

Linguistic Phenomena Annotation Guidelines

March, 2013

Abstract

These guidelines were prepared by Ned Letcher for the development of a corpus of linguistic phenomena to be used for the purpose of investigating techniques for the automatic detection of linguistic phenomena in precision grammars. This project is based at The University of Melbourne and University Washington. For more information please contact Ned Letcher at ned@nedletcher.net.

1 Overview

Firstly, thank you for taking the time to annotate a portion of our linguistic phenomenon corpus! The main aim of this corpus is for it to drive the development of techniques for the automatic detection of linguistic phenomena within precision grammars. We believe that such techniques will increase the utility of both precision grammars and their treebanks as linguistic resources — you are working towards a better future!

In this annotation task you will be asked to annotate utterances with linguistic phenomena that occur within them. By *linguistic phenomena* we mean interesting constructions or features et cetera that linguists concern themselves with when documenting languages. Below you will find general instructions on how to perform the annotation as well as specific instructions for each phenomenon.

2 Annotation Instructions

For each item, you are required to annotate all instances of the five phenomena from Section 3 found within the item. This means that if an item is not annotated with a particular phenomenon, this item will be interpreted as not exhibiting that phenomenon. We recommend doing each phenomenon in a separate pass, as it can be hard to

keep track of the criteria for every phenomenon simultaneously. This means that after reading the remainder of this section you can read the instruction for each phenomenon as you get to it.

2.1 How to Annotate

For each instance of a phenomenon, you will annotate the span of the surface string that you deem to be relevant to the phenomenon. For example, in (1) you would annotate the span *a car was hit by a truck*, as a passive clause.

- (1) As I walked down the street, [a car was hit by a truck,] and rolled over thrice.

You should try to annotate the minimum span involved in the phenomenon, however annotations must still be complete. For example, when annotating an entire clause, you should include adverbials at the beginning and end of an item, even though they may not appear to relate to the phenomenon in question — they still are a part of the clause. It will mostly be the case that annotations of the selected phenomenon will often coincide with perceived constituent boundaries, however this may not necessarily be the case across all phenomena/languages.

The annotation tool you will be using is brat, a tool which runs in the browser. In order to annotate a span, simply select the desired text span with the cursor; then in the resulting pop-up box select the phenomenon to annotate the span with. Double-clicking on an existing analysis will allow you to move the span or delete it. For further help, see the brat manual¹.

2.2 Punctuation

Punctuation that occurs within a span will obviously be included within that span, however the question arises of what should be done when punctuation occurs at the beginning or end of a span. In general our approach is that punctuation should bind tightly and be included within a span if it sits adjacent to the span and is not clearly associated with the preceding/following word. For instance, in (2) when annotating interrogative clauses, the span that would be annotated is “*What time is it?*”.

- (2) [“What time is it?”] I asked.”

¹<http://brat.nlplab.org/manual.html>

Note that this tight binding of punctuation should be respected even if the data has been tokenized so as to put whitespace between words and punctuation characters. If it is ambiguous which direction a punctuation character should bind, then it should attach to the token to its left.

2.3 Coordination

When a candidate for annotation involves a coordination, annotate as two separate instances of the phenomenon if the coordination is at the top level of the constituent pertaining to the phenomenon in question. If the coordination occurs below the top level constituent then annotate this as one instance. For instance, when annotating complement clauses, you would annotate (3a) as two instances and (3b) as one. Likewise, when annotating relative clauses, (3c) would be annotated as two instances and (3d) would be annotated as one instance.

- (3) a. I thought [*that we would go to the shop*] and [*that we would get a sausage roll.*]
- b. I thought [*that we would go to the shop and get a sausage roll.*]
- c. I like people [*who tell the truth*] and [*who eat shrimp.*]
- d. I like people [*who tell the truth and eat shrimp.*]

2.4 Ambiguity

In the case of genuine ambiguity, such as prepositional attachment ambiguity, annotate under the basis of attaching higher rather than lower.

3 Phenomenon-specific Instructions

The following sections outline how we are defining each phenomenon and how they should be annotated. Read through the overview of each phenomenon before beginning its annotation paying particular attention to the instructions in the *How to annotate it?* sections, but don't feel like you must internalize everything before beginning — in some cases there is more information than you will need for most annotations.

3.1 Passive Clause

The passive construction is a valence-changing construction that alters the way thematic roles are mapped onto grammatical roles so as to foreground the patient role.

- (4) a. Scientists discovered the Higgs Boson.
b. The Higgs Boson was discovered (by scientists).

(4) shows a transitive active sentence and the corresponding passive form in which the patient, *the Higgs Boson*, has moved from direct object to become the subject, and the agent has moved from the subject to the oblique prepositional phrase *by scientists*.

The passive construction can be broken up into two distinct processes. The first, agent demotion, sees the agent demoted from subject to an oblique, or completely omitted, having the effect of reducing the valency of the predicate by one. The second, patient promotion, promotes the patient from direct object to subject. For a clause to be counted as a passive construction it must at a minimum involve the process of patient promotion.

Watch out for clauses like (5) where the participle *broken* is not functioning as a passive participle but as an adjective derived from a passive participle. Where it is clear that such participles are functioning as adjectives, the clause will be ignored. Sometimes the syntactic environment together with textual context is insufficient to resolve this ambiguity, in which case mark it as a passive.

- (5) The door was quite broken.

3.1.1 How to annotate it?

- Annotate the span covering the entire passive clause.
- Include the agent phrase if there is one.
- Passive relative clauses should be included.
- Exclude clauses with adjectives formed from past participles.

3.2 Imperative Clauses

In this study we will restrict our attention to clause forms dedicated to serving an imperative function, which is to say that they are uttered when appealing to an audience to bring about a desired state of affairs and the audience is deemed to be control of this state of affairs. Such an example is presented in (6).

