

Treebank of Learner English  
**Annotation Manual - DRAFT**  
Version 1.0, UD Release 1.3 (May 2016)

Yevgeni Berzak      Jessica Kenney      Carolyn Spadine  
Jing Xian Wang      Lucia Lam      Keiko Sophie Mori  
Sebastian Garza      Boris Katz

May 11, 2016

# Contents

<b>1</b>	<b>General</b>	<b>2</b>
<b>2</b>	<b>Format Overview</b>	<b>2</b>
2.1	Metadata . . . . .	2
2.2	POS and Dependency Annotation . . . . .	3
2.3	Corrected Version . . . . .	3
<b>3</b>	<b>ESL Specific Guidelines</b>	<b>4</b>
3.1	Instructions . . . . .	4
3.2	ESL Metadata . . . . .	8
<b>4</b>	<b>Clarifications for English UD Guidelines</b>	<b>13</b>
4.1	General Clarifications . . . . .	13
4.2	UD EWT Treebank and UD Guidelines Differences . . . . .	14
<b>5</b>	<b>Other</b>	<b>16</b>
5.1	Sentence Segmentation . . . . .	16
5.2	Capitalization . . . . .	16
<b>6</b>	<b>Quality Assurance</b>	<b>17</b>
6.1	Review Protocol . . . . .	17
6.2	Resolution Protocol . . . . .	17
<b>7</b>	<b>Annotation Utilities</b>	<b>19</b>
<b>8</b>	<b>Appendix A - Common Annotation Errors</b>	<b>20</b>
<b>9</b>	<b>Appendix B - FCE Error Tags</b>	<b>22</b>

# 1 General

Annotation manual used for annotating the Treebank of Learner English (TLE)[1]. The TLE contains manually annotated POS tags and Universal Dependencies (UD) trees for 5,124 sentences from the Cambridge FCE[2] learner corpus.

The annotation follows the English Universal Dependencies Guidelines. Conventions that address non canonical structures arising due to grammatical errors are described in ESL specific guidelines.

The FCE corpus was segmented to sentences with an adaptation of the NLTK sentence Punct tokenizer, and word tokenized with the Stanford PTBTokenizer. The sampled sentences contain at least one grammatical error that is not punctuation, spelling or generic missing/replace/remove.

## 2 Format Overview

### 2.1 Metadata

The metadata precedes the POS and dependency annotation. Metadata lines start with #.

1. ID: FCE file identifier.
2. ERR: FCE Error codes in the order of their appearance in the sentence.
3. ERR-NAMES: the corresponding error names.
4. SENT: the sentence in the FCE XML format.
5. SENT-TOKENS: tokenized version of the original sentence.
6. SEGMENT: manual segmentation into several sentences if needed.
7. TYPO: alternative annotation for ill formed words that happen to form contextually implausible valid words.
8. MHEAD: Additional annotation for cases of missing head word.
9. AMBIGUITY: marking for ambiguous structures that can be annotated differently.

10. IND WORD UPOS POS HIND REL: column names.

## 2.2 POS and Dependency Annotation

The annotation format is based on the CoNLL dependency annotations: one line per word with the following columns.

1. IND: index of the word in the sentence
2. WORD: the word
3. UPOS: Universal Part of Speech Tag
4. POS: Penn Treebank Part of Speech Tag (annotation guidelines)
5. HIND: index of the head word
6. REL: relation label to head word according to the English relations (NOT the universal relations).

## 2.3 Corrected Version

The autofilled corrected version is generated by copying over the relevant annotations from the original version and updating the corresponding word indices.

Approve an existing word annotation by removing the extra # in the last column, or change the annotation as needed. Fill in the annotation from scratch for words that do not appear (or whose head word doesn't appear) in the original version.

## 3 ESL Specific Guidelines

### 3.1 Instructions

#### Literal Annotation

Annotate according to the *literal* reading of the sentence rather than the corrections or other interpretations of potential intended meanings.

- **Argument Structure**

- **Missing prepositions** annotate as if the omission was intended. This may affect the labeling of the verb dependents (in particular tagging nmod as dobj/iobj).

