Annotation Specification – L1ML

Current as of: Tuesday, March 8th, 2016
Updated versions will be posted on the L1ML Github. We will send out a message with any updates that L1ML adds to the specification. Please feel free to contact us at any time with any questions or concerns about the specification.
Stay tuned for an easier-to-read workflow checklist to use while annotating, as well as additional provided examples for each tag.

1) Overview of Project

The goal of the L1ML annotation scheme is to highlight salient grammatical errors in second language (L2) speakers of English's writing in order to train an algorithm that will be able to identify the writer's native language (L1). You, as the annotator, will be tasked with marking up these features using the Multimodal Annotation Environment in tandem with the provided Document Type Definition (.dtd) file (as of writing, this file is called L1ML_v1.0.dtd).

L1ML uses data from the TOEFL11 corpus. This corpus takes responses from the open-ended free response of the Test of English as a Foreign Language, a test used to evaluate the level of English proficiency of an L2 English speaker wishing to enter an American university. Prompts for the free response section encourage extemporaneous, independent writing and are non-specific enough as to allow for a wide range of test taker answers (e.g., Prompt 3: *Do you agree or disagree with the following statement? Young people nowadays do not give enough time to helping their communities. Use specific reasons and examples to support your answer.*)

Essays in the TOEFL exam are scored on a scale from 1-5. Our sample of the TOEFL11 corpus only focuses on those essays that scored in the *Medium* proficiency category (i.e., received a score between 2.5-3.5). Each essay is written by L2 speakers of English who have one of 11 possible L1s (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish).

2) Using the TOEFL11 Corpus

We have split the TOEFL11 Corpus from its original size to include only documents that received a score of *Medium*, and have chosen a sample that evenly distributes across the various prompts and languages present in the original corpus. The list of files, in .csv format, can be found on our Github page under the TOEFL11_part folder. Each file provides the name of the file (in ####.txt format), the given prompt number (P1…P8), the writer's native language, as well as their score.

Written text for each prompt should be read by the annotator at least once before reading through the essays in order to contextualize the writers' responses. Prompts and text files will be available in the section of the corpus that we provide.

3) Tagging Guidelines

Annotating each document will require several passes over the span of the document, focusing on various subsets of errors during each run-through. We have structured this to narrow the scope of the annotator's focus during his or her markup process. A detailed description of each phase of the annotation is provided below. At the end of these phases, for the annotator's convenience, we have provided a step-by-step checklist workflow for the annotators to follow.

3.1) 1st Pass: Misspellings and punctuation

3.1.1) Misspellings

On the first read-through of the document, the annotator should focus specifically on categorizing the writer's misspellings into several categories: Vowel Issues, Homophone, Segment Voicing, Cognate, Other (description of each to follow below). We have provided automated markings of spelling errors in pre-processing. As you read through the document, you will annotate each spelling mistake and choose a correct category.

Important Note: Note that some spelling mistakes, such as homophones, **may not** be picked up by the automatic spell checker. As you read, carefully look for these spelling mistakes that are not marked by the automatic spell checker and mark them as appropriate.

3.1.2) Tag types – Misspellings

The extent tag "Misspelling" should consume the character span of the misspelled token (remember, you should not only focus on the misspellings marked by the automatic checker, but also for any misspellings that may not have been caught). It has a required attribute field where annotators must choose from a set of possible options, listed below.

Misspelling Error Attribute: Possible Choices
These choices have been listed in a hierarchical rank. If a spelling error appears to encompass multiple meanings, choose the attribute that comes *first* in the list. (Example: if a word has vowel issues AND is a homophone, label it with the vowel issues attribute).

Additionally, if you are on the fence or uncertain if a misspelling falls into a certain category, *do not* label it as that category and see if another category might be more appropriate. If no category is appropriate, choose the attribute other.