- (6) Live long and prosper.

In this study we will exclude negative imperatives (prohibitives) and we will also ignore imperative types that gain their imperative force via conversational implicature, unless this is the primary means the language has for realising the imperative. This means that examples such as in (7) will not be counted as imperatives.

- (7) a. You will take a seat.
b. Can you take a seat?

3.2.1 How to annotate it?

- Annotate the span covering the entire imperative clause.
- Do not include prohibitives.
- Do not include utterances which gain their imperative force via conversational implicature.

3.3 Interrogative Clauses

Interrogative clauses are those whose purpose is to solicit information from the addressee. By looking at the types of responses that interrogatives solicit, they can be broken up into three subtypes:

- (8) a. Do you like coffee? (polar question)
b. Do you like coffee, or tea? (alternative)
c. What kind of coffee do you like? (*wh*-question)

There is also a particular type of interrogative construction, the tag question (as illustrated by (9)), which can be thought of as a construction that forms polar interrogatives out of declarative sentences by adding an appropriate word or fixed phrase.

- (9) Coffee is fantastic, isn't it?

In this study we will include all the different kinds of interrogatives identified above: polar, alternative, content questions and tag questions. One particular type of interrogative we will exclude are those that occur as subordinate clauses, such as indirect questions like those found in (10).

- (10) a. The researcher asked *whose coffee he had spilled*.
b. The technician wondered *how much the equipment was worth*.

3.3.1 How to annotate it?

- Annotate the span covering the entire interrogative clause.
- Include polar, alternative, content and tag questions.
- Exclude indirect questions.
- Exclude declarative clauses end in a question mark and do not display the internal structure characteristic of interrogatives.

3.4 Complement Clause

For this study, we are defining complement clauses as subordinate clauses that are characterized by the following two properties:

1. The constituent possesses the internal structure of a clause, with it possessing a predicate at a minimum, and any arguments of this predicate being marked in the same way as they would be in a main clause.
2. The constituent functions as an argument of a verbal predicate in a higher clause.

For example, the clause *that you are a light sleeper* functions as the object of the verb *understand* in (11a) and as the subject of the verb *frustrate* in (11b).

- (11) a. I understand *that you are a light sleeper*.
b. *That you are a light sleeper* frustrates me.

Complement clauses may take the form of a sentence-like clause, as in (12a) — in which case they tend to represent a fact or proposition — or they may be reduced in some way, as in (12b) and (12d), in which the subject position of the complement clause is realised as a gap.

- (12) a. I remembered (*that*) *I don't like durian*. (that-clause)
b. I remembered *to feed the trolls*. (infinitive clause)
c. I remembered *seeing her*. (participial clause)
d. I remembered *his singing the song*. (gerundal)

A potential cause for confusion may be found in English clauses such as *his sinking the Titanic*, which may at first appear to be a nominal phrase due to the superficial similarity with the NP *his sinking*

of *the Titanic*. These can be teased apart, however, by noticing that in the case of the former, *sinking* is a gerundal verb which takes *the Titanic* as an argument, whereas in the latter case *sinking* is a deverbal noun (or nominalized verb) modified by *the Titanic*. A useful test in such cases is whether the constituent will accept an adverb, as this will not be possible if the constituent is an NP.

- (13) a. I enjoyed *his sinking the Titanic*. (complement clause)
 b. I enjoyed *his sinking of the Titanic*. (noun phrase)
 c. I enjoyed *his sinking the Titanic slowly*.
 d. *I enjoyed *his sinking of the Titanic slowly*.

3.4.1 How to annotate it?

- Annotate the span covering the entire complement clause.
- Include any complementizer such as *that* in the span.
- Do not include the verb of the matrix clause in the span.
- Ignore clauses that are the complements of things other than verbs — if the clause attaches to a noun, it is possible it should be annotated as a relative clause.

3.5 Relative Clause

Relative clauses are a specific type of subordinate clause which serve to restrict or elaborate on one of the referents in the higher clause. In this study we will only look relative clauses which modify nouns. The relativisation is achieved via one of the arguments of the subordinate clause sharing its reference with the head noun of the constituent on which the subordinate clause is dependant.

- (14) a. I see the tree *that has a thick trunk*.
 b. He chose the boy *who eats worms*.
 c. He chose the boy *whom I saw*.
 d. You know that band *whose tunes I like*.

Relative clauses also come in the form of reduced relative clauses. This can occur when the relative pronoun is omitted, as in (15a), but also when the relative clauses is formed with the participial form of the verb, as in (15b).

- (15) a. He chose the boy *I saw*.

- b. You know that band *swing the little boy*?

Relative clauses can also be restrictive or non-restrictive. A restrictive relative clause serves to refine and further delimit the reference of the head noun it modifies. A non-restrictive relative clause does not restrict the reference, but instead elaborates upon the information being provided about the referent picked out by the head noun. In English the distinction between the two is made by placing commas around the non-restrictive relative clause, as illustrated in (16).

- (16) a. The dog *which has a long tail* sat down.
b. Their house, *which had a lovely fence*, sat on a hill.

3.5.1 How to annotate it?

- Annotate the span covering the entire relative clause.
- Include relative pronouns linking the relative clause in the span.
- Do not include the head noun that the clause modifies in the span.
- Include non-restrictive relative clauses and reduced relative clauses.
- Do not include the comma to the left of non-restrictive relative clauses in the span.



Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Letcher, Ned

Title:

Linguistic phenomena annotation guidelines

Date:

2013

Citation:

Letcher, N. (2013). Linguistic phenomena annotation guidelines.

Publication Status:

Unpublished

Persistent Link:

<http://hdl.handle.net/11343/33329>

File Description:

Linguistic phenomena annotation guidelines