Example:

12	I	PRON	PRP	13	nsubj
13	have	VERB	VBP	2	conj
14	to	PART	TO	15	mark
15	ask	VERB	VB	13	xcomp
16	you	PRON	PRP	15	<b>iobj #not dobj</b>
17	the	DET	DT	18	det
18	money	NOUN	NN	15	<b>dobj #not nmod</b>
19	of	ADP	IN	21	case
20	the	DET	DT	21	det
21	tickets	NOUN	NNS	18	nmod
22	back	ADV	RB	15	advmod
23	.	PUNCT	PUNCT	2	punct

Similarly, in “Therefore, I am interested in staying a log cabin” cabin should be dobj (rather than nmod) of staying.

- **Unnecessary prepositions** Annotate all nominal dependents introduced by prepositions as nmod, even if the inclusion of the preposition is an error. In the example below, the presence of the extraneous preposition “to” on line 25 means that “him” on line 26 should be marked as an nmod, rather than an iobj.

21	I	PRON	PRP	22	nsubj
22	had	VERB	VBD	5	parataxis
23	to	PART	TO	24	mark
24	give	VERB	VB	22	xcomp
25	to	ADP	IN	26	case
26	him	PRON	PRP	24	<b>nmod #not iobj</b>
27	water	NOUN	NN	24	dobj

Likewise, because of the extraneous preposition “in” on line 11, “summer” on line 13 should be marked as nmod, rather than nmod:tmod.

4	people	NOUN	NNS	2	nmod
5	will	AUX	MD	7	aux
6	be	AUX	VB	7	auxpass
7	invited	VERB	VBD	0	root
8	at	ADP	IN	10	case
9	your	DET	PRP\$	10	nmod:poss
10	camp	NOUN	NN	7	nmod
11	in	ADP	IN	13	case
12	this	DET	DT	13	det
13	summer	NOUN	NN	10	<b>nmod #not nmod:tmod</b>

- **Wrong verb tense** Annotate according to the used tense.

Example:

1	For	ADP	IN	3	case
2	next	ADJ	JJ	3	amod
3	year	NOUN	NN	5	nmod
4	we	PRON	PRP	5	nsubj
5	<b>search</b>	VERB	<b>VBP</b>	0	root
6	for	ADP	IN	9	case
7	a	DET	DT	9	det
8	new	ADJ	JJ	9	amod
9	barwoman	NOUN	NN	5	nmod

10 . PUNCT PUNCT 5 punct

If the tense is ambiguous, use the base form VB over VBP.

Example:

1	How	ADV	WRB	4	advmod
2	has	AUX	VBZ	4	aux
3	it	PRON	PRP	4	nsubj
4	<b>change</b>	VERB	<b>VB</b>	0	root
5	my	DET	PRP\$	7	nmod:poss
6	daily	ADJ	JJ	7	amod
7	life	NOUN	NN	4	dobj
8	?	PUNCT	PUNCT	4	punct

- **Missing head** annotate the most plausible head word in the original sentence. Use the #MHEAD metadata field to mark the word along with its HIND and REL in the corrected sentence.
- **goeswith** when both subparts of *goeswith* are interpretable (with corresponding PTB post tags), **and** the head word retains the POS in the combined word, the POS tagging of the dependent is NOT annotated with X GW.

Examples:

- them PRP self NN/goeswith
- my PRP\$ self NN/goeswith

Otherwise annotate with the usual X GW

- to X/GW day NN
- may X/GW be ADV
- make NN up X/GW
- be X/GW cause NN

## Clarifications and Exceptions

Follow the instructions below for cases in which a literal reading is either impossible or contextually implausible.

1. **Spelling** spelling errors are annotated according to the correctly spelled version of the word. Verify that the error is categorized as a spelling error (S or SX) in the FCE error annotation. In cases where the misspelled word happens to form a valid word mark it in the #TYPO metadata field.
2. **Uninterpretable word formations** errors with word forms that have no existing PTB tag are annotated according to the correction. Otherwise (and when in doubt) annotate according to the literal reading of the word/phrase.

