I. Annotation approach

The first 4 rounds of annotation contained increasing number of examples (from 100 up to 400 a round), first annotated by all three annotators independently, with the follow-up meetings aimed at resolution of controversial cases and improvement of the annotation scheme. These rounds of annotation were used to verify applicability of the scheme to the dataset. Preliminary results (observed agreement and inter-annotator agreement measures) showed that the scheme can be applied to the data and we can reach a high level of agreement. Controversial cases were discussed in the meetings by the annotators and final agreement on the annotation was reached. These 4 rounds were then followed by individual annotation performed by each of the 3 annotators on a separate 1/3 of the remaining data.

II. Annotation types

The following annotation scheme is based on Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., and Smith, N. A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461. European Language Resources Association (ELRA), May.

1. **MW named entities**: *Chancellor of the Exchequer Gordon Brown*

- **Description**: This type covers references to (often unique) entities that exist in real life: people, events, companies, locations, etc. A good test for whether something is an MW named entity is availability of a Wikipedia page. In addition, particular scientific terms (e.g., *Formica fusca*) belong to this category, too.
- **Linguistic form**: NPs consisting of nouns or nouns linked with prepositions, with at least one noun being a proper noun.
- **Examples**: *Alawite sect, Formica fusca, Chapel of Nuestra Señora del Rosario*
- **Recommendation for simplification**: Provide a link to a description, ontology or encyclopedia page.

2. **MW compounds**: *red tape*, *motion picture*, *daddy longlegs*, *Bayes net*, *hot air balloon*, *skinny dip*, *trash talk*

- **Description**: Lexicalised expressions, meaning that the phrase has entered language (may be listed in the dictionary alongside single words) as a whole. A good test is whether the expression is already contained in a dictionary, has its own Wikipedia entry or can be illustrated with an image. Non-compositionality is another good test: e.g., "red tape" ≠ "red" + "tape". However, as discussed in the annotator meetings, there are strongly non-compositional expressions like "red tape" (+other idioms and metaphors) and there are much weaker ones like "higher proportion" which are much more compositional and the level of their degree of lexicalisation is questionable. We rely on

the following rule: 1) if it is useful for the reader to know of such a frequently occurring expression (e.g., “higher proportion”), and 2) if it the words in the expression co-occurring frequently, then we annotate the expression as an MWE compound. Inability to replace individual words within the expression without changing the meaning may or may not work: cf. “red tape” (cannot replace) and “higher proportion” (can replace for “bigger” with a negligible change). However, we agreed that expressions like “collapsed + property sector” are not MWEs because “property sector” may take many other modifiers apart from “collapsed” and there is nothing special in this word combination.

- **Linguistic form**: NPs consisting of nouns (*motion* + *picture*) or adjective + noun (*red* + *tape*), but also occasionally nouns + verbal adjectives (*life* + *threatening*). We’ve agreed that adverbs + verbs (*fatally* + *injured*, *independently verify, officially announced*) should be annotated as not MWE for consistency. **Note**: some examples in the dataset are used with the quotation marks – note that this shouldn’t be taken as evidence that an expression is an MWE. An expression should be judged on its own merit.
- **Examples**: *life threatening, fatally injured, property sector, financial cushion*
- **Recommendation for simplification**: Depending on the degree of non-compositionality, it may be possible to simplify the compound word-by-word (*higher –> bigger*), replace it with a different compound (*property sector –> housing market*), or replace it with a different word altogether (*red tape –> bureaucracy*).

