Methodological aspects of Corpus Pattern Analysis

Ismail El Maarouf, RIILP, University of Wolverhampton

1 Introduction

The work presented in this article is set in the DVC (Disambiguation of Verbs by Collocations) project, which looks more deeply into the role of collocations in Corpus Pattern Analysis (CPA; Hanks 2004; Hanks and Pustejovsky 2005; Hanks 2013). CPA is in the tradition of Corpus Linguistics, especially of Sinclair's lexical analysis (Sinclair 1991, 1998, 2004) and Hunston and Francis's Pattern Grammar (Hunston and Francis 2000). One of the goals of the DVC project is to build a pattern dictionary of 3,000 verbs (PDEV) following the CPA methodology. Given this new context, the whole enterprise is being revisited with the cooperation of lexicographers and computational linguists, from the relevance and organization of resources, to lexicographic principles and practices.

This article analyses methodological aspects of CPA. CPA is a lexicographical technique for dictionary building, mainly applied to verbs (but see Hanks 2013): it uses a rich paraphernalia of categories and structures to create and represent dictionary entries. Such entries are 'burned' onto texts thanks to a numerical editing software. The goal of this article is to inspect the CPA methodology through various sets of issues in corpus analysis. Such an inspection is seen as a step in the formalisation process of the CPA technique, in order to identify and circumscribe areas of fuzzy decisions. Formalisation is desirable for several reasons:

- It should make the technique more easily replicable by other human beings (lexicographers and corpus linguists).
- It should provide a guideline, and act as a referee in difficult cases.
- It may pave the way for automatic analyses in computational linguistics, and make it easier to interpret systems' errors.

It should be noted that the words *formal*, *formalise* and *formalisation* have suffered an unfortunate fate in linguistics and are sometimes regarded with suspicion by corpus linguists. I believe however that formalisation is a fundamental

notion in practical corpus linguistics. To avoid any misunderstanding, formalisation will be broadly defined as the process whereby practices and concepts are elicited, made explicit, analysed, checked for coherence, modelled, and verbalised.

An important issue regarding the formalisation of lexicographic practice is the notion of relevance, which springs up now and then in different contexts. The global process of pattern creation can be roughly compared to a set of progressive choices by lexicographers in performing what they consider to be relevant splits in the sample corpus: they cluster concordances according to perceived similarities. Relevance is thus not exclusively based on intuition, but driven by corpus evidence.

The succession and variety of choices differ from one verb to another because verbs have different types of patterns, and thus different types of clues. As a consequence, the lexicographer may not always build and design an entry in the same way for every verb (and may not start with the same kinds of clues, for example). This is one of the reasons why the choices which are made for a given verb may not be transferable as generic rules onto other verbs. This does not prevent us from imagining different lexicographer profiles which depend on different concordance configurations. For example, some verb concordances have a heavy use of particles (as *burn away/off/out*), while others have a predominant use of 'that-clause' (*to say*), and so on.

The article will illustrate important issues in CPA through the analysis of the verb *to burn*. This verb shows a variety of interesting uses and has a fairly high frequency in the BNC: 5069 occurrences in 100 million tokens. The article first reminds the reader of the goals and models of CPA. It then tackles the three following issues: Semantic Types, Syntactic Alternations, and Exploitations of Norms. This order of presentation is intended to allow the reader to gradually follow the different steps in terms of complexity of analysis. For example, Semantic Types is better dealt before proceeding to Syntactic alternations, which involve Semantic Types in different pattern positions. Therefore, pattern examples have been selected to illustrate the order of presentation, rather than the order of the dictionary entry.

2 Introduction to CPA

2.1 Goals and model

Perhaps the most important notion in CPA and indeed in Corpus Linguistics in general is that of 'pattern'. Patterns are recurrent textual sequences around a key word. There is more to it than that. First, as Sinclair used to point out, structure

and meaning are closely intertwined. Therefore recurrence should be probed for possible sub-patterns onto which meaning could be mapped. A pattern is some sort of idiomatic construction, which may cover a long stretch of text (e.g. 'naked eye', Sinclair 2004: 24–48), the elements of which may be more or less fixed. It is something that is on a higher level than surface text: it is a representation of a cluster of similar occurrences.

A particular problem in corpus-driven lexicography is lexical variation; e.g. in the case of idioms, one may *grasp* at straws or *clutch* at straws, while in the case of nouns, there is often a very large cluster of lexical items that can activate the same semantic value for a verb (e.g. one may grasp at the *bed posts*, but one may equally well grasp at a *railing* or a *mantelshelf* or any of a very large set of other physical objects – always activating the same meaning of the verb *grasp,* in contrast to (say), grasping at *an idea* or *an opportunity*). This simple fact of lexical variation goes a long way to explain why pattern representation is problematic.