Examples:

- (a) IV:Incorrect verb formation:

```
<ns type="IV"><i>mentionned</i><c>mentioned</c></ns>
```

- (b) IJ:Incorrect adjective formation:

```
<ns type="IJ"><i>funier</i><c>funnier</c></ns>
```

Note: don't forget to add the `ua=true` attribute when adjectives "agree" with plural nouns.

- (c) IN:Incorrect noun formation:

```
<ns type="IN"><i>lifes</i><c>lives</c></ns>
```

- (d) IY:Incorrect adverb formation:

Not applicable, happens only four times in FCE.

- (e) DJ:Wrongly derived adjective:

```
<ns type="DJ"><i>unforgettable</i><c>unforgettable</c></ns>
```

Annotate literally cases that can be assigned with standard PTB tags and are contextually plausible:

- ```
<ns type="IJ"><i>more fast</i><c>faster</c></ns>
```
- ```
<ns type="IN"><i>staffs</i><c>staff</c></ns>
```

3. **Contextually implausible word formations** cases of erroneous word formations that happen to form valid words that are contextually implausible are annotated



according to the correction. Similarly to spelling errors, these cases are marked in the #TYPO metadata field with their literal annotation.

Example:

```
#SENT=...a <ns type="DJ"><i>disappoint</i><c>disappointing</c></ns>
<ns type="MP"> <c>,</c></ns> unknown actor...
#TYPO=21 VERB VB
```

20	a	DET	DT	23	det
21	disappoint	ADJ	JJ	23	amod
22	unknown	ADJ	JJ	23	amod
23	actor	NOUN	NN	24	nsubj

Similarly, in the following case:

```
<ns type="IV"><i>mad</i><c>made</c></ns> the concert absolutely perfect
```

the annotation should be according to “made”, while mad should be annotated in the #TYPO field.

## 3.2 ESL Metadata

### #SENT

Contains the sentence with the manual error annotation, as it appears in the FCE. The FCE error tags can be found in Appendix B. Additional documentation of the error annotation can be found here [3].

- **Adding New Errors** If an error is missing in the annotation, add this error to the SENT field in the original and corrected files, with an attribute new=true (i.e. <ns type=ERRTYPE new=true> ... </ns>) to mark that its our addition. Update the corrected sentence if needed.

Example:

```
#SENT=All kinds of shopping have more and more become <ns type="UT">
<i>to</i></ns> a <ns type="W"><i>rather stressful activity</i><c>stressful
activity rather</c> </ns> than an enjoyable <ns type="S" new=true> <i>on</i>
<c>one</c></ns>.
```

- **Changing Errors**

Errors of the types "CE" (compound error), "M" (missing) and "R" (replace), as well as more specific errors such as "RN", "RV" sometimes appear without a correction. When this is the case, and the error is not purely semantic (e.g. not correcting the error results in a syntactically ungrammatical sentence), add an error correction annotation. Try to make the correction "minimal", i.e. a minimal change that would make the sentence grammatical. Adding the correction will typically mean adding `<i></i>` and/or `<c></c>` and providing a corrected string. As these cases are not always well annotated, sometimes you will need to change the scope of the error tag and potentially change the error type. In all cases mark `new=true` in the error tag.

Examples:

```
#SENT=I must admit that <ns type="UP"><i>,</i></ns> that evening wasn't  
the perfect one <ns type="UY"><i>out</i></ns> for me <ns type="CE">but  
I and for my friends either</ns>.
```

Can be changed to:

```
#SENT=I must admit that <ns type="UP"><i>,</i></ns> that evening wasn't  
the perfect one <ns type="UY"><i>out</i></ns> for me <ns type="R" new=true  
><i>but I and for my friends either</i><c>and neither for my friends</c></ns>.
```

```
#SENT=I would also like to ask <ns type="M"></ns> accommodation you offer  
Would change to:
```

```
#SENT=I would also like to ask <ns type="M" new=true><c>about the</c></ns>  
accommodation you offer
```

```
#SENT=In the future <ns type="CE">they both getting closer and closer</ns>.
```

Would change to:

```
#SENT=In the future they <ns type="R" new=true><i> both getting</i><c>will  
both get<c></ns> closer and closer.
```

Cases like:

```
#SENT=so I wanted to spend my last evening <ns type="R">with great pleasure</ns>  
where the sentence is semantically awkward, but is syntactically correct, can be kept  
as is.
```

**Copy the new SENT to the corrected version** and update the annotation in the corrected version to reflect the error correction.

