MARK-UP CONVENTIONS.

Adapted, with the authors' permission, from VOICE (Version 2.1 June 2007)

VOICE. 2013. *The Vienna-Oxford International Corpus of English* (version POS XML 2.0). Director: Barbara Seidlhofer; Researchers: Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka, Nora Dorn.

1. SPEAKER IDs

Example: AD	Speakers are identified by the initials of the pseudonym they have been assigned. The full pseudonym (and information about the speaker) is provided in the 'Contextual Information' accompanying each transcript. The speaker ID is given at the beginning of each turn.
R	Occasionally, the researcher responsible for recording the conversations participates in the conversation. The ID assigned to this speaker is always the same, regardless of the identity of the researcher involved.

2. INTONATION

Example: PP: sorry**?**	Intonation is rarely indicated in the transcripts. A question "**?**" is used in the transcription to mark a rising tone when it is important to disambiguate questions and statements in context.

3. EMPHASIS

Example: yes you will **DE**finitely	If a speaker gives a syllable, word or phrase particular prominence, this is written in capital letters.

4. PAUSES

Example: NT: yeah **(.)** so i mean in theory the seminars were there to **(.)** test your knowledge	Every brief pause in speech (up to a good half second) is marked with a full stop in parentheses.
Example: RJ: yes i wish i'd kept the erm **(1)** the piece of paper	Longer pauses are timed to the nearest second and marked with the number of seconds in parentheses, e.g (**1**) = 1 second, (4) = 4 seconds, etc.

5. GAPS

Example: BV: how did you find the assignment NG: **(1)** it was i think that it was more difficult than the previous one	When there is a gap between turns, this is annotated using the same conventions as for pauses, but annotated at the beginning of the current speaker's turn.

6. **OVERLAPS**

Example: DC: yeah <23> yeah </23> AMT: <23> or </23> race relations	Whenever two or more utterances happen at the same time, the overlaps are marked with numbered tags: <1> </1>, <2> </2>, etc. Everything that is simultaneous gets the same number, <1> </1> marking the first overlap in each transcript, and the highest number the last one. All overlaps are marked in blue
Example: DC: what was the question you did for the first essay (1) can you <5> remember </5> AMT: <5> i actually 5> (.) don't remember it (.) i wrote about the (.) the negro em	All overlaps are approximate. However, where only one or two syllables of a word overlap, the whole word is included within the tags, with the overlapping syllables marked in blue.

7. OTHER-CONTINUATIONS

Example: AMT: because it's ginsberg the one with the idea of the poet being**=** DC: **=**the vision and	Whenever a speaker continues, completes or supports another speaker's turn immediately (i.e. without a pause), this is marked by "**=**"

8. LENGTHENING

Example: JM: if we need to**:** to pass e**:** both separately in order to pass the subject	Lengthened sounds are marked with a colon ":".

9. REPETITION

Example: JW: **er er** and erm there is an (.) an extra **cri-** criterion (.) placed in here	All repetitions of words and phrases (including self-interruptions and false starts) are transcribed

10. WORD FRAGMENTS

Example: OW: so largely that it's it's it's it's it's mainly (.) wh**-** what i look for is mainly argument	With word fragments, a hyphen marks where a part of a word is missing.

11. LAUGHTER

Example: LR: **@@** yeah yeah for **<@> me </@>**	All laughter and laughter-like sounds are transcribed with the @ symbol, approximating syllable number (e.g. ha ha ha = @@@. Utterances spoken laughingly are put between **<@> <@/>** tags.

12. UNCERTAIN TRANSCRIPTION

Example: JW: a:h i (**rarely**) get these	Word fragments, words or phrases which cannot be reliably identified are put in parentheses ( ).

13. UNINTELLIGIBLE SPEECH

Example: IS: cause my my (.) my degree in spain is a five year degree (.) so: i <9> <**un**> xxx </**un**> </9>	Unintelligible utterances are represented by **x's** approximating syllable number and placed between <**un**> </**un**> tags.