Cluster construction and pattern representation are two different tasks in Corpus Linguistics. Cluster construction involves the careful observation of similarities between concordances in order to build a coherent set going beyond surface form similarity. Pattern representation involves the identification and/or creation of suitable categories, relations, and structures circumscribing as precisely as possible the features of the cluster.

CPA uses several sets of categories to model patterns: Forms, Part Of Speech, Sub-valencies, Semantic Types and Contextual Roles. Since the actual goal is to model verb patterns, CPA also relies on the SPOCA (Subject/Predicator/Object/Complement/Adverbial) model derived from Halliday's Systemic Grammar (Halliday and Matthiessen 2004) rather than on Generative Grammar phrase structures. Patterns can therefore be considered as ordered propositional units involving syntactic relations. In practice, other devices are used such as disjunctive bars for position alternatives or parentheses for optionality, so that they are also similar to Regular Expressions, as used in programming. Every pattern is associated with an implicature, providing a definition in natural language that is 'anchored' to semantic types in the pattern (i.e. it reformulates the text around semantic types in a sentence). Other features include voice (typically active, typically passive, etc.), domain, register or idiomatic comments, FrameNet links, positive/negative statement of a position.

The corpus-driven perspective of CPA entails that the pattern dictionary should first and foremost be representative of the corpus under study. The corpus used in CPA is a sample of the British National Corpus, containing only written data randomly sampled and reduced by half to 50 million tokens. The

BNC contains a wide variety of informative domains, such as *science*, *arts* or *world affairs* in a time frame of about 30 years from 1960 to 1993. The dictionary will thus be representative of this period and of these domains, as well as of other dimensions of corpus variation (Biber 1993). Further work includes investigating pattern variation across various domains, genres, registers, and geographical origin.

2.2 Editing environment

A CPA verb entry consists of a list of numbered patterns ordered by the lexicographer and for which proportionate frequency in the sample is given. Table 1 shows the most frequent pattern of the verb *to burn* (for the CPA entry of *to burn*, see Appendix):

Table 1: Pattern 1 of the verb *burn*

1	19%	**[[Human]] burn [[Physical Object | Building]]**
		[[Human]] sets fire to [[Physical Object | Building]] in order to destroy it

A pattern (first row) is always linked to an implicature (second row) explaining the meaning of the pattern in natural language using the semantic types (inside double square brackets). The pattern is ordered, with subjects on the left of the verb and objects on its right. Any pattern position may contain alternative semantic types (disjunctive bars). Cases where semantic types alternate are discussed in Section 2.

The dictionary is linked to an interface called the Sketch Engine (Kilgarriff *et al*. 2004). The Sketch Engine features corpus concordancing: the lines in which a keyword occurs can be sorted right or left, filtered according to specific forms or part-of-speech tags. When the lexicographer has identified a pattern, he/she annotates the corresponding lines with a pattern number. This information is automatically saved and linked to the dictionary. Figure 1 shows an example of concordances for Pattern 1:

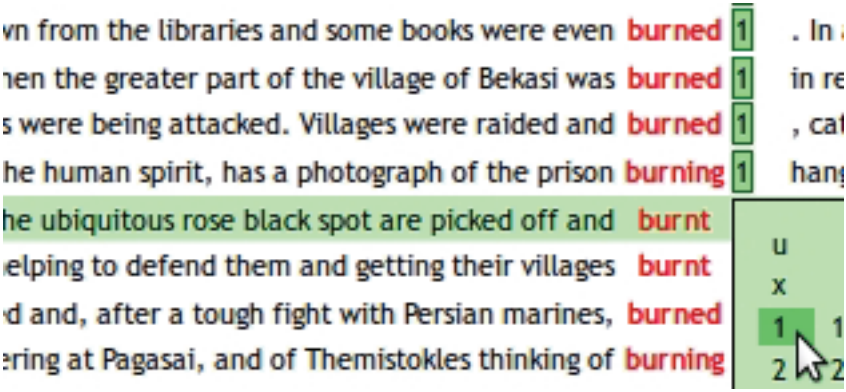


Figure 1: Concordances for Pattern 1

The Sketch Engine also helps the lexicographer to quickly spot a keyword's salient collocates: a parser filters surrounding words according to their syntactic relation to the keyword. Figure 2 shows the most salient words which occur in subject position:

subject	441	4.2
stubble x(57%)	7	3.39
straw x	14	3.34
candle	6	3.05
charcoal	4	2.98
flame	5	2.85
fire	30	2.83
flag x	7	2.82
Rome	6	2.54
fuel	8	2.47
light	11	2.29
eye	7	1.99
oil	4	1.9
house 9(91%)	12	1.67
station	4	1.65

Figure 2: Subjects for burn found by the Sketch Engine

There are boxes for every grammatical relation implemented in the grammar. Each box displays a collocate, followed by its frequency and salience measure in each row. For instance, *candle* was found six times as a subject and its salience is 3.05. A green box (grey in the picture) next to each collocate indicates the pattern number to which it is linked and the proportion of its use in this pattern. It is possible to assign a pattern to all the instances of a collocate thanks to this feature. However, this should not prevent the lexicographer from checking the corresponding concordances. As can be seen, a collocate may not always instantiate one pattern (percentages are not 100%), for various reasons, such as parser errors.