**Importantly** - the original sentences shouldn't change.

Add the following attributes to the "<ns>" tag when applicable.

- **Discourse error** mark words that are erroneous only given the context of the sentence with `disc=true`

Example:

```
I <ns type="TV" disc=true><i>was</i><c>am</c></ns> happy
```

- **Unnecessary agreement** mark unnecessary agreement by adding `ua=true` to the relevant error.

Example:

```
<ns type="IJ" ua=true><i>interestings</i><c>interesting</c>  
</ns> things
```

Do NOT change the error annotation in any other way.

## #TYPO

Fill in the typical UPOS and POS, as well as the corresponding HIND and REL for misspelled and contextually implausible ill formed words in cases where the original word happens to form a valid word. In the annotation itself annotate according to the corrected word form (as indicated in the error annotation). See section 3.1 for more details.

**Format** #TYPO=*IND UPOS POS HIND REL*,...

where IND is the word index. Multiple words are separated by commas (no space after the comma).

Example:

```
#TYPO=15 NOUN NN 11 conj
```

11	breakfast	NOUN	NN	9	dobj
12	,	PUNCT	PUNCT	11	punct
13	lunch	NOUN	NN	11	list
14	and	CONJ	CC	11	cc
15	diner	NOUN	NN	11	conj

## #MHEAD

Mark for words whose most appropriate head is not in the original sentence. See section 3.1 for a general description.

### Format #MHEAD=*IND CHIND CREL*

where *IND* is the index of the word, *CHIND* is the index of the head word in the correction file, and *CREL* is the corresponding relation in the correction. Separate several words with commas (no extra spaces).

In the example below, the tokens “That”, “why” and “.” are missing their head word “is”, which should also be the root of the sentence. To address this structure, we head “That” with the sentence root, and make it the head of “is” and “.” in place of “is”. In the #MHEAD field, we set the head indices of “That”, “why” and “.” to be the index of “is” in the corrected sentence, 2.

```
#SENT=That <ns type="MV"><c>is</c></ns> why I <ns type="TV"><i>couldn't have  
swimmed</i><c>can't swim</c></ns> yet.
```

```
#MHEAD=1 2 nsubj, 7 2 ccomp, 9 2 punct
```

#IND	WORD	UPOS	POS	HIND	REL
1	<b>That</b>	DET	DT	<b>0</b>	root
2	why	ADV	WRB	7	advmod
3	I	PRON	PRP	7	nsubj
4	could	AUX	MD	7	aux
5	n't	PART	RB	7	neg
6	have	AUX	VB	7	aux
7	<b>swimmed</b>	VERB	VBN	<b>1</b>	ccomp
8	yet	ADV	RB	7	advmod
9	.	PUNCT	PUNCT	<b>1</b>	punct

In the next example, 20 is the head index of “of” in the correction file.

```
#MHEAD=17 20 case
```

13	your	DET	PRP\$	14	nmod:poss
14	credibility	NOUN	NN	16	nsubj
15	is	AUX	VBZ	16	aux
16	depending	VERB	VBG	6	parataxis

17	of	ADP	IN	16	dobj
18	.	PUNCT	PUNCT	3	punct

## 4 Clarifications for English UD Guidelines

### 4.1 General Clarifications

- **Determiners without a noun head** (e.g. “which”, “that”) are still tagged as determiners although they act like nouns/pronouns in regards to their relations.

Example: in “First of all”, “all” is tagged as a determiner, and it’s relation is nmod headed by “First”.

Similarly, when there is an adjective that acts as a noun phrase, like “The very poor have a hard time”, it should be tagged as an adjective but its relation should be the same as if it were a noun, so in this example, it is an nsubj of “have”.

