Validation of Corpus Pattern Analysis - Assigning pattern numbers to random verb samples

Silvie Cinkova, Patrick Hanks, October 2010

Motivation of the task

The annotation task described in this manual is designed to validate **Corpus Pattern Analysis (CPA)**, a new approach to the syntagmatic and semantic description of verbs in English and other languages. CPA is based on the observation that, although many words are very ambiguous, **patterns** of word use are only rarely ambiguous. CPA therefore seeks: 1) to identify the patterns of normal use for each word by analysis of actual usage, and 2) to associate meanings with patterns of word use, rather than with words in isolation.

Even a quick glance at a corpus shows that most uses of most words are surprisingly regular, falling into a comparatively small number of patterns. Human beings are creatures of habit. However, when describing these patterns, getting the details right is difficult, and there has to be a mechanism for dealing with unusual (but authentic) uses of words and relating them to normal uses. These mechanisms are provided by the **Theory of Norms and Exploitations** (TNE; Hanks 1994 and Forthcoming).

The first product of CPA is a *Pattern Dictionary of English Verbs* (PDEV; work in progress). This is intended as a basic infrastructure resource, which, in addition to being useful to human learners and teachers of English, will help tackle the problems of word sense disambiguation in computational natural language processing. While we, as human dictionary users, can intuitively appraise CPA as intelligible and likely to be helpful in language learning, we have not yet got any evidence that CPA is any good for machines. The goal of the annotation task described in this manual is to obtain evidence to support the assumption that CPA can mediate the vagueness of natural language to computers – or, alternatively, to deliver a methodologically sound proof that such evidence cannot be gathered in this way. If the corpus annotation undertaken in the present exercise is successful, the annotations and the patterns (revised where necessary) will together serve as a gold-standard data set for machine learning experiments in information retrieval and word-sense disambiguation.

Prototypes and patterns

TNE draws on Prototype Theory and projects it onto patterns of language use. Prototype Theory, as developed by cognitive linguists, explains major aspect of the structure of concepts in the mind (including beliefs about the conventional meanings of words and phrases). TNE relates prototypical meaning concepts to prototypes of phraseology (i.e. linguistic usage), as found in a large corpus. Corpus analysis shows there are not only prototypical uses of words (i.e. normal and conventional uses – **norms**) but also the ever-

present possibility of uses that, in one way or another, deviate from the prototypical patterns and yet are perfectly well-formed and well-motivated utterances. These are mostly creative innovations, but they include also domain-specific patterns. They are called **exploitations**. An exploitation is an utterance that can be related to a corresponding phraseological norm.

Current work in CPA focuses on analysing patterns of English verb use. There are 5756 verbs in the corpus that we are using, BNC50 (50 million words of the British National Corpus). These include all the verbs that are in normal use in English. At the time of writing, 700 of these have been analysed for PDEV. A standard sample for annotation consists of 250 corpus lines for each verb[1]. If more than 25 different patterns are found for a verb, the number of corpus lines in the sample is doubled; if more than 50, it is doubled again, and so on. Most verbs have only a few patterns, but ***take*** has over 200 distinctive patterns. In PDEV, phrasal verbs are treated as patterns of the base verb, not as separate entries. Thus *take off* is a pattern of the base verb *take*, contrasting with another pattern of the same base verb, *take something off*.

For most verbs, only a handful of patterns are frequent, while the rest are more or less rare. Rare patterns are often related in some way to frequent patterns; for example, a secondary pattern may be a conventional metaphor or an inchoative, resultative, or conative alternation.

Components of the verb patterns

Pattern structure

Each verb pattern in CPA is based on the structure of English clause roles described in systemic grammar (sometimes called 'slot-and-filler grammar'): see, for example, Halliday, *Categories of the Theory of Grammar* (1961). For technical reasons, generative grammar is not well suited to the empirical descriptive analysis of natural language. Nevertheless, with a little ingenuity, conversion of PDEV patterns into generative parse trees is possible. This works best if the basic logic of generative grammar is supplemented by selected features of Lexical Functional Grammar, in particular clause roles, which play a central role in CPA but are sadly neglected in generative grammar.

Clause roles in systemic grammar are:

> S – Subject (Agent in dependency grammar) – the semantic subject of the clause (omitted or introduced by the preposition *by* in passive realizations)
> P – Predicator (the verb, together with its auxiliaries if any)
> O – Object (direct or indirect; in CPA, 'direct object' includes the subject of passive sentences)

[1] For some verb, there are fewer than 250 occurrences of the verb lemma in BNC50. In these cases, PDEV tags all the lines available; it does not at present seek to supplement BNC50 with lines from other corpora.

