Guidelines for Linguistic Annotation of Phrases and Expressions in Clinical Records

Aleksandar Savkov, John Carroll, Jackie Cassell

1 Introduction

The Purpose of these guidelines is to introduce the reader to the annotation of general practitioner (GP) notes with syntactic chunks and semantic expressions. To achieve that these guidelines provide some basic grammar and linguistics knowledge, as well as some additional instructions about annotation techniques.

The Task itself amounts to identifying and annotating a number of different linguistic phrases and expressions in GP notes using the web-based annotation tool **Brat**.

The Motivation for this task lies in the crafting of gold standard data for training and evaluating machine learning tools that will automate the process of linguistic analysis. This automated process will ultimately serve as the basis of more complex analysis leading to the automated extraction of information about symptoms and diseases from GP notes.

The Information in these guidelines is distributed in three sections. The **Common Grammar** section introduces the reader to basic notions of grammar and linguistics. This section may be skipped by a reader with prior linguistic experience. The **Chunks** section explains the notion of syntactic chunks and their annotation according to these guidelines. In the last section, called **Annotation**, we discuss the details of a good annotation practice, as well as some of the specific issues and tasks of the annotation of medical records.

Notes on the use of bold and italics in the guidelines. All examples are marked with *italics*. The focus area of the example, e.g. the phrase head, is marked with ***bold italics***. Key points in the guidelines are highlighted in **bold face** only.

2 Common Grammar

This section explains the basic notions of parts of speech, noun and adjective phrases, and main verbs. This information is crucial for a complete understanding of the guidelines. However, readers who are familiar with basic grammar should be able to skip it and continue reading at Section 3.

2.1 Parts of speech

English words are traditionally classified into eight lexical categories, or parts of speech: nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions, and interjections.

- **Nouns** are abstract or concrete entities: a person (*policeman*, *Michael*), a place (*bank*, *Brighton*), a real object (*tie*, *radio*), an imaginary object (*unicorn*), a feeling or an idea (*joy*, democracy), or a quality (*cleverness*). Nouns could be grouped together to form compound nouns as in *bus driver*, *desk lamp*, or *party animal.*
- **Pronouns** are generally used instead of nouns and personal names in different situations with different functionality. Here are examples of all the types: *I*, *you*, *we*, etc.; *me*, *her*, *them*, etc.; *my*, *mine*, *your*, *yours*, *their*, *our*, etc.; *this*, *that*, *these*, *those*, etc.; *anyone*, *anything*, etc.; *who*, *which*, etc.; *who*, *whom*, etc.
- **Adjectives** are words used to describe qualities and attributes of nouns: *green, lazy, tall, heavy, kind.* They also include comparative and superlative forms like *better, best, worse, worst, taller, tallest, etc.*
- **Verbs** are words that indicate an action (*walk*, *write*), an occurrence (*happen*, *occur*), or a state of being (*be*).
- **Adverbs** are qualifiers of adjectives (***slightly*** *green*; ***absolutely*** *fresh*), verbs (*to work* ***efficiently***; ***suddenly*** *disappeared*), clauses, sentences, or other adverbs (*walking* **slightly** *impatiently*). They are usually the answer to the questions *How?*, *Where?*, or *When?*.
- **Prepositions** are words that express some sort of relation, for example a spacial relation is expressed by prepositions such as *to*, *under*, *before*, *inside*, etc.
- **Conjunctions** are words that connect other words and phrases, e.g. *and* and *or.*
- **Interjections** are words of emotional greeting (or exclamation) like *wow*, *tut-tut*, *ugh*

2.2 Phrases

In everyday speech, an arbitrary group of words may be called a **phrase**. However, in linguistics, a phrase is defined as one word or a sequence of words that function as a single unit in the syntax of a sentence.