- **Copula “is” with a clausal complement** when “is” acts as a copula and the complement is a clause (for example, “The goal is to try harder next time”), “is” should be the root of the sentence and the complement should be xcomp or ccomp as appropriate.

- **Coordinating conjunctions** when using coordinating conjunctions, annotate keeping nested structure in mind.

For example, a sentence such as A but B and C that is parsed as A but (B and C) should be annotated as follows:

1	A	NOUN	NN	0	root
2	but	CONJ	CC	1	cc
3	B	NOUN	NN	1	conj
4	and	CONJ	CC	3	cc
5	C	NOUN	NN	3	conj

- **Expletives** Use expl and csubj for templates such as: It would be X to VERB Y.

Example:

7	it	PRON	PRP	10	expl
8	would	AUX	MD	10	aux
9	be	VERB	VB	10	cop
10	interesting	ADJ	JJ	6	ccomp
11	to	PART	TO	12	mark
12	vistit	VERB	VB	10	csubj

- **Titles and Names** Use PROPN NNP compound for titles such as Dr. or Mr. when followed by a name.

Example:

18	killed	VERB	VBN	15	xcomp
19	by	VERB	IN	21	case
20	Mr	PROPN	NNP	21	compound
21	Brown	PROPN	NNP	18	nmod
22	after	ADP	IN	23	mark

- **nmods** If a noun phrase has no case attached, is adverbial in nature, and is temporal, it should be labeled nmod:tmod. Otherwise, if it has no case attached and is adverbial in nature, it should be labeled nmod:npmo.

Example:

1	Finish	VERB	VB	0	root
2	your	DET	PRP\$	3	nmod:poss
3	homework	NOUN	NN	1	dobj
4	next	ADJ	JJ	5	amod
5	week	NOUN	NN	1	nmod:tmod

- **Punctuation** attach punctuation as high hierarchically as will not violate projectivity, i.e. if the sentence is PhraseA , PhraseB, and the head of PhraseA is in PhraseB, then the comma should be headed by the same word as PhraseA.

For sentences that need to be segmented, the final punctuation mark doesn't attach to the root, but to the head with the parataxis relation label.

## 4.2 UD EWT Treebank and UD Guidelines Differences

In several cases the UD English treebank annotation differs from the UD guidelines. In the ESL treebank we **always follow the guidelines**.

- **NAMES** UD guidelines state that names should use the relation “name” with the

head as the first word in the name. The EWT uses the relation “compound” and has the head at the end (which is consistent with how “compound” should be formed in general). We follow the guidelines and use “name”.

Example:

11	Danny	PROPN	NNP	6	ccomp
12	Brook	PROPN	NNP	11	name
13	and	CONJ	CC	11	cc
14	Tina	PROPN	NNP	11	conj
15	Truelove	PROPN	NNP	14	name

- **POSSESSIVE PRONOUNS** UD guidelines say possessive pronouns like “his book” should be tagged as determiners, so POS=PRP\$ should have UPOS=DET. However, the training data uses UPOS=PRON. We follow the guidelines and use UPOS=DET.

Example:

9	with	ADP	IN	10	case
10	two	NUM	CD	8	nmod
11	of	ADP	IN	13	case
12	my	DET	PRP\$	13	nmod:poss
13	friends	NOUN	NNS	10	nmod
14	your	DET	PRP\$	15	nmod:poss
15	musical	ADJ	JJ	16	amod
16	show	NOUN	NN	8	dobj

- **EXPLETIVES** The training data doesn’t tag expletive “it” as REL=expl, it instead uses REL=nsubj.

For example: “It was said that we should leave”: the EWT annotates “it” = nsubj and “leave” = ccomp, while the correct annotation is “it” = expl and “leave” = csubjpass.



## 5 Other

### 5.1 Sentence Segmentation

In some cases the sentence tokenizer under-segments sentences. When you encounter this situation, fill in the `#SEGMENT` field.

**Format** `#SEGMENT=IND ROOTIND, ...`

Where *IND* is the index of the first word in the new sentence, and *ROOTIND* is the index of the root word in the new sentence. Multiple segmentations are separated by commas (no space after the comma).