C – Complement (a phrase that is coreferential either with the subject of the sentence, as in *He is happy*; *he is the President*, or with the direct object, as in *They elected him President*; *it made him happy*)
A – Adverbial (usually a prepositional phrase, a particle, or one of a small set of adverbs, as in *She drove to London, she drove home, she drove off*).

This is the basic framework underlying all PDEV patterns. Absence of a direct object can be part of a pattern, affecting its meaning, so it is stated explicitly: [NO OBJ]. On the other hand, absence of an adverbial does not normally affect the meaning and so it is not normally stated.

Each clause role in a pattern is 'populated' by a paradigm set of collocations—words that regularly occur in a particular clause role (or 'argument') in relation to a particular verb. The relevant collocations of a verb in a clause role are usually nouns that share some basic aspect of their meaning, which can be expressed as a **semantic type**. Semantic types are stored in a hierarchically structured **shallow ontology**. See below for further details.

Each verb pattern is accompanied by an **implicature**. For instance, one of the patterns of the verb *translate*, with its implicature, is as follows:

PATTERN: **[[Human]] translate ([[Document]]) (from [[Language 1]]) (into [[Language 2]])**
IMPLICATURE: [[Human]] expresses the meaning of [[Document]] in [[Language 1]] in the words and phraseology of [[Language 2]]

Thus, the pattern states regularly occurring contexts that co-determine its meaning. The meaning is paraphrased in the implicature, which is 'anchored' to the pattern by means of the semantic arguments, which are expressed in double square brackets.

The combination of pattern and implicature is called a category. A **verb entry** consists of one or more **categories**, each of which in turn consists of a pattern and an implicature.

Alternations

Most syntactic alternations (e.g. causative/inchoative, resultative, conative) are stated as separate patterns, if there is enough evidence to justify treating them at all. However, if only one or two occurrences of an alternation are found in a sample, they may be treated as **exploitations** of a normal pattern and tagged as syntactically anomalous uses (e.g. '1.s').

Major exceptions are: 1) the active/passive alternation, 2) the indirect object alternation, and 3) the reciprocal alternation.

1) Passive verbs are normally treated in the same pattern as their active counterparts. The predicator is stated in the passive form only when a particular verb usage is normally passive, so that the active form sounds unnatural (e.g. 'be bothered'). A modal verb and/or a negative particle may also form part of the predicator in a pattern (e.g. 'can't be bothered').

2) With verbs of giving, there is a regular alternation in English between a dual object construction (SPOO, as in *He gave her an apple*) and one with an object and adverbial (SPOA, as in *He gave an apple to her*). Such cases are treated in CPA as part of a single pattern, in which the alternating clause element is tagged as an "Indirect Object".

3) Another exception concerns reciprocal verbs. The alternations of reciprocal verb patterns are syntactically complex, with only slight differences of emphasis (e.g. *John met Alice; Alice met John; John and Alice met; John met with Alice; Alice met with John*). Significant realizations of reciprocity will be indicated in the pattern in the revised pattern editor currently in preparation.

Implicatures

Each pattern has at least one implicature. The implicature is a paraphrase of the pattern using a different verb and different phraseology. As far as possible, relevant arguments of the pattern are repeated in the implicature (see, for example, the two occurrences of [[Document]] in the ***translate*** example above). This has the effect of 'anchoring' the implicature to the pattern.

Sometimes the implicature is enriched with a **secondary implicature**. A secondary implicature adds complementary information to the primary implicature. In principle, there can be any number of secondary implicatures for a pattern, though in practice rarely more than one is stated.

> PATTERN: **[[Human | Action | Drug]] alleviate {pain | anxiety | illness | ...}**
> IMPLICATURE: [[{Human | Action | Drug]] causes {pain | anxiety | illness | ...} to become less intense
> SECONDARY IMPLICATURE: [[Human]] is typically a Health Professional

Collocations: semantic types and lexical sets

Semantic Types

Collocations that have a distinctive semantic feature in common are grouped together according to their **semantic type**. Semantic types represent cognitive concepts such as Human, Institution, Animal, Event, etc. They represent 'folk concepts' that play a central role in the way words are used. They owe little or nothing to scientific conceptualizations such as *mammal* or *animate*.

