Error annotation in adjective–noun (AN) combinations

This document describes the annotation scheme devised for annotating errors in AN combinations and explains how the inter-annotator agreement has been estimated.

1 Data

We present and release a dataset of AN combinations which, on the one hand, exemplify the typical errors committed by language learners in the choice of content words within such combinations, and, on the other hand, are challenging for an EDC system.

To satisfy the first condition, we examined the publicly available CLC-FCE dataset (Yannakoudakis *et al.*, 2011), used the error annotation in this dataset (Nicholls, 2003), and analysed the typical errors in AN combinations committed by language learners. We have compiled a list of 61 adjectives that are most problematic for learners (e.g., *big*, *large* and *great* that are frequently confused with each other, *classic* and *classical*, *economic* and *economical*, etc.).

Using this set of 61 adjectives, we have extracted AN combinations from the unannotated part of the Cambridge Learner Corpus (CLC),[1] a large corpus of texts produced by English language learners, sitting Cambridge Assessment's examinations.[2] We have focused on AN combinations previously unseen in a native English corpus, as we hypothesise that they would have a higher chance of containing an error. Such combinations are more challenging for EDC algorithms. In our research, we used the British National Corpus (BNC)[3] to select the corpus-unattested combinations.

We have compiled a set of 798 AN combinations. An annotation scheme has been devised to annotate these examples as correct or incorrect, and for the incorrect combinations, to identify the locus of error (adjective, noun or both) and the type of confusion (incorrect synonym, form-related word, or non-related word). The most appropriate corrections are included in the dataset.

Misspellings: Even though the nouns in the collected AN combinations have been checked against nouns extracted from the native corpus, some misspellings are still present in the data. The following approach has been taken to deal with this issue: an AN combination containing a misspelled noun is removed from the list of candidates if the same combination with the correctly spelled noun is found in the native corpus, otherwise, the combination is kept in the list.

Our observations on the data reveal that misspellings can be grouped under the following two types: misspellings resulting in a non-word or word of a different PoS (e.g., **looser* for *loser*), and misspellings resulting in a different word of the same PoS (e.g., **principal* for *principle*). We have adopted the following approach to these cases: in the first case, a combination should be treated and annotated ignoring the fact that the words are misspelled (i.e., annotating the ANs as if the words are spelled correctly). The corrections for the misspelled words may optionally be provided in the correction field. In the second case, however, we suggest using the tag for the form-related confusions (tag `F`, as explained below).

[1] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/

[2] http://www.cambridgeenglish.org

[3] http://www.natcorp.ox.ac.uk/

Alternative spelling: In case British vs. American spelling differs for a word (e.g., *centre* vs. *center*), AN combinations in both spelling variations are checked against native corpora to make sure they do not appear there.

2 Annotating combinations out of context

An AN combination in the dataset is first annotated on its own, without any reference to the context where it occurs. Since the focus of this research is on finding anomalies in content word combinations, we try to avoid annotating contexts rather than AN combinations by first annotating the combinations in isolation.

At this level of annotation, one of only two tags should be used:

- **`C`** – an AN is correct
- **`I`** – an AN is incorrect

When running the annotation script, the annotator is first prompted to annotate the AN with one of the two tags. Only after that, a set of contexts is presented for further annotation. The annotation tool is designed so that the upper level tag (`C` or `I`) is automatically added to the first field in every context.

3 Annotating combinations in context

We distinguish between phrase-level and word-level tags. Phrase-level tags describe an AN combination as a whole, whereas word-level tags are related to particular words within a combination (i.e., an adjective or a noun).

3.1 Phrase-level tags

Tags `C` and `I` provided by the annotator when annotating an AN combination in isolation are phrase-level tags. In addition, when turning to context, an annotator might realise that no further annotation for the combination can be provided, because:

- the context as a whole does not make any clear sense or is too ungrammatical to get the meaning. The learner's communicative intent is *unclear* or too hard to guess. In that case, tag **`U`** should be added to the very first field, after the already provided tag `C` or `I` (resulting in `CU` or `IU`). No correction is needed in that case;
- the context reveals that the only way the combination can be corrected makes it inappropriate for this annotation, and the combination should be *excluded* from the set. The most typical cases for this tag to be used include the ones when correction requires insertion of some words between the adjective and the noun, when a parser error has occurred and the combination is not an AN (i.e., the given adjective is not related to the given noun with a grammatical relation), and when correction involves a change in the PoS of the word(s) and results in a combination which is not an AN combination anymore. In that case, tag **`O`** should be used. Since the use of this tag means that the combination should be excluded from the annotation set, the tag should

only be used when the annotator considers the combination inapproapriate for annotation. The tag should be added to the very first field, after the already provided tag `C` or `I` (resulting in `CO` or `IO`). In addition, a comment specifying the reasons for choosing this tag should be provided in the same field in brackets after the tag. Correction is not required.