In Figure 2, x stands for non-verbal use, such as nominalisation as in *straw burning*.

(1) Straw *burning* for energy also has great potential.

These boxes help the lexicographer to draw a first 'sketch' of relevant semantic types, based on the similarities shared by the collocates. For example, I can think of the clusters {*fire*, *flame*}, {*light*, *candle*}, or {*charcoal*, *fuel*, *oil*}, illustrated in the following examples:

(2) The **fire** was *burning* steadily, the flame of the oil-lamp dipped and bobbed and the sweet smell of incense filled the air.

(3) Tall **candles** *burned* in the blacked-out windows.

(4) Some 800,000 gallons (3.5 million litres) of **oil** and petrol *burned* fiercely.

The CPA lexicographic environment is a very handy tool which ties the dictionary to the corpus.

2.3 Limitations

CPA is still at an early stage and, though the environment is rich in terms of information clues, improvements are planned in the near future.

First, once defined, a pattern cannot be decomposed in order to access parts of it: it is a flat 1-level structure, mapped onto several instances. Thus, a lexicographer may have to create two patterns which are very similar. It might be interesting to investigate the benefits of different representational structures, such as trees, to create and visualise patterns.

The consequential second limitation is that patterns are therefore not interconnected to each other: the only encoded pattern structure is ranking (by frequency for example). It would be interesting to investigate the kind of relations

existing between patterns (collocational, syntactic, semantic, and so on) because they could be used for structuring a verb's entry, bearing in mind the user's interest.

Another limitation is that pattern elements (subjects, objects, etc.) cannot be independently labelled in the corpus: the only possible link between the entry and the corpus ties a pattern to a verb occurrence. It would be desirable to specify in more details how pattern elements are realised, because it would provide more clues concerning the choices of the lexicographer. If this feature were enabled, it would also open the possibility for the lexicographer to add specific comments onto pattern elements.

Finally, general comments on verb uses in the concordance are limited. The lexicographer may only record whether an occurrence is an instance of a norm, or if it is an exploitation: exploitations cover anomalous syntactic structure, anomalous semantic argument or figurative uses. Hanks proposes a much larger typology of exploitations, including *irony*, *zeugma* and *hyperbole* (Hanks 2013). It would be interesting to include them in future versions of CPA since they are, at present, not annotated. Other types of exploitations may also spring up from further verb explorations, so the lexicographer should be allowed to enhance his/her own corpus-driven typology of exploitations. More generally, the lexicographer may also feel the need to add comments, which are actually only available at the level of the pattern (not of the instance).

This said, the dictionary and its editing interface are still at an early stage and it is not yet possible to measure the impact of these limitations on pattern construction. These limitations may also have a positive impact on the creativity of the lexicographer.

An important question to be answered in the DVC project, is whether two different lexicographers using this environment would agree on the identification of patterns and on the annotation. The next sections will look at possible areas of confusion or disagreement.

Previous experiments on Inter-Judge Agreement in CPA annotation were globally promising (Cinkova *et al*. 2012). Most disagreements observed were accidental errors due to lack of detail in guidelines, or confusions between norm instance and norm exploitation. But the judges generally agreed on the choice of the pattern. These experiments justify a detailed analysis of difficult areas, so as to provide a solid basis in preparation for the larger annotation campaign of the DVC project (3,000 verbs).

3 Relevance with semantic types

A crucial component of CPA is its emphasis on massive usage of semantic categories, called semantic types. The list of these types has been progressively compiled and they are organised in a shallow ontology. These types (e.g. [[Human]], [[Building]], [[Process]], etc.) refer to properties shared by a number of entities, the referring words of which are regularly found to participate in several pattern positions.

3.1 From lexical items to semantic types

The most 'cognitively salient' noun related to the verb *to burn* is *fire*. Hanks makes an interesting distinction between 'cognitive salience' and 'social salience' (Hanks 2013: 21). He suggests that

> as far as the lexicon is concerned, social salience (in the form of frequency of use) and cognitive salience (in the form of ease of recall) are independent variables, or perhaps even bear an inverse relationship: that is the more frequently a lexical item is used, the harder it becomes to call to mind and talk explicitly about all the normal uses of it.