2.2.1 Noun Phrases

A noun phrase (NP) is a unit centred around one noun or pronoun or a gerund, which is called the **head** of the noun phrase. The rest of the phrase consists of **modifiers** that give further information about the head. So if a noun denotes an entity (e.g. *dog*), the noun phrase provides us with more information about that entity. For example, in the sentence *John has a big brown dog* the noun phrase *a big brown dog* gives us information about the colour and size of a dog owned by John. Other examples of the information conveyed by modifiers are attributes (*the green mile*), location (*the door in the floor*), ownership (*my girl*), quantity (*seven samurai*), and other more complex notions (*the girl who played with fire*). Usually the entity that is in the focus of the noun phrase is a noun or a pronoun, but it could also be the present participle form of a verb (as in *I love **reading***), which is called a **gerund** (see Section 2.3 for more information).

2.2.2 Adjective Phrases

Adjective phrases (AP), are syntactic constructions with a head and zero or more modifiers. The head of an adjective phrase is naturally an adjective, e.g. *very **fast***. The number of words in APs may vary just like the one of NPs. APs could also include some modifiers of the adjective head as in *The river is **crystal clear***, *wound **severely infected***, *The wine tastes **very good*** and *Patient feels **slightly constipated***.

2.3 Verbs

For the purpose of these guidelines we discuss three types of verbs depending on their semantic and syntactic roles in a sentence or a clause. The first group is that of the **main verbs** which express the central action, occurrence, or state of being of the sentence or the clause. The second group, which are called **auxiliary verbs**, carry additional grammatical information about the main verb such as tense, (passive) voice, or modality. It is important to note that some auxiliary verbs can be also main verbs and even be used twice in the same sentence with different function or meaning as in sentence 2. below. Some main verbs, called **phrasal verbs**, on the other hand, can be comprised of a verb and a preposition or a particle as illustrated in example 6. below. The following examples show verbs in SMALL CAPS and main verbs in **BOLD**:

1. *The bears* ARE **EATING** *the berries in the garden.*
2. *The bears* HAVE *always* **HAD** *berry snacks in the summer.*
3. *The bears* WILL **EAT** *berries.*
4. *The bears* CAN **EAT** *berries if they **find** any.*
5. *The bears* SHOULD *not* **EAT** *too many berries.*
6. *The bears* COULD **RUN INTO** *berry bushes.*

The third type of verbs, which are called **raising verbs**, always appear in conjunction with the main verb as shown (in bold) in the following sentences:

7. *The bears* **NEED** *to* EAT *berries.*

8. *The bears* **HAVE** *to* EAT *berries.*

9. *The bears* **APPEAR** *to* EAT *berries.*

10. *The bears* **SEEM** *to* EAT *berries.*

2.4 Gerunds

Gerunds are a special case where verbs in present participle form (ending with *-ing*) act as nouns and form noun phrases. Here are some examples of base NPs with gerunds:

11. *This house needs* ***cleaning****.*

12. *The patient has* ***normal bowel emptying****.*

13. ***Apple picking*** *is fun.*

3 Annotation Types

In this section we describe the different types of annotations and we provide guidance for their correct annotation. The annotations are divided into two groups: phrase chunks and expressions. The former includes syntactic chunks based on noun phrases, adjective phrases and main verbs, while the latter includes locative, temporal, quantitative and "on-examination" expressions.

3.1 Base Noun Phrase Chunks

Base NPs are a subset of NPs that **do not contain other NPs**. A useful trick for identifying them is to watch out for prepositions, which should not be present in them. For example, ***The cat*** *in* ***the mirror*** is not a single base NP chunk, but two of them (in bold). To help identify base NPs we provide a list of modifiers that can appear within them along with examples:

- determiners
  - articles: ***the*** *dog*, ***a*** *cleaning*
  - demonstrative pronouns: ***this*** *girl*, ***that*** *man*
  - possessive determiners: ***my*** *homecoming*, ***your*** *dog*, ***our*** *car*, *the* **police officer's** *wife*, *the* **neighbour's** *constant complaining*
  - quantifiers: ***some*** *people*, ***every*** *day*, ***most*** *children*, ***any*** *student*, ***all*** *birds*, ***no*** *coffee*, ***five*** *cakes*, ***10*** *miles*
- adjectives preceding the head, such as *large*, *beautiful*, *sweeter*, *excruciating*, *soothing*
- nouns immediately preceding the head, such as *college* in ***a college*** *student*; note that the number of preceding nouns is unconstrained, so they could be stacked as in ***bus*** *driver*, ***school bus*** *driver*, ***city school bus*** *driver*

Annotation Two things need to be considered when annotating a base NP: its head and its borders. First, there should be just one head in the NP. In most cases this means just one noun, pronoun, or a gerund inside it. For example, *a good **dog***, *twelve angry **men***, ***someone***. However, one has to be careful not to mistake the nouns modifying the head as heads. Consider the base NP *the black dragon **tattoo***, in which the word *dragon* modifies *tattoo*, which is the head of the base NP. The same logic is applied when dealing with more than one preceding nouns, e.g. *the school bus **driver***. One should also bear in mind the opposite situation, where two base NPs are listed one after another without the usual punctuation, as well as the mixture of both. Consider the sample clinical text snippet *c/o fever cough back pain*, in which there are three base NPs: *fever*, *cough*, and *back pain*. Another matter to consider is the length of base NPs, which in some cases could be quite substantial, e.g. *the long winding high mountain roads*.

3.2 Adjective Chunks

We define **adjective chunks** (AP) to be adjective phrases that act as predicative expressions. In simple terms this generally means when they follow a copula verb. The copula verbs used most often in the notes are *to be* and *feel*, although there are also others: *seem*, *look*, etc. In grammatical text copula verbs should always be present in predicative constructions (*The carpet is **red***; *The sky looks **dark***), but in the notes they are often omitted and the construction is implied by the context: *blood pressure **normal***; *abdomen **tender***; *[baby] coos and **alert** and **happy***. Here are some examples of typical AP cases (in bold) with comments:

1. *patient is **anxious*** — standard predicative construction following the verb *be*
2. *chest **clear*** — omitted copula verb (*be*)
3. *leg feels **much better*** — the determiner *much* is part of the AP
4. *finger is **severely infected*** — the adverb *severely is part of the AP*
5. ***worried** wants to be admitted today* — predicative construction (*patient is worried*) was reduced to only an adjective phrase

Annotation When annotating an adjective chunks the annotator should take care to include all modifiers of the adjective (usually adverbs) as in example 3. and 4. above. The annotator should always bear in mind predicative constructions and not mistake NPs with adjective modifiers for adjective phrases *The patient has **high fever***.

3.3 Main Verb Chunks

Main verb chunks usually contain only the main verb. The only other words that may be included are adjacent prepositions or particles in the cases of phrasal verbs, like *show up*, *take care*, *calm down*, etc. Gerunds (see end of Section 2.3) should not be confused with main verbs, they should be annotated as part of base noun chunks (see Section 3.2).

Annotation The annotation of a main verb is more or less a straight-forward matter as its scope is usually a single word. The cases where its annotation scope spans over other words are those of phrasal verbs that include particles (*show up*, *calm down*) and/or prepositions (*power through*). However, those particles should not be annotated when they are not adjacent to the main verb as in *Please,* ***calm*** *him* ***down***. Verb negation should not be included in the main verb chunk annotation also when it is contracted (***is**n't*).