3.2 Word-level tags

Word-level tags are to be added to the individual words (adjective, noun, or both), depending on the locus of an error, in the provided annotation fields. Correction(s) are to be provided in the correspondent correction field(s). An annotator should make choice between three tags:

- **`F`** should be used when a correction (a supposedly intended word) is close to the one used by the learner in form;
- **`S`** should be used if a correction is semantically related to the used word. To minimise the risk of different interpretations by different annotators, we limit the set of possible semantic relations to synonyms and hyper-/hyponyms. In addition, when looking for possible synonyms and hyper-/hyponyms, we use an on-line thesaurus `thesaurus.com`.
- **`N`** should be used when the only possible correction is not close to the used word neither semantically, nor in form.

4 Corrections

Corrections should be provided in the special fields for the adjective, noun, or both, depending on the locus of the error. The provided corrections should correspond in type to the tag chosen for the word. If multiple corrections are possible, they can all be listed in the correction field separated by " / ", but only in case when they are all of the same type (e.g., a set of synonyms). In case, an AN can be corrected in several ways and the corrections are of different types, an annotation procedure defined in the next section should be followed, and the choice between the possible corrections is to be made accordingly.

5 Annotation procedure

This section specifies certain order for how the tags should be chosen. Presumably, this would help to standardise the annotation procedure and improve the inter-annotator agreement.

- **Step 1**: First of all, a choice is to be made between `C` for the correct and `I` for the incorrect combinations. Since at this point the combination is presented in isolation, its interpretation is not dependent on any specific context. Therefore, if an annotator can think of a possible interpretation under which the combination is semantically acceptable, it should be tagged with `C`. If an interpretation is hard/impossible to arrive at, the AN should be tagged with `I`. Additionally, an AN should be tagged as incorrect, if a semantically acceptable interpretation is too hard to arrive at, even if it does exist when one of the two words is taken in some rare meaning (e.g., domain-specific one). The motivation for that is that we do not want to include particularly

rare language phenomena/interpretations, and chances that a learner knows about these specific interpretations are low. In case an annotator would like to change their decision after looking at the context (e.g., realising that some interpretation is indeed plausible), the context annotation window can be simply closed without saving the annotation. After that, the annotator can start annotating the same combination again.

- **Step 2**: After the initial annotation with `C` or `I` is provided in the first window, a context annotation window is opened where every context contains `C` or `I` in the first field. First, the combination is to be considered as a whole – if the context does not help in deciding whether it is used correctly or not, tag `U` is to be added (`CU` or `IU`), no correction required. If the context reveals that the combination is not appropriate for the annotation, tag `O` is to be added (`CO` or `IO`), a comment should be provided in the same field in brackets, no correction is required. At this point, the annotation for the AN can be finished.

- **Step 3**: If the context helps interpreting a combination as correctly used, and the annotator considered the combination to be correct in isolation, no further actions are to be taken (the combination would have tag `C` assigned to the first field, no correction is needed), and the annotation can be finished.

- **Step 4**: In case, an annotator considered a combination to be incorrect in isolation, or the context reveals that a generally correct combination (tag `C`) is incorrectly used, the first decision to be made is that about the locus of the error. If the locus of the error (adjective, noun, or both) is evident to the annotator, they should proceed to annotating the erroneous word(s). If there are several ways of correcting the AN, the following rules should be followed:

  - choose the minimum possible correction, i.e., correct as few words as possible;
  - given a choice of correcting an adjective or a noun, choose the adjective-related correction. The motivation for that is that the AN combinations were chosen, in the first place, so that they contain one of the most frequently misused adjectives rather than nouns.

- **Step 5**: Next, an appropriate tag and correction should be chosen for the erroneous word(s). At this point, the choice is to be made between three tags (`F`, `S` and `N`) for both generally correct (tag `C` in the first field) and generally incorrect (tag `I`) combinations. The tags / possible corrections are to be applied / looked for in the following order: `F>S>N` (i.e., tag `F` is preferred over tag `S`, and tag `S` is preferred over tag `N`). The motivation for that is as follows:

  - We suppose that learners rarely commit inexplicable errors and a word used is somehow related to the intended one in a learner's mental lexicon. Therefore, tag `N` (unrelated word) and the correspondent correction should be used only when every other interpretation has failed.
  - It can also be hypothesised that learners more easily confuse words related by form or similar pronunciation than those related semantically. The first type of confusion does not (necessarily) require any knowledge of the possible word meanings, whereas in the second case, to be able to confuse two semantically close words a learner should be aware of the possible meanings of both words. Therefore, given a choice between tag `F` and `S` with the correspondent corrections, an annotator should choose the first option.