Vowel Issues: Chosen if the misspelled word has included incorrect vowels or missing vowels.
EX. A dack was quacking.
EX2. My favrite food is bananas.
Dack: Misspelling. Misspelling Error: Vowel Issues

Homophone: Chosen if the word has been spelled as a homophone of the intended word.
EX. I went to the park and saw many people their.
Their: Misspelling. Misspelling Error: Homophone.

Segment Voicing: Chosen if the word has used segment voicing of consonants somewhere in the word.
EX. My daughter attends sghool in Waltham.
Sghool: Misspelling. Misspelling Error: Segment Voicing.

Cognate: Chosen if the word is spelled similarly enough to be a **possible** cognate of English in the native speaker's language.
EX. I live in an appartement in New York City.
Appartement: Misspelling. Misspelling Error: Cognate.

Other: Chosen if the spelling mistake fits none of the above categories. *All other categories should be considered first before Other!*
EX. Teh game is so fun.
Teh: Misspelling. Misspelling Error: Other

3.1.3) Awkward Punctuation

Additionally, you will need to look for punctuation that appears in unnatural places. We have defined this as punctuation such as semi-colons, periods, colons, commas, exclamation points, and question marks that appear in unnatural and ungrammatical places.

3.1.4) Awkward Punctuation Tag

Awkward punctuation is an extent tag that should take up the character span of the awkward punctuation. This tag only applies to punctuation that appears in a bizarre or even ungrammatical place.
EX. We. went to the Red Sox game. .: Awkward Punctuation.
Note: As before, if you question or are uncertain if a punctuation is awkward or ungrammatical, *do not mark it.*

3.2) 2nd Pass: Nouns, Determiners, and Noun/Predicate Adjective Prepositions

On your second read-through, focus specifically on errors that deal with nouns, determiners, and prepositions. These will be marked for you using a POS-tagger during pre-processing.

3.2.1) Determiner Errors

Determiner errors will take the form of a mismatch between a determiner and its corresponding noun. This will be encapsulated using a Determiner extent tag, a Noun extent tag, and a DetNounLink linking tag that connects the two. Determiners and their nouns only need to be marked *if one or more of them has an error.*

3.2.2) Determiner – Extent Tag

Note #1: ANY Determiner tag will require a corresponding Noun tag and DetNounLink to go along with it. A determiner tag should consume the entire determiner present.
EX1. These girl from Paris is very friendly.
If more than one determiner is present, then each determiner should be tagged *with its own tag*.
Ex2. These two girl from Paris are very friendly.
Each determiner tag comes with *two optional attributes*, shown below.
3.2.2.2) Determiner – Attributes: Correct Form and OtherCorrectForm
If the correct form of a determiner can be repaired given the context (i.e., you can infer what the correct form of the determiner should be), choose it from the provided list of determiners: {a | an | the | some | this | that | these | those | NUMBER | POSSESSIVE | OTHER}
NUMBER refers to any real number that could be used as a modifier.
POSSESSIVE refers to any genitive pronoun or phrase.

OTHER refers to any determiner that you feel should have been used by the writer but is *not* present in this list. In that case, fill it in using OtherCorrectForm. If you feel there should be no determiner present at all, you can write *NONE* in this field. (Also, make sure to select the ExtraDeterminer field on the Noun).
Note #2: If you are uncertain whether the error is in the determiner or the noun, choose the one which would most preserve the grammaticality of the sentence. In EX1, *girl* and *is* agree in number, so we have evidence that the intended determiner was *this*. If there is insufficient evidence or you still remain uncertain, place the error in the noun (see Noun section below)

Note #3: If a determiner is *missing*, this should be encoded within the noun on which it would be dependent, if it were present.