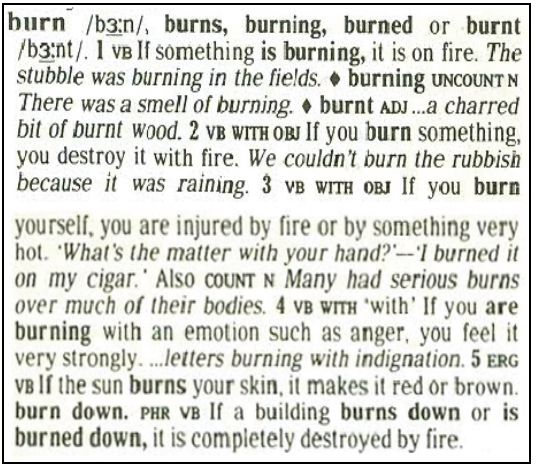
Figure 2 shows that the noun *fire* could both be cognitively and socially salient as a collocate (in subject position) of the verb *burn*, since it is both the most frequent subject, and is cognitively related to the verb *to burn*. However, the CPA pattern this word is used in only accounts for six per cent of the data. This can be explained in the following way: *fire* is similar to very few other collocates (*flame*, *bonfire*, etc.) and therefore constitutes almost the only member of its semantic type. On the other hand, other semantic types (such as [[Human]]) accumulate a much wider range of collocates, and are therefore more frequently involved in patterns. Word sketches may help to spot patterns very quickly, but be misleading for norm identification. This is because word sketches are based on lemmas, and not on semantic types.

Part of the concordances for the word *fire* as subject are illustrated in Figure 3:

of the Church. The **fire**,	*burning*	its way through the structures of
the air. The **fire** was	*burning*	almost smokelessly, with just a thin trail
old ash. The **fire** was	*burning*	steadily, the flame of the oil-lamp dipped
the house the **fire** was	*burning*	and a couple of sticks of jharo , resting
of the few **fires** still	*burning*	. Kalchu shut the door behind them.
us, reflecting the **fires** that	*burned*	in every grate.
Nov. 6 The **fires** had	*burned*	for approximately eight months,
been capped. The **fires** had	*burned*	for approximately eight months,
in the evening as **fires**	*burned*	throughout the city,
Aristotle noticed long ago: ' **Fire**	*burns*	both in Hellas and in Persia;
up the **fire** which already	*burnt*	brightly in the grate.
ten separate **fires**, which have	*burned*	through the floors,' he added.
at last was the **fire**	*burning*	hot, high and welcoming.
after midnight, leaving their **camp-fires**	*burning*	, made off to the south.
bothy home. A **turf fire**	*burned*	merrily in the end fireplace,
next door a **log fire**	*burned*	. My mother-in-law, remembering that
gas and coal fires to	*burn*	safely. Never block ventilation bricks.