Example:

`#SEGMENT=3 7`

1	Your	DET	PRP\$	0	root
2	sincerly	ADV	RB	1	advmod
3	Madrid	PROPN	NNP	7	nsubj
4	is	VERB	VBZ	7	cop
5	a	DET	DT	7	det
6	big	ADJ	JJ	7	amod
7	city	NOUN	NN	1	parataxis

### 5.2 Capitalization

If a sentence is in all-caps, convert the words in the annotation (not in the metadata) to their proper capitalization. Example “I LIKE EATING CUPCAKES WHILE ANNOTATING” should be changed to “I like eating cupcakes while annotating.”.

## 6 Quality Assurance

### 6.1 Review Protocol

Each sentence annotation (both original and corrected) is reviewed by another annotator (“reviewer”). Review sentences are randomized, and the review process is double blind.

In the original file, a # sign appears at the last column of each line. Approve the word annotation by deleting the #.

If you would have annotated the word differently, add the annotation in the following format

*UPOS POS HIND REL*

Use \* instead of the token correction for items that remain the same.

For example:

```
1    big    ADJ    JJ    2    amod    #
2    city   VERB   NNS   5    nsubj   #
```

Should be corrected as:

```
1    big    ADJ    JJ    2    amod
2    city   VERB   NNS   5    nsubj  #NOUN NN * *
```

Make sure to add all the relevant changes in the original sentence to the **corrected sentence** using the same format.

In addition to reviewing the annotation itself, correct the format of the **metadata** if needed (for the metadata, change it directly, no need to do any special marking).

### 6.2 Resolution Protocol

#### General

In the resolution process, you take the role of a “judge” and resolve all the disagreements between annotators and reviewers. Additionally, you add reviews for annotation errors that were missed by the reviewer.

For each disagreement token, resolve the review in an additional 8th column using + to approve review and - to reject review (i.e. approve annotation).

For example, in this case the judge accepts the review for the word “city”:

1	big	ADJ	JJ	2	amod				
2	city	VERB	NNS	5	nsubj	#NOUN	NN	* * +	

For partial approval mark +/- four times separated by spaces. For example, if you reject the HIND in the review, but approve everything else, mark + + - +. If you have a third opinion that is neither the annotation nor the review, specify it instead of +/- . For example, if you reject the review and think HIND (that is marked 9 in the annotation and 11 in the review) should be 10, mark - - 10 -.

### Adding Reviews

If you encounter an issue that was missed by the reviewer, add a review in the 7th column (in the regular UPOS POS HIND REL format), starting with @ instead of #, and approve your own review with a + in the 8th column. If you notice a plausible alternative annotation for an ambiguous structure, add it even if you think that the existing annotation is better, reject your own review in the 8th column, and add a marking to the #AMBIGUITY as explained below.

### Marking Ambiguities

In the role of a judge, mark alternative word annotations for structural ambiguities which arise from annotator-reviewer disagreements or otherwise present in the sentence. Mark the alternative annotation to the one that results from applying the resolution decision in the #AMBIGUITY field (the alternative can be the annotation, the review, a combination of the two, or something else). Mark both “standard” English ambiguities (such as PP attachment) as well as cases with conflicting cues in grammatical errors, where several potential literal readings are possible.

**Format** #AMBIGUITY=*IND UPOS POS HIND REL*,...

where IND is the word index. Multiple words are separated by commas (no space after the comma).

Example (PP attachment ambiguity):

```
#SENT=What kind of people will be invited <ns type="RT"><i>at</i><c>to</c>
</ns> your Camp <ns type="UT"><i>in</i></ns> this summer?
```

```
#AMBIGUITY=13 NOUN NN 10 nmod
```

1	What	DET	WDT	2	det
2	kind	NOUN	NN	7	nsubjpass
3	of	ADP	IN	4	case
4	people	NOUN	NNS	2	nmod
5	will	AUX	MD	7	auxpass
6	be	VERB	VB	7	cop
7	invited	VERB	VBD	0	root
8	at	ADP	IN	10	case
9	your	DET	PRP\$	10	nmod:poss
10	Camp	NOUN	NN	7	nmod
11	in	ADP	IN	13	case
12	this	DET	DT	13	det
13	summer	NOUN	NN	7	nmod
14	?	PUNCT	.	7	punct

## 7 Annotation Utilities

Download the annotation utilities. Get announced updates by typing “git pull”.

