

Transcription Conventions for Exam Scripts (UNNC CAWSE) v.1

Basic principles:

1. The content is transcribed as closely as possible to what is seen.
 - **English variety:** Both American English and British English are transcribed as they are observed in the scripts.
 - **Capitalisation and punctuation:** Type the words and punctuation as they are seen in the script. Typical examples can be found in Table 1:

Table 1 Common issues of capitalisation and punctuation identified in the UNNC CAWSE

Definition	Examples in the UNNC CAWSE
1. Proper nouns are not capitalised or capitalised incorrectly.	<i>iphone, ipod, ipad, or Iphone, Ipod, Ipad</i>
2. Sentences do not finish with full stops and/or begin with capital letters.	<i>they changed their logo because of public opinion, ..., the characteristics of micro-apartment that use share facilities is another key to solve the problem. compared to the normal-size apartments units,...</i>
3. A capital letter (not as part of a proper noun) appears in the middle of a sentence.	<i>So the new logo Here.</i>
4. Punctuation is not used correctly, e.g. a full stop in the middle of a sentence.	<i>In terms of the design of logo, it is obvious that there are many legal. issues associating with it. especially, the <dvl>copy right{copyright}</dvl>.</i>

- **Spelling errors:** Those are kept as they are but are annotated. See #2 below. Note that other types of deviations are NOT annotated. For example, grammatical errors such as tense (e.g. *he talks yesterday**) are not corrected.
 - **Exception:** Words that students crossed out (e.g. ~~linguistics~~) or put a cross before or after (e.g. x corporate, media x) are omitted.
2. Orthographic Deviations (same as Transcriptions Conventions for the Spoken Subcorpus v.1), annotations are carried out for the following items:
 - **Unintelligible words:** The tagset <ut>x</ut> is used to indicate unintelligible words. The number of 'x' approximates the number of words written.
 - **Orthography:** Only formal deviations at the level of orthography are annotated at this stage. This refers to words that deviate from standard orthography. The online Oxford Dictionary (<https://en.oxforddictionaries.com/>) is used as the reference. Ill-formed words can include a range of different conditions from simply missing or misplacing a letter to words that do not follow the rules from morphology or word formation. To annotate those deviations, the format below is used:

2.1 If the transcriber is able to infer the correct word, the deviated word is marked with <dvl>DEVIATION {CORRECTION}</dvl>.

Example: *At the same time, providing the same facilities the normal size <dvl>accomdadation{accommodation}</dvl> has is also a key to maintain equality in the whole city.*

Transcription Conventions for Exam Scripts (UNNC CAWSE) v.1

2.2 If the transcriber is unable to infer the correct word, 'x' is used to replace the correction in the form of <dl>DEVIATION{x}</dl>.

Example: *The construction of the architecture would be designed <dl>standable{x}</dl> and the regular check and protect should never be ignored.*

Table 2 Taxonomy of ill-formed words annotated in the UNNC CAWSE

Category	Sub-category	Definition	Examples in the UNNC CAWSE DEVIATION{CORRECTION}*
Misspelling (M)	M1 Omission	Letter(s) missing	<i>coffe{coffee}</i>
	M2 Redundancy	Letter(s) redundant	<i>discussed{discussed}</i>
	M3 Substitution	Letter(s) incorrectly replaced by other letter(s)	<i>seccessful {successful}</i>
	M4 Transposition	Letter(s) incorrectly placed	<i>aduiences{audiences}</i>
	M5 Mixed	Combination of at least two types of errors above	<i>Austrila{Australia}</i>
Word Formation (F)	F1 Derivation	Incorrect use of affixes in the creation of words	<i>unforgettable{unforgettable}, fastly{fast}</i>
	F2 Inflection	Incorrect use of inflections, which indicate the grammatical function of a word	
		Nouns: two inflections 1. possessive 's 2. plural -s, -es, -ies.	Noun: <i>Beatles's{Beatles'}, nutritions{nutrition}</i>
		Verbs: Four inflections 1. 3 rd person singular present tense -s, -es, -ies 2. present participle -ing 3. past tense -ed, -ied 4. past participle -ed, -ied, -en	Verb: <i>launchs{launches}, referring{referring}, rised{risen}</i>
	Adjectives: Two inflections 1. comparative -er, -ier 2. superlative -est, -iest	<i>hoter{hotter}, earlest{earliest}</i>	
	F3 stem	Incorrect use of stems. Often involved with incorrect use of part of speech (POS)	<i>differents{differences}, acrossing{across crossing}</i>
Proper nouns (PN)		Incorrect spellings involved with proper nouns	<i>Stephen Jompth{Steve Jobs}, Starbuck{Starbucks}</i>
Compound words (CW)		Involved with word segmentation: splitting or joining words incorrectly	<i>copy right {copyright}, eventhough {even though}, non standard {non-standard}</i>
Incorrect in the context (including a slip of the		This refers to the condition when the spelling of a word looks legitimate (i.e. in reference to the Oxford	Ex1: <i>Starbucks was founded in <dl>settle{Seattle}</dl>, ...</i> Ex2: <i>The brand of the Apple is really successful. Everyone can't</i>

Transcription Conventions for Exam Scripts (UNNC CAWSE) v.1

pen) (IC)		<p>dictionary) but it does not make sense in the context. This is different from other conditions listed above, for instance in that an automatic spell checker would not pick up the error.</p>	<p><i>forget its logo because it's fun and <dv>agentic{energetic}</dv>.</i> <i>Ex3: As an important factor to make the company become great successful, the <dv>log{logo}</dv> of Apple is very simple and memorable.</i> <i>Ex4: The mermaid as a <dv>well known{well-known}</dv> image once was criticized due to the <dv>nicked{naked}</dv>body.</i></p>
-----------	--	--	---

- * 1. The format follows the same rules as in the transcription and annotation conventions described in Section 2 where 'x' is used to indicate uncertain/unknown correction.
- 2. The taxonomy above is adapted from Bestgen and Granger (2011).