3.4 Expressions

Temporal Expressions are words, phrases or clauses that **contain information related to time**. Some refer to a specific moment or a period in time related to an event discussed in the sentence like *in seven minutes*, *an hour ago*, *yesterday*, *next year.* Others refer to the duration of an event, for example, *for three days*, *lasting two weeks.* The third type of temporal expressions describes frequency of repetition: *twice a day*, *every week*, *biannually*, etc. Temporal expressions may not always refer to time units directly, sometimes they refer to the time or duration of other events as in *last time*, *when they were young*, *while the sun was up.* There are also temporal expressions that are more vague and indefinite like *recently* and *already.*

Locative Expressions cover two types of expressions related to location. The first type points to the locus of a medical finding (infection, bruise, pain, etc.). The second type points to real places such as hospitals and geographical entities. The expressions may be a one of the modifiers of a base NP as in ***back*** *pain*, or a whole prepositional phrase such as ***in the hospital.*** There are cases in the notes where loci are expressed with omitted or ungrammatical syntactic constructions, for example ***chest*** *pain* ***lower left quadrant***.

Quantitative expressions represent some sort of **quantity or measurement**, like number (*five spots*), weight (*5kg*, *ten grams*), volume (*10cc*, *a pint*), length (*12cm*, *three inches*), etc. The quantity in a quantitative expression doesn't need to be explicit, it may be vaguely defined or inferred as in *several inches* and *a few kilos.* It is important to emphasise that the items that are quantified should also be part of the annotation, i.e. *a few* ***kilos***, *12****cm***. Quantities of time like *two hours* could also be regarded as quantitative expressions, but for the purposes of annotation they should **NOT** be annotated as quantitative expressions (see Section 4.3). There are also numbers that are not quantities, but identifiers or placements in a sequence, e.g. *group 3*, *second testing*, *phone 012345678.* Such number occurrences should not be annotated as quantities.

On-Examination **expressions** are different versions of the the expression *on examination* that **marks the border between the complains and the examination observations**. These expressions are constructed and/or abbreviated in different ways, e.g. *o/e* or *during examination.* The task of the annotator is to identify such expressions in the record.

4 Annotation Process

The process of annotation is the assignment of labels, called tags, to parts of the text based on the definitions and instructions in the sections above. This section discusses some technical issues and rules of conduct for this process, common problems and a few useful annotation tips.

4.1 Annotation Tasks

The first stage of the annotation process, called **prime annotation**, is the stage when two or more annotators annotate the same data independently, assigning the annotation tags listed below. After the prime annotation process is complete, the results go through a process, called **referral**, that resolves any disagreement between the prime annotations to ensure the quality and consistency of the annotation.

List of annotation tags:

- Noun Phrase Chunk (NP)
- Adjective Chunk (AP)
- Main Verb (MV)
- Locative Expression (LE)
- Temporal Expression (TE)
- Quantitative (QE)
- On-Examination Expression (OE)

4.2 Prime Annotation Tips

This section gives directions about specific data issues and discusses some general good annotation practices. It also gives some tips to help ensure more consistent annotation results.

We recommend that annotation is done one record at a time, considering the whole record and not just parts of it. This means that the annotator should read the whole record and try to understand its meaning before starting to annotate.

It is important to remember that lists of items of the same type should be annotated separately. For example, the Christmas presents in the following sentence are three different NP chunks: *Johnny got* ***a teddy bear****,* ***a remote-controlled car****, and* ***a hokey stick*** *for Christmas.*

If an expression seems complicated and the annotator is unsure how to deal with it, he or she can start by annotating to its left and right thus closing down its word span and making the task easier.

Annotation is not an exact science and sometimes the descriptions in the guidelines won't fit perfectly. In those cases the annotator should make an approximate decision, which is acceptable as long as it doesn't stray too far from the guidelines. The annotator should also have in mind that the decisions they make will undergo a referral process that is meant to improve the quality of annotation in exactly such ambiguous or unclear situations.