The semantic types are stated in a finite inventory, which constitutes a "shallow ontology" of about 200 items. The inventory is kept under review during the creation of new patterns. Occasionally, a new semantic type is added, so strictly speaking, the ontology will not be definitive until all verbs have been processed. However, revisions are rare and small in practice. Now that 700 verbs have been compiled, the current inventory of semantic types can already be efficiently used with confidence for both creating new patterns and for relating random corpus concordances to the already existing patterns.

The shallow ontology of CPA semantic types has as its top type [[Anything]]. Some verbs take literally anything – entity, event, or state; concrete or abstract; anything – as an argument.

> PATTERN: **[[Anything]] amaze [[Human]]**
> IMPLICATURE: [[Anything]] causes [[Human]] to be very surprised

However, most verbs have a distinct preference for a smaller set of lexical items in each clause role. The purpose of the semantic types in CPA is to make it possible to state the semantic preferences that determine the range of nouns and noun phrases that are normally found in a particular clause role. In the CPA shallow ontology, the top type [[Anything]] is divided into [[Entity]] and [[Eventuality]]. [[Eventuality]] is divided into [[Event]] and [[State]], and so on down each branch of the hierarchy to quite a delicate level of generalization, such as [[Road Vehicle]] or [[Musical Instrument]]. In addition there are two minor subparts to the ontology: [[Part]] and [[Property]]. As in other ontologies, each semantic type inherits the formal property of the type above it in the hierarchy. The CPA shallow ontology differs from other ontologies in that it is driven by the empirical needs of semantic analysis of corpus data. It makes no attempt to show how modern scientific terminology represents entities and events in the world. Thus, it is curiously unbalanced. For example, there are senses of verbs such as ***bark*** or ***saddle*** that expect [[Dog]] or [[Horse]] in a particular clause role. There are many words and names that denote horses and dogs, so [[Dog]] and [[Horse]] must be recognized as semantic types. On the other hand, in general English there are no verbs that require a distinction between jackals and hyenas, so these are not semantic types.

When two or more arguments[2] have the same semantic type, they are distinguished by numbers.

> PATTERN: **[[Human 1 | Animal 1 | Institution 1 | Document]] signal ([[Human 2 | Animal 2 | Institution 2]]) [[Information | Eventuality]]**
> IMPLICATURE: [[Human 1 | Animal 2 | Institution 1 | Document]] communicates [[Information]] about [[Eventuality]] (to [[Human 2 | Animal 2 | Institution 2]]) by gestures, language, or other means.

[2] The terms **argument** and **argument slot** are used here informally as approximate equivalents to 'clause role (of a particular verb pattern)'.

This pattern also shows that an argument may be realized by any of several alternating semantic types.

A companion document, *Guidelines for Applying Semantic Types in CPA*, is currently in preparation. At the same time, a few of the types that were coined in the early stages of the project (e.g. [[Abstract]]) are being reviewed.

Lexical sets

The number of collocates that normally populate an argument slot in relation to a particular verb varies greatly. Some verb patterns admit virtually any noun phrase as a collocate in an argument slot; others take only a very small set of lexical items as normal collocates in that slot. If a suitable semantic type is not available at an appropriate level of delicacy, the salient collocates may simply be listed, thus:

> PATTERN: **[[Human | Action | Drug]] alleviate {pain | anxiety | illness | ...}**
> IMPLICATURE: [[Human | Action | Drug]] causes {pain | anxiety | illness | ...} to become less intense

Lexical sets are grouped together inside curly brackets. In CPA, curly brackets are used to group things together and to identify a lexical item that fulfills a clause role; they have no other special significance.

A lexical set may consist of only one word (a single lexical item). This is usually but not necessarily true of idioms and light verbs, for example pattern 29 of ***take***:

> PATTERN: **[[Human]] take {responsibility} for [[Anything]]**
> IMPLICATURE: [[Human]] accepts the duty of doing whatever is necessary to ensure that [[Anything]] is OK
> SECONDARY IMPLICATURE: If [[Anything]] turns out to be bad or not OK, [[Human]] may be fired or otherwise punished

Some lexical sets are bigger than others. The pattern manager lists only the two or three most salient items in a lexical set. If there are more items in the lexical set, the pattern manager indicates this by three dots.