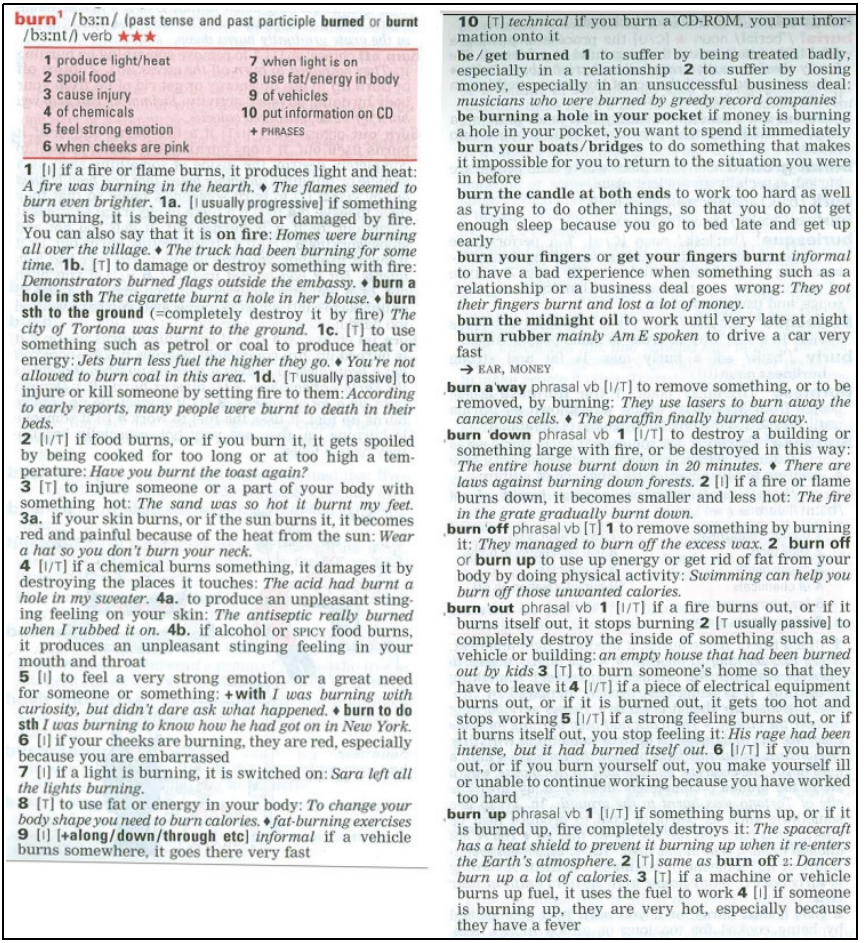
Figure 3: Concordances for fire burning

These occurrences of *fire burning* could be split into at least two sets: first, those which are controlled fires, as suggested by collocates such as *steadily, safely, grate,* or *welcoming*; second, uncontrolled fires, as in *burns its way through* or *burns through the floors*. The controlled/uncontrolled distinctions of *fire* are attested in the definitions provided by the *COLLINS COBUILD student's dictionary* (1993) and of the *MACMILLAN English Dictionary for Advanced Users* (2005), reproduced for convenience in Figures 4a and 4b:



burn /bɜːn/, **burns, burning, burned** or **burnt** /bɜːnt/. **1** VB If something **is burning,** it is on fire. *The stubble was burning in the fields.* ♦ **burning** UNCOUNT N *There was a smell of burning.* ♦ **burnt** ADJ *...a charred bit of burnt wood.* **2** VB WITH OBJ If you **burn** something, you destroy it with fire. *We couldn't burn the rubbish because it was raining.* **3** VB WITH OBJ If you **burn** yourself, you are injured by fire or by something very hot. *'What's the matter with your hand?'—'I burned it on my cigar.'* Also COUNT N *Many had serious burns over much of their bodies.* **4** VB WITH 'with' If you **are burning** with an emotion such as anger, you feel it very strongly. *...letters burning with indignation.* **5** ERG VB If the sun **burns** your skin, it makes it red or brown. **burn down.** PHR VB If a building **burns down** or **is burned down,** it is completely destroyed by fire.

Figure 4a: COBUILD dictionary entry for burn



burn[1] /bɜːn/ (past tense and past participle **burned** or **burnt** /bɜːnt/) verb ★★★

1 produce light/heat
2 spoil food
3 cause injury
4 of chemicals
5 feel strong emotion
6 when cheeks are pink
7 when light is on
8 use fat/energy in body
9 of vehicles
10 put information on CD
+ PHRASES

1 [I] if a fire or flame burns, it produces light and heat: *A fire was burning in the hearth.* ♦ *The flames seemed to burn even brighter.* **1a.** [I usually progressive] if something is burning, it is being destroyed or damaged by fire. You can also say that it is **on fire**: *Homes were burning all over the village.* ♦ *The truck had been burning for some time.* **1b.** [T] to damage or destroy something with fire: *Demonstrators burned flags outside the embassy.* ♦ **burn a hole in sth** *The cigarette burnt a hole in her blouse.* ♦ **burn sth to the ground** (=completely destroy it by fire) *The city of Tortona was burnt to the ground.* **1c.** [T] to use something such as petrol or coal to produce heat or energy: *Jets burn less fuel the higher they go.* ♦ *You're not allowed to burn coal in this area.* **1d.** [T usually passive] to injure or kill someone by setting fire to them: *According to early reports, many people were burnt to death in their beds.*
2 [I/T] if food burns, or if you burn it, it gets spoiled by being cooked for too long or at too high a temperature: *Have you burnt the toast again?*
3 [T] to injure someone or a part of your body with something hot: *The sand was so hot it burnt my feet.* **3a.** if your skin burns, or if the sun burns it, it becomes red and painful because of the heat from the sun: *Wear a hat so you don't burn your neck.*
4 [I/T] if a chemical burns something, it damages it by destroying the places it touches: *The acid had burnt a hole in my sweater.* **4a.** to produce an unpleasant stinging feeling on your skin: *The antiseptic really burned when I rubbed it on.* **4b.** if alcohol or SPICY food burns, it produces an unpleasant stinging feeling in your mouth and throat
5 [I] to feel a very strong emotion or a great need for someone or something: **+with** *I was burning with curiosity, but didn't dare ask what happened.* ♦ **burn to do sth** *I was burning to know how he had got on in New York.*
6 [I] if your cheeks are burning, they are red, especially because you are embarrassed
7 [I] if a light is burning, it is switched on: *Sara left all the lights burning.*
8 [T] to use fat or energy in your body: *To change your body shape you need to burn calories.* ♦ *fat-burning exercises*
9 [I] **[+along/down/through etc]** *informal* if a vehicle burns somewhere, it goes there very fast
10 [T] *technical* if you burn a CD-ROM, you put information onto it

be/get burned **1** to suffer by being treated badly, especially in a relationship **2** to suffer by losing money, especially in an unsuccessful business deal: *musicians who were burned by greedy record companies*
be burning a hole in your pocket if money is burning a hole in your pocket, you want to spend it immediately
burn your boats/bridges to do something that makes it impossible for you to return to the situation you were in before
burn the candle at both ends to work too hard as well as trying to do other things, so that you do not get enough sleep because you go to bed late and get up early
burn your fingers or **get your fingers burnt** *informal* to have a bad experience when something such as a relationship or a business deal goes wrong: *They got their fingers burnt and lost a lot of money.*
burn the midnight oil to work until very late at night
burn rubber *mainly Am E spoken* to drive a car very fast
→ EAR, MONEY