Use `training_data_searcher.py` to see how various patterns are annotated in the UD EWT English training data.

Run `anno_checker.py` before submitting annotation files.

Run `metadata_checker.py` and `review_checker.py` before submitting reviews.

Run `resolution_checker.py` before submitting resolution files.

## 8 Appendix A - Common Annotation Errors

A list of common annotation errors. Watch out for these errors during the review process.

1. **SYM** make sure that symbols are marked as such.
2. **possessive nouns** make sure the ' is labeled case.
3. **compound:prt** use this relation when appropriate.
4. **so** may have been annotated as CONJ CC # cc instead of ADV RB # advmod. If you can replace it with “so that”, then its SCONJ.
5. **so that** should be SCONJ and SCONJ with “that” as mwe.
6. “it” expletives are PRON PRP (not PRON EX) and “there” expletives are PRON EX.
7. “it is/was said/reported/written that X”, “it” is *expl* (not nsubj/nsubjpass) and head of X is *csubjpass* (not ccomp).
8. **out of** is NOT case mwe, it is case case with both headed by the same word.
9. **go** or **do** when followed by a gerund: the gerund should almost always be a VERB VBG, not NOUN NN.
10. **even though** is not mark mwe, it is **advmod mark**.
11. **looking forward to VERB** verb relation should be **advcl**, not xcomp.
12. **which** as in “I ate some cake, which turned out to be a terrible idea”, which should be **advcl**.
13. **PROPN** for all proper nouns. Adjectives that are part of a proper noun should also be tagged PROPN.
14. **Organization names** (UD guidelines reminder) “when organization names have a clear syntactic modification structure, the dependencies should reflect that structure using regular syntactic relations”.

Examples:

1	Camp	PROPN	NNP	2	compound
2	California	PROPN	NNP	0	root

1	The	DET	DT	4	det
2	International	PROPN	NNP	4	amod
3	Arts	PROPN	NNPS	4	compound
4	Festival	PROPN	NNP	0	root

1	Ministry	PROPN	NNP	0	root
2	of	ADP	IN	4	case
3	the	DET	DT	4	det
4	Interior	PROPN	NNP	1	nmod

15. If there is an auxpass there should be an nsubjpass nearby (and vice versa).
16. **superlatives and comparatives** should be marked JJR, JJS, RBR or RBS.
17. **called/become** both take xcomp.
18. **“never” and “not”** are neg, not advmod.
19. **“no”** is a a neg, not a det.
20. **“closed”** when modifying a noun is ADJ.
21. **“that”** should not be an expletive.
22. **“what”** introducing dependent clauses, goes with the lower clause.
23. words that introduce verbs may have been marked as case instead of mark.
24. subordinating conjunctions should be SCONJ not ADP.
25. **mwe’s** make sure all the expressions in the **mwe list** are marked as such.