> PATTERN: **[[Human]] take {risk | chance | ...}**
> IMPLICATURE: [[Human]] does something that may have bad consequences, in the hope of obtaining some [[Benefit]]

Here, the comparatively rare lexical alternation ‘take a gamble’ has not been stated explicitly by the pattern manager.

A lexical set may alternate with a Semantic Type:

PATTERN: **[[Human | Animal]] bleed [NO OBJ] {from [[Body Part]] | from {wound | injury | laceration}}**
IMPLICATURE: [[Human | Animal]] loses blood from a wound or injured [[Body Part]]

Semantic roles

The semantic type of an argument is sometimes complemented with a **semantic role**. The semantic type captures the 'formal' quale of the argument, which is an intrinsic property of nouns normally found in that argument slot. On the other hand, a semantic role captures what may be an extrinsic property of the nouns in the same slot, namely one that is assigned to them in context even if it is not an intrinsic property. Compare the two (invented) example sentences below:

Mr. Woods sentenced Bailey to 3 years.
The judge sentenced the old villain to a term of imprisonment.

In the first example, the semantic roles of the arguments ('judge', 'criminal', 'punishment') are implied, i.e. they are assigned by context; in the second, they are realized explicitly. The meaning of the verb is the same in both sentences. The pattern in both cases is:

PATTERN: **[[Human 1 = Judge]] sentence [[Human 2 = Criminal]] to [[Event = Punishment]]**

The semantic role can also capture the 'semantic prosody' of an argument, e.g. good or bad. Thus, semantic roles can have the form of adjectives; e.g. *Bad, Valuable, Sad*. Any noun, adjective, or noun phrase may be stated as the semantic role of an argument, if it captures a generalization that is true in context.

Like semantic types, semantic roles allow alternations, thus:

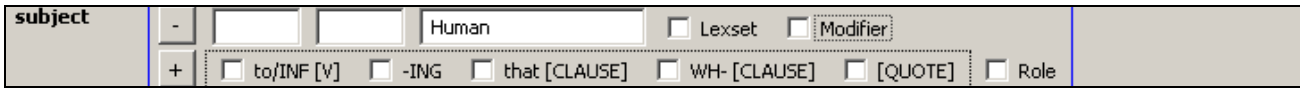
PATTERN: **[[Human]] construct [[Artifact | {Route = Road | Canal | Railway} | Building]]**
IMPLICATURE: [[Human]] creates [[Artifact | {Route = Road | Canal | Railway} | Building]] by putting together several diverse components

Verb arguments

Semantic types, semantic roles, and lexical sets are representations of nouns or noun phrases. CPA also captures verb complements, such as infinitives, that-clauses, and wh-

clauses, as well as *-ing* clauses and direct speech quotes for each semantic type. They are described by their syntactic form:

PATTERN: **[[Human | Document | Institution]] acknowledge {that-CLAUSE}**



Optional arguments

Round brackets indicate that the given argument is optional; i.e. there are regularly occurring concordances in the corpus in which the given argument is omitted and the omission does not affect the implicature.

Semantic type ambiguity

Sometimes it is impossible to decide which semantic type to assign and, at the same time, the choice of semantic type determines the choice of the implicature. E.g.:

The AAA launched their education programs.
Pattern 1: begin or initiate an endeavour
Pattern 2: begin to produce or distribute; start a company

In this case, is a *program* an [[Event]] or is it a [[Product]]? Event implies Pattern 1, Product implies Pattern 2. In such cases, preference is given to the more frequent of two competing patterns.

The task: assigning concordance lines to patterns

Introduction

The patterns have been compiled on the basis of corpus evidence. During the compilation of verb patterns, lexicographers sorted concordances by assigning pattern numbers to them as tags on the basis of perceived similarity of pattern meaning. The tags were assigned partly using a random sample of concordances and partly using concordances pre-sorted by the Sketch Engine. CPA requires that, as far as possible, every line in a selected random sample (the number of corpus lines being declared in the pattern editor) must be assigned to a pattern or declared to be a tagging error or unclassifiable. The random sample is the basis for comparative frequencies (expressed as percentages in the pattern editor), not all tagged corpus lines.

Procedure

The annotator working on the validation task (henceforth, the 'validator') will not see the entire original lexicographer's annotation, the **reference random sample (RRS).** This RRS will be divided into two parts: S1 (200 concordances) and S2 (50 concordances). The validator can refer to S1 for guidance. The validator will receive S2 without tags. The validator's task is to tag each line in S2 with the appropriate pattern number.