14. PRONUNCIATION VARIATIONS AND COINAGES

Example: MA: i'm worried about the assessment o:n on **<pvc> jan {june} </pvc>**	Striking variations on the levels of phonology are marked as <**pvc**> </**pvc**>. What is heard is represented in spelling according to general principles of English orthography. If a corresponding existing word can be identified, this existing word is added between curly brackets { }.

15. ONOMATOPOEIC NOISES

Example: SB: so if you just say what's <spel> k k h </spel> and then **<ono> buf </ono>** you get a lot	When speakers produce noises in order to imitate something instead of using words, these words are represented in spelling according to general principles of English orthography between <**ono**> </**ono**> tags.

16. NON-ENGLISH SPEECH

Example: PP: in spa:nish we said **<L1sp> anquilosado </L1sp>**	Utterances in a participant's first language (L1) are put between tags indicating the speaker's **L1.**
Example: AM: **<LNsp> anquilosado </LNsp>** i don't know what it is	Utterances which are neither English nor the speaker's first language are marked **LN** with the language indicated.

17. FIRST LANGUAGE INFLUENCE

Example: HV: i've instructed them not to focus on details	Words and phrases that reflect the morphology or phraseology of the first

(.) but on the the <**L1dnlinf**> red line </**L1dnlinf**> (.) of their lecture Example: EB: > if you make plagiarism (1) that means that you're you may be ex- <**L1spinf**> expulsed </**L1spinf**> from the university	language are put between tags indicating the **inf**luence of the speaker's **L1**.

18. SPELLING OUT

Example: DC: <spel> c l o s u r e </spel> closure	The <**spel**> </**spel**> is used to mark words or abbreviations that are spelled out by the speaker, i.e. words whose constituents are pronounced as individual letters.

19. SPEAKING MODES

Example: JW: i think i'm on here somewhere (.) erm (1) er <**soft**> sorry that's (.) the email writing (.) list </**soft**> Example: JW: and if we look at erm (.) er this is (.) erm (1) er <**reading aloud**> documentation and presentation other skills </**reading aloud**>	Utterances that are spoken in a particular mode (fast, soft, reading aloud from a text, etc.) and are notably different from the speaker's normal speaking style are marked accordingly. The description of the speaker's demeanour provided in the 'Contextual Information' that accompanies each transcript helps define what is 'normal' for each speaker in each conversation.
<fast> </fast> <slow> </slow> <loud> </loud> <whispering> </whispering> etc.	The list of speaking modes is an open one.

20. BREATH

Example: DC: **hh** what was the question you did for the first essay	Noticeable breathing in or out is represented by two or three **h's** (hh = relatively short; hhh = relatively long)

21. SPEAKER NOISES

<coughs> <clears throat> <sniffs> <sneezes> <yawns> etc.	Noises produced by the current speaker are transcribed. The list of speaker noises is an open one.

22. ANONYMIZATION

Example: JW: what (.) will generally happen erm i've sent two dissertations to (.) erm er doctor <**pseud**> williams </**pseud**> Example: EC: you have to take into consideration first of all first years (.) am are getting used to the whole system getting used to the university (.) getting used to <31> <pseud> ireland </pseud> </31>	This project seeks to protect the anonymity of the people that took part in these conversations, and that of others who may be mentioned. Names of people, institutions, exact locations, etc. are replaced by pseudonyms and put between <**pseud**> </**pseud**> tags. The names of the towns or universities where the recordings took place are replaced by the name of the country and likewise put in <**pseud**> </**pseud**> tags.

23. TRANSCRIPTION BORDERS

Example: <**beg** UE2_00:31>	The **beg**inning of each transcript is noted by indicated by indicating the code for the transcript and the exact position of the track on each recording in minutes and seconds.
Example: <**end** UE2_11:00>	The **end** of the transcript is noted in the same way.

24. TIME STAMPS

Example: 05:30 (the quotation)	Each transcript is time-stamped every 30 seconds, with the word or phrase that is uttered at that 30 second interval in parentheses after the minutes/seconds.

ADDITIONAL INFORMATION

Any words or phrases that we judged likely to prove obscure to the reader are glossed in footnotes in each transcript. For example, 'hazop' = 'Hazard and Operability'.