3. **verb-particle**: *pick up*, *dry out*, *take over*, *cut short* -> **Merged** with “other phrasal verbs” = “verb-particle and other phrasal verbs”

- **Description**: This class of expressions is easy to identify – they consist of a verb and a particle. PoS-taggers should be capable of identifying these correctly.
- **Linguistic form**: Verb + particle
- **Examples**: *wind down, give in, give up.* **Note**: Because of the disagreement in the status of such words as *short* in *cut short* and because of the considerable overlap with the category of “other phrasal verbs”, the two were merged.
- **Recommendation for simplification**: Individually, neither “give” nor “in” may be complex, while together “*give in*” might be. We therefore need to identify them as a phrase and simplify them as a phrase, too: it’s possible that some verb+particle may be simplified another verb+particle expression, or we might need to find one word replacement as in *“give in” –> “surrender”.*

4. **verb-preposition**: *refer to*, *depend on*, *look for*, *prevent from*

- **Description**: This class of expressions is also easy to identify – they consist of a verb and a preposition. PoS-taggers should be capable of identifying these correctly.
- **Linguistic form**: Verb + preposition
- **Examples**: *laced with, billed as*
- **Recommendation for simplification**: Full MWE may be replaced with a simpler MWE of the same syntactic structure. Attention should be paid to grammatical constraints.

5. **verb-noun(-preposition)**: *pay attention (to)*, *go bananas*, *lose it*, *break a leg*, *make the most of*

- **Description**: This class covers expressions similar to MW compounds – it should be a frequent phrase that the reader will benefit from knowing about. The requirement is that it starts with (the grammatical head is) a verb. Such examples should cover typical selectional preferences: e.g., one would expect to see "surgery" after "underwent".
- **Linguistic form**: Verb + NP (+ preposition)
- **Examples**: *underwent surgery*
- **Recommendation for simplification**: Like MW compounds, these expressions will have varying degrees of non-compositionality: for fully non-compositional ones, we'll have to replace them as a whole (*"lose it" –> "become angry"*), while for others it might be possible to replace word-for-word (*"go bananas" –> "go crazy"*). Attention should be paid to grammatical constraints.

6. **support verb**: *make decisions*, *take breaks*, *take pictures*, *have fun*, *perform surgery*

- **Description**: The condition for this class is that the verb is one of the "light" verbs or "support verbs" – either one of the auxiliary ones (*"do", "have"*) or one that is frequently used to form various expressions (e.g., *"take", "make", "give"*).
- **Linguistic form**: Verb + NP
- **Examples**: *make clear*
- **Recommendation for simplification**: Most often, such expressions will require a full replacement: e.g., *"make decisions" –> "decide", "take breaks" –> "rest"*, etc.

7. **other phrasal verb**: *put up with*, *miss out (on)*, *get rid of*, *look forward to*, *run amok*, *cry foul*, *add insult to injury*, *make off with*. **Note**: The use of this category is deprecated: the examples given in the guidelines are quite idiomatic and don't show up in data too often; there is enough confusion between this and "verb+particle" category (e.g., in *set aside, stay put, turn around*); and there is not much benefit for the TS system in distinguishing between the two categories –> merged with "verb+particle".

8. **PP modifier**: *above board*, *beyond the pale*, *under the weather*, *at all*, *from time to time*, *in the nick of time*

- **Description**: Expressions that contain preposition as a grammatical head.
- **Linguistic form**: Preposition + NP
- **Examples**: *on board, at stake, without exception*
- **Recommendation for simplification**: Simplification may involve elaboration using a relative clause

9. **coordinated phrase**: *cut and dry*, *more or less*, *up and leave*

- **Description**: These phrases may be considered a subtype of fixed phrases – they are very frequent, widely used expressions, that contain coordination. Not any coordinated

phrase will be a fixed coordinated phrase (e.g., we decided not to annotate “repentant and rededicated” as this is not a fixed phrase).
- **Linguistic form**: PoS + coordinative conjunction + PoS (PoS may be noun, verb, adjective, adverb, etc. but it will be the same one on both sides of the conjunction)
- **Examples**: *shock and horror, duck and cover*
- **Recommendation for simplification**: Simplification would typically involve replacement of the whole MWE; additional explanation may need to be provided in case of fixed phrases.