3.2.3) Noun tags

Nouns should be tagged if they, their modifying determiners, their modifying prepositions, or some combination thereof have one or more errors.
When tagging nouns, several areas to note:

1. Only the head noun phrase should be tagged. If this is a noun-noun compound, include the maximal noun-noun compound. Otherwise, do not include other modifying phrases.
   EX. These *chemistry class* at Brandeis had too many students.
   (Here, *chemistry class* would be annotated as a noun.)
2. If a noun is part of a coordinating clause, use the span feature to get a disjointed span. Do not include the coordinating word. If it seems more appropriate to use more than one noun tag (i.e., the two noun phrases each have their own determiner or preposition, each with its own, separate problem) then more than one noun tag can be used.
   EX. My most good *friend* and *brother* came with me.
   (Annotate *friend* and *brother*).
3. Nouns that are missing determiners or prepositions (or both) need to have this information marked *on the noun* and does not get marked with a determiner or preposition tag. If there is a mismatch between noun and preposition and/or noun and determiner, then *both* the preposition and the noun and/or the determiner and the noun are annotated.
   EX. Apple at Waltham is very yummy.
   Noun: DetError – Missing Determiner
   Noun: PrepError – Incorrect Preposition
   Preposition: Correct Preposition – "from" (CDATA)
   PrepositionLink: arg0 "prep" at
   Arg1 "mother" Apple
4. Nouns may have issues with their subject-verb agreement. At this stage, this does not need to be encoded. In the third pass, nouns and verbs who disagree in subject will be marked with an SVDisagreement tag. (See Section 3.3 for details)

3.2.4) Noun attributes

There are three categories of noun attributes, described in more detail below. All are optional and should only be chosen if the phenomenon in question is observed. More than one category may be chosen.

PrepError (implied): There is a problem with a preposition. NOTE: For the sake of our annotation, infinitival "to" should be treated as a preposition.
~MissingPreposition: There is a preposition that is not present in the sentence that should be. No other link or extent tag is required.
EX. It was time go to school, so I leave.
Noun: PrepError – Incorrect Preposition

~ExtraPreposition: There is a preposition present when no preposition is required. A tag on the noun (with this attribute) should be included, as well as a tag on the Preposition. The two should be linked using a corresponding PrepositionLink tag.
EX. A letter for about students….
Noun: PrepError – Extra Preposition
Preposition (no CDATA required)
PrepositionLink: arg0 "prep": for
arg1 "mother": letter
~IncorrectPreposition: The preposition provided is incorrect for the given context. As above, there should be a tag on the preposition, with a corresponding PrepositionLink tag.
EX. Young people about America don't help their communities.
Noun: PrepError – Incorrect Preposition
Preposition: Correct Preposition: "from"
PrepositionLin: arg0 "prep": about
Arg1 "mother": people

DetError (implied): There is a determiner issue in dependency with the noun.
~MissingDeterminer: There is no determiner included when there should be. No other link or extent tag is required.
EX. Car is very useful in modern society.
Noun: DetError -- MissingDeterminer
~ExtraDeterminer: A determiner has been included when no determiner is required. A Determiner tag with the corresponding DetNounLink tag (explained below) should be included.
EX. There are a lot of cars in the society.
Determiner: WrittenForm "the" CorrectForm "OTHER" OtherCorrectForm "NONE"
Noun: DetError – ExtraDeterminer
DetNounLink: arg0 "article" – the
Arg1 "noun" -- society
~IncorrectDeterminer: The provided determiner is incorrect for the given context. A Determiner extent tag with corresponding DetNounLink link tag should be included.
EX. These cat are very friendly.
Determiner
Cat: DetError -- IncorrectDeterminer
DetNounLink – arg0 "article" – these
Arg1 "noun" – cat

PlError: There is a problem with the count of a noun.
~NotSingular: This should be chosen if a noun has been made plural, but should, in fact, be singular in a grammatical sentence.
EX. Furnitures is not well made in my country.

Noun: PlError -- NotSingular
~NotPlural: The converse of above. This should be chosen if a noun has been made singular, but it should be plural.
EX. It is important to take some risk.
Noun: PlError -- NotPlural
NOTE: For PlError, if it is ambiguous or difficult to repair the count of the noun, simply skip this category.