burn a'way phrasal vb [I/T] to remove something, or to be removed, by burning: *They use lasers to burn away the cancerous cells.* ♦ *The paraffin finally burned away.*
burn 'down phrasal vb **1** [I/T] to destroy a building or something large with fire, or be destroyed in this way: *The entire house burnt down in 20 minutes.* ♦ *There are laws against burning down forests.* **2** [I] if a fire or flame burns down, it becomes smaller and less hot: *The fire in the grate gradually burnt down.*
burn 'off phrasal vb [T] **1** to remove something by burning it: *They managed to burn off the excess wax.* **2** **burn off** or **burn up** to use up energy or get rid of fat from your body by doing physical activity: *Swimming can help you burn off those unwanted calories.*
burn 'out phrasal vb **1** [I/T] if a fire burns out, or if it burns itself out, it stops burning **2** [T usually passive] to completely destroy the inside of something such as a vehicle or building: *an empty house that had been burned out by kids* **3** [T] to burn someone's home so that they have to leave it **4** [I/T] if a piece of electrical equipment burns out, or if it is burned out, it gets too hot and stops working **5** [I/T] if a strong feeling burns out, or if it burns itself out, you stop feeling it: *His rage had been intense, but it had burned itself out.* **6** [I/T] if you burn out, or if you burn yourself out, you make yourself ill or unable to continue working because you have worked too hard
burn 'up phrasal vb **1** [I/T] if something burns up, or if it is burned up, fire completely destroys it: *The spacecraft has a heat shield to prevent it burning up when it re-enters the Earth's atmosphere.* **2** [T] *same as* **burn off** 2: *Dancers burn up a lot of calories.* **3** [T] if a machine or vehicle burns up fuel, it uses the fuel to work **4** [I] if someone is burning up, they are very hot, especially because they have a fever

Figure 4b: MEDAL dictionary entry for burn

As can be seen from the definitions, *fire* can also be used for electric or gas devices, in which case the literal meaning is almost absent. Interestingly, none of these definitions provide the information that *fires* are also used for light, and not only heat, though these two properties tend to go together; the sun, for example, provides light and heat at the same time. Going back to the evidence, comparatively few instances account for heating, but describe fire-burning as a general process.

At the time of designing the entry for *burn*, the semantic type [[Fire]] already existed. A semantic type is created when it has been repetitively observed in patterns and is considered relevant. [[Fire]] is considered as a process of combustion, which may be controlled or uncontrolled. This type may well be used for previously seen concordance lines. Two other instances share a similar use, illustrated in Figure 5:

tourists. Inefficient **stoves** are kept	*burning*	day and night to provide heat and food
rose again and the **brazier**	*burned*	hot. We warmed our hands as though

Figure 5: Concordances for Pattern 8

In these cases, *stoves* and *braziers* are not literally burning, but it is the fire on/in it which burns. There are two possible solutions: either considering these examples as instances of another category, or as instances of [[Fire]]. The closest existing ontological categories are [[Artefact]] and [[Furniture]], which include very different entities. *Stove* and *brazier* would share some features of [[Artefact]], [[Furniture]] and [[Fire]]. But new semantic types should not be created because of ontological incompatibilities. CPA favours a prototypical approach to meaning, where lexical units do not hold all the necessary conditions for membership. Given the low frequency, these two examples could therefore be considered as instances of the *fire-burn* pattern. As a consequence, the semantic type [[Fire]] will also contain *burning artefacts*, that is, controlled machines used to produce heat and light. Finally, other physical objects than stoves can be said to burn: *candle* and *light* (Figure 6):

light before the rood to	*burn*	from the second peal to matins,
film with only one **light**	*burning*	. The police were worried that squatters
by stagnant ponds, the **lights**	*burn*	into the night, and the driving engines
provide only a **candle** to	*burn*	during the daily mass at the church
for Amnesty. Keep the **candle**	*burning*	in the '90s Keep the 30th Anniversary
away. There were no **candles**	*burning*	in the windows. Today the new Little
10 circular tables. Tall **candles**	*burned*	in the blacked-out windows.
Martyr's Cross where their **candles**	*burn*	to a carpet of wax.
steps are lined with **candles**,	*burning*	all the time, so that now they rise

Figure 6: Additional concordances for Pattern 8

This is the same problem as before except that there is more evidence of a pattern this time, and that there exists a semantic type to capture a generalisation of these lexical items, namely [[Light Source]]. This semantic drift between heat and light is interestingly captured by Pattern 8 which covers all the examples seen in this subsection (see Table 2):

Table 2: Pattern 8 of the verb *burn*

8	5%	**[[Fire | Light Source]] burn [NO OBJ]**
		[[Fire | Light Source]] is in a state of combustion, producing intense heat or light

3.2 Generalisation with semantic types

The lexicographer refers to the ontology in order to choose an appropriate semantic type for a given pattern position. In this section, I will describe some difficulties which arise in choosing the appropriate level of generalisation. Example (5) illustrates Pattern 1 (Figure 1), where *castle* is a lexical instance of [[Building]] as *Duke of Argyll* is of [[Human]].