Inter-annotator disagreements will be analysed by SC and PWH. When necessary, the patterns will be revised and a new 50-concordance sample will be generated for validators to tag it according to the revised patterns and PWH will revise the original RRS created by him according to the improved patterns.

Norm-conformant concordances

A norm-conformant concordance line will receive the number of the corresponding pattern. A concordance line is norm-conformant when:

1) it has the same implicature as the pattern;
2) it has the same number of arguments as the pattern presents[3];
3) prepositional arguments match the prepositions in the pattern,
4) nouns conform to the listed semantic types[4]
5) the arguments get the same semantic role
6) the non-noun arguments match the form description defined by the pattern (e.g. that-CLAUSE)

Grammatical ellipsis

In some cases, the subject of the verb under observation is missing for grammatical reasons. This happens in the following cases:

Imperative (2nd person): ***Speak*** *to your friend!*
Verb control: *Peter decided to* ***speak*****.**
Phased predicators: *Peter ceased to* ***speak*** *to his friend*. *Peter stopped* ***speaking*** *to his friend*.

These cases are still regarded as norm-conformant.

[3] Grammatical ellipsis of subject in imperative and verb control (called 'phased predicators' in systemic grammar, e.g. '*she intended to treat him well'*) still count as norm-conformant use. For more detail see the following section, on ***Grammatical ellipsis***.
[4] This rule has one exception, namely the semantic coercion of a noun and its modifier; e.g. [[Drink]] => [[Container]] of a [[Drink]]. See Section ***Semantic Type Coercion between a verb argument and its modifier*** for more detail.

Contextual clues, contextual ellipsis

Any contextual as well as common-knowledge clues within the scope of the surrounding text necessary for good understanding of the concordance are legitimate to use. The contextual hints can be used

1) to resolve contextual ellipsis
2) to specify the domain or register and disambiguate semantic types/roles

For instance, it is legitimate to use the knowledge that a given person name denotes a politician and associate the sentence

Mrs. Thatcher abstained again

with the correct pattern *abstain from a vote*.

Another example:

In ***ride***, both a horse and a bicycle can be ridden; they are treated as different patterns. The sentence *I will ride* is therefore ambiguous between two patterns. However, if the word *bicycle* is present in the wider context, a sentence such as: *I will ride and you will follow me* inevitably triggers the bicycle pattern, even though the default in the ellipsis of this argument is normally the horse frame (where [[Horse]] is marked as optional).

It is also completely acceptable to have a look at the source document information if it helps to decide on the relevant pattern.

However, we would like to discourage the validators from spending time on acquiring encyclopedic information in an attempt to resolve unintelligible concordances. If the validator is not able to assign a pattern number reasonably quickly using the context, the origin of the source document, and world knowledge, then the unintelligible concordance line should be marked as as ".u".

Pronouns as arguments

In many cases, an argument is rendered by a pronoun. Personal and relative pronouns are usually anaphoric, whose antecedents are easily traced in the context. When assessing whether or not a concordance matches a given pattern, we always consider the semantics of the antecedent of a pronoun as well as that of the pronoun itself.

that-clauses

The subordinating conjunction *that* introducing a noun clause is often omitted in English. This is often a source of confusion for unwary analysts. Both the following are examples of a verb with a that-clause:

I told them that they should stay.
I told them they should stay.

Direct speech and indirect speech

With reporting verbs, direct speech is tagged as [QUOTE], even if quotation marks are not present. Note that the reporting verb may be at the beginning, the end, or in the middle of the direct speech; also the normal order subject and verb is sometimes inverted.

"I won't do it", said Mr. Smith, "unless you accept my conditions." – [QUOTE]
Mr. Smith said, "I won't do it unless you accept my conditions." – [QUOTE]
"I won't do it unless you accept my conditions," said Mr. Smith. – [QUOTE]
I won't do it, Mr. Smith said, unless you accept my conditions. – [QUOTE]

A clause not enclosed by quotation marks counts as a *that*-CLAUSE whether or not it contains *that*, unless its subject is in the first person, even if the reporting verb is in the middle or at the end.

Mr. Smith says he won't do it unless they accept his conditions. – that-CLAUSE
He won't do it, Mr. Smith says, unless they accept his conditions. – that-CLAUSE
He wouldn't do it, Mr. Smith said, unless they accepted his conditions. – that-CLAUSE
He won't do it unless they accept his conditions, Mr. Smith said. – that-CLAUSE

Semantic Type Coercion between a verb argument and its modifier

People typically drink a [[Beverage]], but we can also talk about drinking a cup, a glass, or a bottle of a beverage. Cups, glasses, and bottles have the semantic type [[Container]], but in case of *drink*, they do not denote containers, but an amount of a beverage.

