the first thing I I will consider is the quality of the er*: *product*
2.12 Unintelligible	Unintelligible speech is represented by <ut>x</ut> regardless of word length. The number of 'x' approximates the number of words heard. For example, <ut>xxx</ut> indicates three words.	*can you can you discuss the article* <ut>xx</ut>
2.13 Word Fragments/ Truncation	A hyphen is used to mark word fragments (such as self-interruption or self-correction), which indicates a truncation at the beginning or the end of a word.	=*wha- what you mean biology*
3. Vocal (non-verbal)		
3.1 Exhalation/ Inhalation	Noticeable breathing is marked by 'hh'.	*I think* (1.3) *it depends* hh *er*: *they can the school test can test something*
3.2 Laughing	3.2.1 If the current speaker is laughing, use <laughing> in the utterances where the laughter occurs.	*so confused* <laughing>
	3.2.2 If more than one person is laughing, speaker codes are added in the tag to specify who is laughing.	<s0001 and t0001 laughing>
	3.2.3 If everybody is laughing, use <all laughing>.	<all laughing>
3.3 Pauses	A pause equivalent to or longer than one second is represented by the duration of time within parentheses (), correct to one decimal place.	hh (1.4) *just er*: *ok animals human*
4. Non-vocal		
4.1 Non-vocal Communicative (NVC) Events	Three frequent NVC events are noted: <writing>, <nodding> and <shaking head>. No other NVC event is indicated.	s0005: <nodding>
5. Others		
5.1 Anonymisation	Anonymisation: The tagset <anm>x</anm> is used to anonymise the data for the sake of ethics, e.g. names of students, staff, family, friends, or any other sensitive information. The number of 'x' approximates the number of words heard. If it is unclear whether it is sensitive information, the rule of thumb for the transcribers is to simply anonymise the	t0001: *great what is the South African student's name* s0001: *er* <anm>*xx*</anm>

	names. However, since the data is collected from UNNC, UNNC and its associated proper nouns such as building names do not need to be anonymised.	
5.2 Time Stamps	5.2.1 For classroom video recordings, time stamps for every ten minutes are added.	<10 mins> <20 mins>
	5.2.2 For audio recordings of speaking assessment, preparation times are edited out, but the duration of time is indicated in the tag <prep time XX min(s)>.	<prep time 1 min> <prep time 2 mins>
5.3. Deviation	• Two types of deviations are covered here: lexical and pronunciation (including a slip of the tongue which may otherwise be considered transcribing errors). • 1) Lexical errors in relation to forms (i.e. formal misselection and misformation) and semantics (i.e. collocation/word choice) and 2) pronunciation errors which may hinder comprehension are preliminarily tagged and will be illustrated below. • Note that this is not intended to be used for grammatical errors (e.g. subject-verb agreement).	
	5.3.1 If 1) there is a clear deviation (in terms of pronunciation or lexis) and 2) both the intended and deviation words are clear in the context, the tagset <dvp/dvl/dv>DEVIATION{CORRECTION}</dvp/dvl/dv> is used. The corrected word is provided in the curly brackets {} following the deviation, unless the speaker corrects it themselves immediately. The attribute 'dvp' is used for pronunciation deviation while 'dvl' refers to lexical deviation. When it is unclear which category the deviation falls into or the deviation may involve both pronunciation and lexical levels, then 'dv' is used.	**Pronunciation deviation (with an issue of intelligibility) <dvp>:** *it said that our* <dvp >*hurt* {*heart*}</dvp> *er have four chambers* **Lexical deviation <dvl> (1) Slip of the tongue**: *the description of the movement of the water between ocean er* <dvl>*background*{*underground*}</dvl> *water and er atmosphere* **(2) Part-of-speech (POS)** *er: some music bands will er: maybe* <dvl>*creative*{*create*}</dvl> *some music* **(3) Derivation** *I stay in the library* <dvl>*alonely*{*alone*}</dvl> **(4) Inflection** *if you have so many* <dvl>*childrens*{*children*}</dvl>*in your in family* **(5) Collocation/ word choice:** *I think sometimes we buy a* <dvl>*purchase*{*product*}</dvl> *not only for its good quality*

		Mixed/uncertain deviation <dv>: *maybe it's very bad for the* <dv>*economic*{*economy*}</dv> *er: to the country*
	5.3.2 If the deviation of utterance is unclear/un-transcribable or the transcriber is uncertain about the intended correct word, 'x' is used. The number of 'x' approximates the number of words heard, e.g. <dvp>xx{CORRECTION}</dvp> or <dvp>DEVIATION{x}</dvp>.	**Pronunciation deviation <dvp>:** <dvp>*x*{*osmosis*}</dvp> *is a kinds of maybe a material and which only let's er water to go through it* **Lexical deviation <dvl>:** *I think art is a* <dvl>*virus*{*x*}</dvl> *it can include traditional art and also the art er graffiti*