## 9 Appendix B - FCE Error Tags

'AGN': 'Noun agreement', 'AGREEMENT'  
'AGV': 'Verb agreement', 'AGREEMENT'  
'CL': 'Collocation error', 'STYLE'  
'CN': 'Noun countability error', 'NOUN'  
'ID': 'Incorrect formation determiner', 'DET'  
'IV': 'Incorrect formation verb', 'VERB'  
'MA': 'Missing pronoun', 'PRONOUN'  
'MD': 'Missing determiner', 'DET'  
'MT': 'Missing preposition', 'PREP'  
'MV': 'Missing verb', 'VERB'  
'RA': 'Replace pronoun', 'PRONOUN'  
'RD': 'Replace determiner', 'DET'  
'RJ': 'Replace Adjective', 'ADJ'  
'RN': 'Replace noun', 'NOUN'  
'RT': 'Replace preposition', 'PREP'  
'RV': 'Replace verb', 'VERB'  
'RY': 'Replace adverb', 'ADV'  
'UA': 'Unnecessary pronoun', 'PRONOUN'  
'UD': 'Unnecessary determiner', 'DET'  
'UT': 'Unnecessary preposition', 'PREP'  
'UV': 'Unnecessary Verb', 'VERB'  
'TV': 'Tense verb', 'VERB'  
'AGA': 'Pronoun (anaphoric) agreement', 'AGREEMENT'  
'AGD': 'Determiner agreement', 'AGREEMENT'  
'AGQ': 'Quantifier agreement', 'AGREEMENT'  
'AS': 'Incorrect argument structure', 'VERB'  
'CD': 'Wrong determiner because of noun countability', 'NOUN'  
'CE': 'Compound error', 'NOUN'  
'CQ': 'Wrong quantifier because of noun countability', 'NOUN'  
'DA': 'Wrongly derived pronoun', 'PRONOUN'  
'DD': 'Wrongly derived determiner', 'DET'  
'DV': 'Wrongly derived verb', 'VERB'  
'FD': 'Wrong form determiner', 'DET'

'FJ': 'Wrong form adjective', 'ADJ'  
 'FY': 'Wrong form adverb', 'ADV'  
 'W': 'Word order', 'WO'  
 'X': 'Incorrect formation of negation', 'NEG'  
 'DJ': 'Wrongly derived adjective', 'ADJ'  
 'DN': 'Wrongly derived noun', 'NOUN'  
 'DY': 'Wrongly derived adverb', 'ADV'  
 'FN': 'Wrong form noun', 'NOUN'  
 'FV': 'Wrong form verb', 'VERB'  
 'IJ': 'Incorrect formation adjective', 'ADJ'  
 'IN': 'Incorrect formation noun', 'NOUN'  
 'IQ': 'Incorrect formation quantifier', 'QUANT'  
 'L': 'Inappropriate register', 'COLLOCATION'  
 'MC': 'Missing conjunction', 'CONJ'  
 'MJ': 'Missing adjective', 'ADJ'  
 'MN': 'Missing noun', 'NOUN'  
 'MQ': 'Missing quantifier', 'QUANT'  
 'MY': 'Missing adverb', 'ADV'  
 'RC': 'Replace conjunction', 'CONJ'  
 'RQ': 'Replace quantifier', 'QUANT'  
 'UC': 'Unnecessary conjunction', 'CONJ'  
 'UJ': 'Unnecessary adjective', 'ADJ'  
 'UN': 'Unnecessary noun', 'NOUN'  
 'UQ': 'Unnecessary quantifier', 'QUANT'  
 'UY': 'Unnecessary adverb', 'ADV'  
 'AG': 'Argeement', 'AGREEMENT'  
 'FQ': 'Wrong form quantifier', 'QUANT'  
 'DQ': 'Wrongly derived quantifier', 'QUANT'  
 'IY': 'Incorrect formation adverb', 'ADV'  
 'FA': 'Wrong form pronoun', 'PRONOUN'  
 'DI': '?', 'OTHER'  
 'IA': 'Incorrect formation pronoun', 'PRONOUN'  
 'DT': 'Wrongly derived preposition', 'PREP'  
 'UP': 'Unnecessary punctuation', 'PUNCT'  
 'MP': 'Missing punctuation', 'PUNCT'



'RP': 'Replace punctuation', 'PUNCT'  
'S': 'Spelling', 'SPELLING'  
'SA': 'American spelling', 'SPELLING'  
'SX': 'Spelling confusion error', 'SPELLING'  
'M': 'Something missing', 'OTHER'  
'R': 'Replace', 'OTHER'  
'U': 'Unnecessary', 'OTHER'

## References

- [1] Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. Universal dependencies for learner english. *arXiv preprint*, 2016.
- [2] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics, 2011.
- [3] Diane Nicholls. The cambridge learner corpus: Error coding and analysis for lexicography and elt. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 572–581, 2003.