(5) In 1685 the **castle** was *burnt* by the **Duke of Argyll** and fell into ruin.

A semantic type will always be more general than the actual lexical set which fits in a pattern position (in a given corpus). For example, the lexical sets preferred as [[Building]] objects of the verb *burn* will never be exactly the same as those of the verb *leave*. But, since a corpus is a sample, the lexicographer can authorise more lexical items than only those observed, by means of the generalisation which enables the semantic type. A semantic type thus authorises more members than actually observed by the lexicographer. Hanks (Hanks and Jezek 2008) refers to this property of lexical sets as 'shimmering', meaning that some words are preferred to others in a specific pattern position though they may all belong to the same semantic type. "Lexical sets [in a given pattern position] are not stable paradigmatic structures" (Hanks and Jezek 2008: 399).

This discrepancy between lexical sets and semantic types renders the choice of the appropriate level of generalisation even more subtle. Indeed, a very abstract type like [[Anything]] or [[Animate]] does not take full benefit of the richness of the semantic ontology and does not capture a verb's collocational preferences. The lexicographer must try to find the most appropriate level of abstraction. It is therefore often preferable to have several alternative semantic types rather than one broad category which would encompass more than their addition. Examples (6) and (7) illustrate this point because their respective

objects belong to two different semantic types, namely [[Energy]] and [[Food]] of pattern 17 (see Table 3).

(6) Even people of the same age and sex can differ considerably in the **energy** they *burn* up to keep their bodies going.

(7) Too many diets also have the disadvantage that they are associated with a drop in metabolic rate; that is in effect, the rate at which you *burn* up **food**.

Table 3: Pattern 17 of the verb *burn*

17	2%	pv[a] **[[Animate]] burn [[Energy | Food]] ({up | off})**
		[[Animate]] 's [[Body]] makes use of [[Energy | Food]], typically by doing exercise

a. **pv** here stands for 'phrasal verb'.

[[Energy]] and [[Food]] are very different entities in the ontology (Figure 3) but both are attested as objects of *burn* in this specific meaning of bodily elimination. If the lexicographer were to combine both types under [[Entity]], he/she would authorise the inclusion of many more entities than only [[Food]] and [[Energy]]. For example, the semantic type [[Entity]] would authorise use of the words *table* and *idea,* neither of which is appropriate in this context (Figure 7):

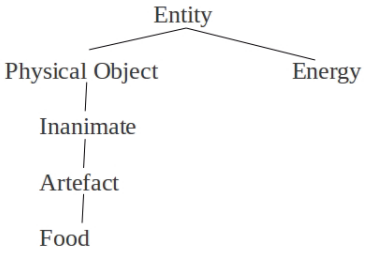


Figure 7: Ontological relatedness between [[Food]] and [[Energy]]

Both solutions are of course possible, but the more distant two semantic types are in the ontology, the less relevant it is to merge them in a more abstract type.

Another case worth mentioning is when both a semantic type and its hypernym are included in the same pattern position. This may seem to be an incongruence at first sight, but it is justified on the ground of collocational strength. In CPA, it is important to cover the main preferences of a verb. Thus, if a semantic type (corresponding to various lexical items) happens to be very frequently used in a given pattern position, it should be indicated in the pattern, even if its own hypernym has already been reported in the same configuration. For example, pattern 9 of the verb to burn combines with [[Building]] and [[Location]] and the latter is the hypernym of the former (see Table 4).

(8) The only barbecue we ever had was when the **house** next door *burnt* down.

(9) Although he captured some knights and *burned* down the **town**, he was unable to take the castle.

Table 4: Pattern 9 of the verb *burn*

9	8%	pv **[[Human]] burn [[Building | Location]] {{down} | to {the ground}}**
		[[Human]] sets fire to [[Building | Location]] and completely destroy it

In this case, it would not be wrong to leave [[Location]] as the only possible semantic type in Object position, and it would also be too restrictive to keep only [[Building]]. But the presence of both is justified by the salience of one subtype of [[Location]], which is [[Building]]. In other words, if a subtype's use is dominant, it should be recorded in the pattern.

Assigning semantic types to pattern elements is a subtle task. Experiments have shown that:

- one lexical item may well be a prototypical representative member of a semantic type such as [[Fire]],
- it is sometimes better to keep two alternative semantic types than a broader category (e.g. [[Fire]] and [[Light Source]] or [[Food]] and [[Energy]]),
- it is also conceivable to put both hypernym and hyponym in the same pattern position ([[Building]] and [[Location]]).