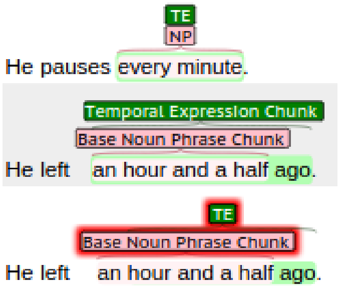
Consistency is the annotator's best friend. The annotator should make sure that they handle similar situations in the same way across the whole data. Going back to fix things is inevitable, but it gets harder towards the end of the annotation, so we recommend paying extra attention to cases that seem difficult in the beginning, when going back and fixing all previous occurrences is still feasible.

In cases of uncertainty we recommend being conservative. When none of the annotations seem to fit, no annotation should be made. Also in the cases of conjunction (see Section 4.4), chunks should be annotated separately unless it is clear they are part of entities that should be annotated together, e.g. *bits and pieces*, *this and that*, *black and white*.

Finally, we advise the annotators to not worry too much. They shouldn't "overthink" problems as this might lead to confusion. If they encounter more difficult cases, they should look for them or a version of them here in the guidelines or just make a note of them and carry on.

4.3 Priority and Embedding of Annotation

The phrase chunks defined in these guidelines are very restricted. They have a very basic form with just a few words. We call chunks and/or expressions embedded if one of them contains the other. They coincide if they include the same words. Embedding and coinciding of annotations is **allowed when a phrase chunk or a main verb is contained in an expression** or the other way around (see 1. and 2. in Example 1). However, while embedding phrase chunks into expressions is allowed, their **partial overlapping is not** (see 3. in Example 1). Therefore, in order to avoid errors in identifying the span of the expressions, we recommend that the annotation of phrases and main verbs precede the annotation of expressions. The rules for embedding annotation are incorporated in **Brat** (see Section 4.8), which warns the annotators when they make overlapping annotations by **highlighting it in red**.



He pauses every minute.

He left an hour and a half ago.

He left an hour and a half ago.

Example 1: Examples of annotation overlapping. NP chunks are denoted with bold face and temporal expressions are underlined

4.4 Including Conjunctions

Conjunctions are words such as *and* and *or*, which are used to join words, phrases, clauses, and sentences together. In most cases they are not included in the annotation of phrase chunks (NPs, APs, MVs), except when they connect modifiers (*the tall and handsome man*; *a green or blue jacket*; *slow but steady pace*; *scary yet exciting adventures*). Normally when conjunctions connect two phrases as in the sentence *We have* ***a cat*** *and* ***a dog***, the phrases are annotated separately. Only when they connect phrases that usually go together and have become fixed expressions like *black and white* or *bits and pieces*, they are annotated as one phrase.

However, if in doubt about whether a conjunction should be included in a phrase, it is recommended that the more conservative decision is taken, namely separate annotation.

The guidance above is not relevant to conjunctions within expressions (see Section 3.4), which are not defined so strictly.

4.5 Redacted Text

The clinical records data contained sensitive information that was redacted and replaced with the tilde symbol (∼). In most cases the context gives enough information to make a good guess what sort of words are missing – usually either names or places. Such redacted words should be annotated as if they were apparent. The annotator should also make a note of their guess about such words. For example, in *as per dr.* ∼∼∼∼∼∼ *advise* the doctor's name was removed, but it should still be annotated as an NP chunk. The annotators should use upper case letters with surrounding angle brackets to denote abstract entities in their guesses. For example, *<NAME>'s* will be a good guess for the redacted text from the example above.

4.6 Abbreviations and Acronyms

Abbreviations and acronyms need to be considered and annotated as their full forms. For example, the phrase *poss ovarian* should be annotated as an adjective phrase, and *FBC* (*full blood count*) should be annotated as a noun phrase. However, acronyms of phrases and sentences that could not be annotated with one tag like *spt* (*seen patient today*) should not be annotated. As a rule of thumb the annotators should think about acronyms as their full forms and annotate them accordingly **only** if the whole acronym can be annotated with one annotation.

4.7 Punctuation and Special Symbols

The nature of the clinical records data uses punctuation and special symbols in two different ways: 1. in their classical context and usage, and 2. as an abbreviation or a substitute for words and/or expressions.