She drank 8 glasses of spirits

therefore matches the pattern "[[Human]] drink [[Beverage]]", even though there is no mention in the pattern of [[Container]] in the pattern.

The semantic head of a noun phrase containing the preposition *of* may be to the left or the right of the preposition, so this sentence can be analysed as "she drank spirits"; spirits is a

central and typical member of the lexical set [[Beverage]]. In such cases, the noun before *of* has a partitive or quantifying function. A partitive example is the noun *slice*.

He ate four slices of toast

matches the pattern "[[Human]] eat [[Food]]" because *toast* can be parsed as the semantic head of the direct object.

However, this analysis holds even for the following sentences, which exhibit ellipsis as well as coercion:

She drank 8 glasses
She drank 8 pints
He ate four slices.

Because glasses and pints are regularly used as partitives for beverages and slices for certain kinds of food, these sentences can be marked as regular pattern-conformant uses. There is no need to tag them as exploitations.

Another examples of coercion is:

*the 28-year-old stockbroker was **riding** his first ever winner.*

Here *winner* = [[Horse]]. Corpus evidence shows that this is a regular (domain-specific, horse-racing) meaning of *winner*, so it does not need to be treated as an exploitation. Instead, it will simply be one of the words that populate the semantic type [[Horse]]. Words are polysemous, so there is no limit to the number of semantic types that a noun may have.

Semantic Type Coercion between a verb and its argument

Coercion can also occur one "level up", compared to the case above. We can say things like:

He enjoyed his soup.

This is actually a context-dependent semantic shortcut, where the actual event is underspecified. The sentence literally means:

He did something to the soup with enjoyment.

The actual predicate is to be inferred from the context. Even without context we can usually guess a default event which is assumed when no counter indexes are found, as, in this particular case, eating the soup. See the following examples:

I enjoyed the soup.

= (default) *I ate the soup with enjoyment. I enjoyed eating the soup.*
= (said by a cook) *I cooked the soup with enjoyment. I enjoyed cooking the soup*.

I enjoyed the performance.
= (said by a non-artist) *I watched the performance with enjoyment. I enjoyed watching the performance.*
= (said by an artist) *I made the performance with enjoyment. I enjoyed making the performance.*

The coercion in verb as it is described in this section is to be marked with **.c.**

Noun uses that don't fit

If the validator finds a concordance line containing a word that does not fit any of the semantic types in the patterns:

a) if the meaning is clearly one that is covered by the implicature of an existing pattern, mark it as "**.a**" for 'anomalous argument'. If many such lines are found for any given pattern, the lexicographer will consider whether to change the scope or focus of the pattern.

b) if the concordance line does not fit any pattern or if its meaning is not clear, it should be marked as "**.u**" ("unclassifiable"). By doing this, the validator makes an explicit indication that there is no appropriate pattern for this concordance. This means either that the validator is not able to understand the context well enough to associate it with a given pattern or that the validator is sure that none of the patterns available matches this use.

c) if the noun suggests a verb coercion, assign "**.c**". Verb coercion means in short that the noun would prefer a different predicator, which is implicit from common knowledge or the context, and the present predicator adds some complementary semantic features to the event (see Section ***Semantic Type Coercion between a verb and its argument***)

Exploitations

Exploitations of normal uses are marked by adding .a, .f, .c or .s to the pattern number.

Anomalous argument (.a)

The mark **".a"** indicates an '**anomalous argument'** or '**honorary member'** of a lexical set. By 'honorary member' we understand a noun that is an unusual, unexpected member of the relevant lexical set and does not have the right semantic type, although the meaning clearly is that stated in the implicature. An example is the following corpus extract:

*Rashid Solh, the prime minister, spoke darkly of the arrival in Lebanon of several hundred Israeli agents provocateurs whose mission was to destroy the republic. The plot had **arrived** at Beirut.*

The meaning of the second sentence is clear, and indeed it almost matches the following pattern:

[[Human | Vehicle | Animal]] arrive [NO OBJ] {at [[Location]]}

However, a plot is neither a Human not a Vehicle nor an Animal. Moreover, it is very unusual to talk of a plot 'arriving' at a place. This not what plots do (normally). So it is an anomalous argument – an unexpected argument. It is not a metaphor – a figurative usage, because "the plot" (whatever it may have been) was something that clearly arrived in a physical location.