3.2.5) DetNounLink

DetNounLink is a link tag that should be used between determiners and their corresponding nouns. This tag should only be used if:

a) There is a determiner and a noun present
b) There is a mismatch between them, or there is a determiner present when no determiner should be present

For determiners that haven't been included that *should be* included, use the MissingDeterminer category of the Noun attribute DetError.

For DetNounLink, the Determiner and Noun tags extent tag should be linked and labelled with their corresponding 'article' and 'noun' labels within DetNounLink.

3.2.6) Preposition

As above with the determiner tags, any *missing prepositions*, i.e. any prepositions that are not included that should be should be marked on their corresponding 'mother' head or dependency. In any other case, when the preposition has been included and it shouldn't be, or it is a mismatch, the preposition should be labelled using the preposition tag. If possible, fill in the *correct form of the preposition* using CDATA. All prepositions should be linked using the Preposition tag.

For this round, only focus on prepositions that are dependent on *nouns* or *predicative adjectives*. As before, only prepositions that have *some issue with their dependency* need to be marked. Problems with prepositions related to verbs will be captured on the third pass.

3.2.7) PrepositionLink

The PrepositionLink tag should be used to link together a preposition and its corresponding 'mother'. For this round, that 'mother' will be either a *noun* or a *predicative adjective* (explained in more detail below). A PrepositionLink will always require two extent tags to be linked. Be sure to label the phrasal head 'mother' with the 'mother' attribute and the preposition with the 'preposition' attribute.

3.2.8) Predicative adjectives (Adj tag)

IMPORTANT NOTE: The adjective tag should only be used with *adjectives* that --
1) Fall as the predicate of the sentence.
2) *Have a problem with a preposition in their dependency.*

EX. It is really necessary at me to buy a car.
Adj: PrepError – IncorrectPreposition
Prep: CorrectPreposition – "for" (CDATA)
PrepositionLink: arg0 "prep" at
arg1 "mother" necessary

The extent tag should consume the entire adjective. No modifiers should be included.

As with nouns, the predicative adjectives have an attribute for PrepError that includes MissingPreposition, ExtraPreposition, and IncorrectPreposition. As above, if there is no preposition and there should be, use MissingPreposition. No other link or extent tag is required. However, if there is an extra or incorrect preposition, use ExtraPreposition and IncorrectPreposition, respectively, along with their corresponding Preposition and PrepositionLink tags.

3.3) 3rd Pass -- Auxes, Verbs, Subject-Verb Disagreement, Inconsistent Tense, and Missing Copulae

In this third pass, please focus on issues surrounding verbs and their prepositional dependencies. Instructions for the specific types of errors we'd like to focus on are expounded upon below.

3.3.1) MissingCopula

If a copula (i.e., the verb to be) is missing when it should be present, mark *the whitespace* where you feel the verb should fall using the extent tag *MissingCopula*. Type in the required field CorrectForm with what your intuition is for the correct verb form.
EX. That girl __ scary.
MissingCopula: CorrectForm "is" (CDATA)

3.3.2) Verbs

Verbs should be marked if there are errors:

- In verb form
- With a preposition that is dependent on them
- With agreement with the subject
- If the tense does not remain consistent within the sentence in a grammatical or plausible manner