10. **conjunction/connective**: *as well as*, *let alone*, *in spite of*, *on the face of it*/*on its face*

- **Description**: This group is mostly identifiable by its role in the sentence as a connector. Otherwise, it might consist of a various number of words or various parts-of-speech.
- **Linguistic form**: Various parts of speech, often involves adverbs (“*well*” in “*as well as*”) and prepositions (“*in*” and “*of*” in “*in spite of*”)
- **Examples**: *such that, to this end*
- **Recommendation for simplification**: These expressions cannot be simplified word-by-word; instead they may require syntactic rather than lexical simplification.

11. **semi-fixed VP**: *smack* <one>*’s lips*, *pick up where* <one> *left off*, *go over* <thing> *with a fine-tooth(ed) comb*, *take* <one>*’s time*, *draw* <oneself> *up to* <one>*’s full height*

- **Description**: These expressions are similar to other phrases with verb being the grammatical head. The definitive characteristic is that it is a frequently used phrase of unlimited length, as the direct or indirect objects depending on the verb may attach several modifiers: e.g. verb + noun = “*smack lips*” needs to attach modifiers e.g. verb + modifier + noun = “*smack ones’ lips*”. **Note**: We only put idiomatic non-compositional expressions into this category, and we treat other VPs as not MWEs.
- **Linguistic form**: Verb as a grammatical head with further dependents which may include multiple NPs with modifiers. Dependents may be direct or indirect objects.
- **Examples**: *offered his sympathies, flexed their muscles*
- **Recommendation for simplification**: Many of such expressions will be metaphorical like “*flex one’s muscles*”. While simplifying we need to know which constituents belong to this expression. Care should be taken when simplifying the phrase to ensure agreement with the non-fixed unit.

12. **fixed phrase**: *easy as pie*, *scared to death*, *go to hell in a handbasket*, *bring home the bacon*, *leave of absence*, *sense of humor*

- **Description**: Frequent, often non-compositional, expressions. Can be treated as a supertype for other types like coordinated phrases. Seems to include various types of linguistic forms, but some clear subtypes include NPs with prepositions (*leave of absence, sense of humour*), and simile (*easy as pie, quacks like a duck*). We’ve also agreed that foreign terms like *en route* and *et al* will by default be classified as fixed phrases.

- **Linguistic form**: Various. One subtype includes NPs with prepositions. Another subtype includes simile (*as*, *like*)
- **Examples**: *quacks like a duck, state of shock, the tide has turned*
- **Recommendation for simplification**: The simplification will depend on each individual case; in addition, in some cases the simplification might involve elaboration or explanation instead.

13. **phatic**: *You're welcome. Me neither!*

We didn't have examples for this type in the data.

14. **proverb**: *Beggars can't be choosers. The early bird gets the worm. To each his own. One man's* <thing1> *is another man's* <thing2>.

We didn't have examples for this type in the data.

15. **Not MWE:** All other examples where the annotators selected multiple unrelated words that don't form an expression. Examples include "*authorities should annul the*", "*IP address is blocked*", etc. **Note**: some examples in the dataset are used with the quotation marks – note that this shouldn't be taken as evidence that an expression is an MWE. An expression should be judged on its own merit.

16. **Not MWE but contains MWE(s):** This additional type is included in the annotation scheme because often the annotators selected a longer phrase that contains both free words and proper MWEs: for example, "*Clarinet Concerto and Clarinet Quintet*" is a combination of two MW named entities; "*collapsed property sector*" is a combination of a separate word and an MW compound, and so on. In such cases, the MWE sub-unit should be classified and simplified according to the categories above. **Note**: Care should be taken about incomplete MWEs: sometimes a phrase looks like an MWE (e.g., "Peter Doskozil") but it is actually part of a longer MWE (e.g., "Hans Peter Doskozil"). As we don't want to annotate parts of bigger phrases, we tag such cases as "Not MWE"