Square bracket prefixes should be excluded from the annotations. For example, *[D]****Difficulty*** should be annotated as a NP-chunk starting after the closing bracket.

Normal punctuation

As a rule of thumb punctuation that is inside the span of an annotation should be left there (e.g. hyphens, (*on-line*), apostrophes (*Jimmy's*), commas, etc.) and punctuation that is on the annotation fringes should be excluded (e.g. quotation marks, braces, etc.). An exception to the last rule is the apostrophe sign in sentences like *The dog is* ***my neighbours'****.*

Special symbols and punctuation

As mentioned before, often punctuation and other symbols like a plus, a minus, slash, etc. are used for some peculiar unorthodox purposes. Annotators should try to use their best judgement in identifying the purpose of the symbols in these cases. Symbols that are used to convey the meaning of words should be treated as such. For example, question marks could replace the word *possible*, and a series of plus signs could indicate an increase in some value. We encourage the annotators to use their judgement in the annotation of such symbols, but we emphasise the need for consistency in their decisions.

4.8 Brat Annotation Tool

The annotations will be recorded using the **Brat** annotation tool, which is a web-based tool that allows annotators to access data from and input annotation to a remote server. No local installation is required, only **JavaScript** and **cookies** need to be enabled for the successful loading of the tool in the browser . The full functionality and performance of the tool is only guaranteed when using the latest version of one of the two supported browsers: **Google Chrome** or **Safari**. Even if the tool seems to run using other browsers, such as **Internet Explorer** or **Firefox**, they should not be used, because its performance there is not predictable and it may end up damaging the input or even the existing data.

Prime annotation
The prime annotators will be assigned a personal folder with documents each containing a small batch of clinical records. The folder should be loaded using the **Collection** button in the left upper corner of the interface. Annotations are created by selecting the portion of a text using the mouse and choosing the appropriate annotation type from the pop-up menu. Annotations can be edited or deleted by double-clicking on them at any time. We recommend reading through the tutorial available at `http://weaver.nlplab.org/~brat/demo/latest/#/` for a better and more practical understanding of the annotation process with Brat.

Annotation referral
The annotation referral process aims at selecting the best version of an annotation out of all prime versions. The referee should edit merged versions of all annotations, exposing all conflicting annotations. They should resolve the conflicts by **deleting** the less appropriate annotations.

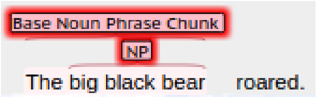
4.9 Annotation Referee

This section gives instructions for the **annotation referee** and should not concern prime annotators.

The most important notion that we want to emphasise here is that the annotation referee must NOT add any information, but only choose between already existing annotations without changing them. In the cases of only one existing prime annotation for a certain bit of text, the annotation should remain unchanged. If all prime annotations seem wrong, the one that seems nearest to a correct annotation among them should be chosen.

When comparing annotations of roughly the same chunks or expressions, the annotation label is more important than the annotated word sequence. For instance consider the sentence in Example 2. If one annotation identifies the word sequence *big black bear* as an NP chunk and the other identifies the word sequence *the big black bear* as an AP chunk, the NP chunk annotation is considered better, because even though it should include one more word, it is labelled correctly as a NP chunk.

The word sequence of an annotation on the other hand is important to the referee in cases of annotations with the same label. Then the annotation that includes the word sequence closest to the correct one is considered better. For example, consider the NP chunk *the big black bear* in Example 2, which is identified in two different ways (1. and 2.). The annotation in 2. is considered the correct (or the better) choice, because it includes all the words of the NP chunk.



The big black bear roared.

Example 2: Word span of annotation example

Naturally, in cases of uncertainty with regard to the word span, the annotation referee is advised to be conservative and keep to smaller word spans and the label choice of the majority (if applicable).