Verbs have the following attributes:
MissingSubj: If a subject is missing, this attribute should be changed to Yes. No other link or extent tag is required. The default value is No.
EX. Rains a lot this year.
Verb: MissingSubj: Yes
PrepError: As above, if a verb is missing a preposition, choose MissingPreposition. No other link or extent tag is required. However, if a preposition is *present* and there is some type of problem (mismatch or extra), then the preposition should be tagged with Preposition and a corresponding PrepositionLink tag is needed.
EX. I talk my job a lot with my cowokers.
Verb: PrepError – MissingPreposition
WrittenForm & CorrectForm: If the written form can be categorized into one of the categories provided, or if you have an intuition for what the correct form should be, choose it from the list. If you have an intuition for what the form should be, but it is not provided in the list, choose other. Otherwise, this field may be left blank. (NOTE: For the purposes of our annotation, base form will constitute the morphological base (i.e., all verb forms that are phonologically equivalent to the base form, even if they are inflected. In the below example, see would count as the base form, as it is phonologically equivalent to the infinitival form). *NOTE: If you feel there is ambiguity or uncertain which tag to pick,*
EX. She see my brother often.
Verb: WrittenForm: base
CorrectForm: 3rd sg
WrittenTense: If and only if there is a problem with the tense of the verb, choose one of the tenses provided here. If the problem is with the future tense, select only the verb, not the modal. *NOTE: If you are uncertain or unsure of which tag to pick, skip this field.*
EX. Last year, I write an essay for my English class.
WrittenTense: Present

NOTE: Only main verbs should be marked with the verb tag. Auxes have their own special tag, which will be explained below.

3.3.3) Auxes

Auxes should only be marked if there is some problem in the agreement between the Auxes and the Main Vs that they modify, or if they contribute to an Inconsistent Tense within a sentence.
Auxes have an included *WrittenForm* and *CorrectForm*, as well as *WrittenTense*, as above. Note that *WrittenTense* will not apply *ever* to a modal. If it is possible to ascertain the *WrittenForm* and *CorrectForm*, choose the correct form from the list, as you would with verbs above.

NOTE: There is no need to link an Aux to its main V, as they will be within a reasonable locality as to ascertain this information without a link.
EX. I has written many essays.
Has: Verb: MissingSubj: No. WrittenForm: 3rd sg CorrectForm: base
Noun
SVDisagreement: arg0 "subject" I
Arg1 "verb" has

3.3.4) Subject-Verb Disagreement

Use this link tag if there is a disagreement in person and number between a subject and its corresponding verb. Both the Noun subject and its corresponding Verb will need to be marked with a Noun or Verb/Aux tag, respectively.
As above, only the heads of noun phrases should be tagged, and coordinated NPs should be marked only with their heads as well.
The SVDisagreement should be linked *to the corresponding finite verb*. This could be a Verb tag OR an Aux tag. Use the link tag to connect the subject and the verb, and make sure each is labelled with the 'subject' and 'verb' tag, respectively.

3.3.5) InconsistentTenseLink

Use this link if the tense of a given sentence changes abruptly as to be bizarre, ungrammatical, and/or outside the realm of semantic possibility. For the ease of annotation, these inconsistencies should only be considered *intrasententially* (that is, within a given sentence.) We will define a sentence as ending with a period – other punctuation, such as a colon, semi-colon, or comma signal the continuation of a sentence.

The *finite* verbs of a sentence should be linked using InconsistentTenseLink (thus, they can be either Auxs or Verbs, depending on the situation.) These should be labelled with their corresponding labels "verb1", "verb2" (up to a maximum of 6 verbs).
EX. Last year, I write an essay before class and then I showed it to the teacher.
Verb "write": MissingSubj: No. WrittenTense: Present
Verb "showed": MissingSubj: No.
InconsistentTenseLink: arg0 "verb1" write
arg1 "verb2" showed

3.4) 4th Pass – Wrong POS

This final pass through the document is to look for any errors in part of speech that:

- Have not been captured by any of the above tags
- Constitute an error in the usage of the word. The wrong part of speech is being used in a given context, as though it has been analyzed by the learner as having a different part of speech.

EX. In order to success in life, one must work hard for to success.
Here, it appears that success has been identified as a verb, so it would get tagged using the WrongPOS tag.

WrongPOS has two attributes that need to be filled in:
WrittenPOS: Identify, as best as you can, the part of speech of the word included by the writer. If you are unable to identify it, or it is not on the list, choose OTHER.
CorrectPOS: Identify, as best as you can, the part of speech of the word that the author *should have written* in this context. If you are unable to identify it, or it is not on the list, choose OTHER.