6 Annotation tool

To annotate the data, we have used a special annotation tool. First, an AN combination is presented in a separate window, and an annotator is asked to press C if the combination is correct, or I otherwise. If C or I is pressed, the context annotation window is opened. In case the initial annotation is wrong (an annotator would like to change the initially provided tag to a different one, e.g., I to C), the context annotation window should be closed, and the annotator should be able to open and reannotate the same combination.

Tag C or I provided by the annotator in the first window is automatically added to the first field in the context annotations. In case the annotator would like to use any other phrase-level tags (e.g., U or O), these should be added to the same field as C or I. Word-level tags (F, S or N) and corrections should be provided in the correspondent adjective or noun-related fields.

The final annotation is stored in the format: [general level annotation (C/I/CU/IU/CO/IO)] - [adjective-related annotation (J/JF/JS/JN)] - [noun-related annotation (N/NF/NS/NN)].

7 Format of the XML file

The annotated dataset is stored in an XML file. Each annotated AN is stored as a separate node <AN/> with the attributes *id* storing a unique AN identifier, and *lem* storing the lemma representation of the AN. The second part of the *id* separated by a colon (e.g., *1:0* and *1:1*) points to the multiple entries with the same lemma.

The annotation is provided in the child node <annotation/> with the three fields of the annotation pointing to the out-of-context annotation, adjective-related and noun-related annotation: for example, C-J-N is used for ANs correct both out of context and in context, while C-JS-N is used for ANs correct out of context but incorrectly used in context, where the correction would require changing the adjective for a semantically related one.

Corrections are provided in the text field of the same node in square brackets. Symbol = is used in cases where the originally used word is correct.

An optional child node <comment/> is used when additional explanation is provided, for example, if an annotator used tag O.

The <context/> field contains the context of use extracted from the CLC. Additional tags are used to point to the AN: <e/> for the annotation with the attribute *t* for the tag, <c/> for corrections and correction-related comments, and <i/> for the originally used adjectives and nouns.

8 Inter-annotator agreement

The dataset was primarily annotated by a professional linguist. However, to ensure that the annotation scheme is clear and efficient, the dataset was split into 100 and 698 ANs, and the 100 ANs were first annotated by the same professional annotator and three other annotators. We have measured the

observed inter-annotator agreement and estimated the Cohen's *kappa* (κ) for each pair of annotators. The κ is estimated as follows:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{1}$$

where $Pr(a)$ is the observed agreement between the annotators, and $Pr(e)$ is the probability of agreement expected by chance. $Pr(e)$ is estimated using the observed annotation to calculate the probabilities of each annotator randomly selecting each category. $\kappa = 1$ when two annotators are in complete agreement (i.e., $Pr(a) = 1$), and $\kappa = 0$ when there is no agreement other than what would be expected by chance (i.e., $Pr(a) = Pr(e)$).

We use the mean values of κ to check whether the annotation scheme is clear, and the mean values for the observed inter-annotator agreement to set the upper bound for our experiments.[4]

We measure agreement at the two annotation levels: for the *out-of-context* annotation we only consider the ANs annotated with the tags `C` for the correct and `I` for the incorrect ANs, whereas for the *in-context* annotation we distinguish between the `C-J-N` annotation for the ANs correct in context and all other combinations of tags (including `O` and `U`) for the ANs incorrect in context.

The observed inter-annotator agreement values within each pair of annotators are presented in Tables 1 and 2 for the *out-of-context* and *in-context* annotations, and the κ values – in Tables 3 and 4.

The mean κ values of 0.6500 and 0.4917 for the *out-of-context* and *in-context* annotation are interpreted as substantial and moderate/medium agreement between annotators and show that the annotation scheme can be effectively applied to annotate the dataset.

The mean values of 0.8650 and 0.7467 for the observed inter-annotator agreement are used as an upper bound in our experiments.

[4]We have also estimated the inter-annotator agreement for the two levels of annotation using Fleiss' *kappa* (Joseph L. Fleiss 1971. *Measuring nominal scale agreement among many raters*. Psychological Bulletin, Vol. 76, No. 5, pp. 378–382), which is used for measuring reliability of annotation for more than two annotators. However, we found that Cohen's *kappa* and observed agreement for pairs of annotators are easier to interpret.

Annotator	1	2	3	4
1	–	0.86	0.86	0.85
2		–	0.92	0.89
3			–	0.81
4				–

Table 1: Observed inter-annotator agreement, *out-of-context* annotation.

Annotator	1	2	3	4
1	–	0.76	0.72	0.74
2		–	0.78	0.76
3			–	0.72
4				–

Table 2: Observed inter-annotator agreement, *in-context* annotation.

Annotator	1	2	3	4
1	–	0.64	0.62	0.60
2		–	0.80	0.73
3			–	0.51
4				–

Table 3: *Kappa* values, *out-of-context* annotation.

Annotator	1	2	3	4
1	–	0.52	0.44	0.48
2		–	0.56	0.52
3			–	0.43
4				–

Table 4: *Kappa* values, *in-context* annotation.