Figurative uses (.f)

Conventional metaphors form patterns in their own right. For example, the following sentence is clearly a metaphorical use of the verb *arrive*:

*Jane and I quickly **arrived** at joint decisions about the project*

However, this matches pattern 2 "[[Human | Institution]] arrive [NO OBJ] {at [[Concept = Considered Opinion]]}". It is conventional and therefore a pattern.

In cases of non-conventional metaphor, we mark concordances as pattern number + ".f"

*Nobody **arrives** at ICI board level without some steel and determination in his character.*

This is ".f" (figurative) because the arrival is one that happens in terms of someone's career development, not in terms of physical movement to a place. "Board level" in a large company is not a [[Location]].

A high number of "f"-concordances in the sample may signal that a secondary pattern has been missed. If so, it will be added during the pattern revision stage.

Unusual syntax (.s)

In ordinary discourse, writers and speakers sometimes omit a word. The omission affects the focus rather than the meaning.

*We **punish** too much—and in particular, we **imprison** too much.*
—(BNC) J. Dignan, 1992. *The Penal System.*

Here the implied objects of punishing and imprisoning are "people" or "anyone" or, implicitly from the context, "anyone who has broken the existing laws". These are syntactic exploitations: intransitive uses of verbs that are normally transitive. Both will be tagged as "1.s".

Not only the phraseology associated with particular lexical items, but also whole constructions, may be exploited.

"I would also like to apologize about losing you and Ema the house."
—Marian Keyes, *The Other Side of the Story*, p. 632.

The background is that, in the novel, the speaker and her boyfriend bought a house together but then failed to generate enough income to keep up the mortgage payments, so they 'lost' the house—it was repossessed.

Here the verb *lose* is being used ditransitively, which is not one of its normal patterns. It is coerced into being a ditransitive verb because it is a being used as a complementary antonym of ***give.*** If you can give someone a house, you can also lose them a house. Exploitation of grammatical constructions is less common than exploitation of phraseological patterns.

Tagging errors and other noise (x)

For the CPA validation task it is important to tag only genuine verb uses. There are some misprints in the corpus and tagging errors in the Sketch Engine grammar. Because of these problems, it occasionally happens that some noun and adjective uses of a word are presented in the concordances for a verb. These should be tagged as "x".

It is also important to make a distinction between a real use of a verb for some communicative purpose and a mere **mention** of the verb. These too should be marked with 'x'. Examples are the following:

—"*Now let us look at idioms such as 'grasp the nettle'."*
This is a mention, not a use, of the verb ***grasp***. Tag it as "x".

—"*Frankie goes to Hollywood*" is the name of a band, not a use of ***go***. Tag it as "x".

In contexts where a use of the past or present participle of a verb borders on being an adjective or noun, it can be difficult to decide whether a concordance should be marked as an example of a pattern or as "x".

- A rule of the thumb is: when an occurrence of a participial form can be reasonably matched with an existing verb pattern, it should be interpreted as a verb occurrence.

There are, however, a few "hard rules" for sorting away a noun or adjective use and tagging them as "x":

- When an *–ing* form is preceded by a determiner it is always regarded as a **noun**: *Which/some/no/any **reading.*** This applies even for the passive construction *there was (no) **–ing*** in the sense *(No) **–ing** took place/could take place.*

- When an *–ing* form of a lexical verb is preceded by a possessive determiner (possessive pronoun, genitive), it is a **noun**. E.g.: *I hate his **coming** late*. We treat this as a noun use of ***coming***. Compare *I hate him **coming** late*, which we treat as a verb use of ***come***.

- When an *–ing* form takes a direct object and is not preceded by a determiner, it is a verb: ***building** bridges.* On the other hand, when an *–ing* form renders the direct object with the prepositional phrase ***of***, it is a **noun**: ***building** of bridges.*

- A participle that is frequently used in the attributive position with a noun is an adjective, e.g.: *surprising evidence, an interesting article, a wicked guy.*

Unclassifiables (u)

Unclassifiables, marked with "u", are definitely verb uses of the verb being tagged, but they do not even come close to matching any of the normal patterns. In case of strong doubt, it is better to mark a concordance as "u" rather than to assign it to a pattern where it does not fit.