However, a change in the pattern which also entails a change in its meaning is considered as good evidence to create a new pattern. This is not always the case, especially for syntax.

4 Syntactic constructions

A CPA verb pattern is an abstraction of one or several syntactic constructions. Its representation is a fixed, ordered set of grammatical and semantic categories, and is based on the SPOCA model (Hanks 2013: 94):

- S – subject (almost invariably, a noun group)
- P – predicator (a verb group, including auxiliaries)
- O – object (a noun group; a clause may have 0, 1, or 2 objects)
- C – complement (an adjective or a noun group that is co-referential with either
- the subject or the object)
- A – adverbial (in systemic grammar called adjunct, a term that unfortunately
- has a different meaning in generative grammar)

Patterns capture the typical uses of verbs and are not driven by syntactic theory: syntactic features are not systematically recorded in patterns. This section offers an example of alternations which do not influence pattern creation (passive/active), and others which do (inchoative/agentive). It will also touch upon the issue of syntactic exploitations.

4.1 Passive voice

The most frequent English syntactic alternation is the active/passive voice. CPA, like 'Pattern Grammar' (Hunston and Francis 2000), does not generally encode the passive voice in the pattern. In other words, all patterns expressed in the active voice with a direct object could theoretically be passivised.

> If the subject of a passive verb is the same as the direct object of the same word as an active verb, it is generally not necessary to propose separate patterns for the active and passive uses of that verb. The semantic relationship is entirely regular and predictable, the same meaning of the verb being activated whenever the passive subject and the active direct object are members of the same lexical set or have the same semantic type.
>
> (Hanks 2013: 188)

Putting a verb in the passive voice is a change of focus rather than a change of pattern meaning. In some cases, however, it may be *relevant* to build a passive pattern. This is signalled by changing the pattern verb form to *be verb+ed*.

Two passive patterns have been identified for the verb *to burn*. Pattern 16 combines with *off* and mainly involves [[Stuff]] (solid or liquid unspecified material, etc.) being removed from a surface (see Table 5 and Figure 8). No active equivalent was found:

Table 5: Pattern 16 of the verb *burn*

16	1%	pv **[[Stuff | Physical Object]] be burned [NO OBJ] {off}**
		[[Stuff | Physical Object]] is removed from [[Surface]] by burning

The same **dirt** could be	*burnt*	off by the use of heat
rock from which the **peat cover** has been	*burnt*	off. Peat fires kill all vegetation
He cannot sleep. His **eyelids** have been	*burnt*	off. He puts a scarf around his head
the **grass** at the city's other cemeteries has been	*burned*	off. That's because the water table

Figure 8: Concordances for Pattern 16

The only other pattern using the particle *off* is Pattern 17 (*infra*): this pattern is in the active voice and involves [[Energy]] as subjects (eliminated by exercise), which is not really related to pattern 16, though both make use of the preposition *off*.

The second passive, Pattern 14, does not have an active counterpart, but a very close inchoative pattern (Pattern 13; see Subsection 4.2). Since passive/inchoative alternation is not predictable, these patterns were kept separate. It is interesting that both uses avoid the cause of *burning+out* (see Tables 6 and 7, and Figure 9):

Table 6: Pattern 13 of the verb *burn*

13	1%	pv **[[Human]] burn [NO OBJ] {out}**
		[[Human]] is exhausted, typically because of hard work over a long period of time

Table 7: Pattern 14 of the verb *burn*

14	1%	pv **[[Human]] {be | get} burned [NO OBJ] {out}**
		[[Human]] is or becomes exhausted, typically because of hard work over a long period of time

Peter Stanford sounds even further	*burnt*	out than Graham Greene who
between shows, the **Girls** were	*burnt*	out coping with double and tripling.
We were really young, and were almost	*burnt*	out at 21 years old,' explained Setzer
stories of the **youngster** being `	*burnt*	out' by the gipsy life of constant
We were all	*burned*	out, exhausted by the battles
the sort of situation where **bands** get	*burned*	out.'
and **they** can get quickly	*burnt*	out; the Italian football landscape is
Wild Thing on his guitar -- **I** felt	*burnt*	out the next morning, but everyone
Many of them got	*burnt*	out. The loneliness, the below-the-poverty-line

Figure 9: Concordances for Pattern 14

4.2 Causative/inchoative alternations

A frequent English alternation found in *burn* and many other verbs is the so-called causative/inchoative alternation, or in COBUILD terminology, 'ergative'. Examples (10) and (11) provide examples for the verb *to burn*.

(10) it was rumoured that the **Louvre** had been *burned* down as well.

(11) the old **Palace of Westminster** which had *burned* down in 1834

The first example is causative (though the agent is not present) because the verb is passive and the action was started rather than starting by itself (inchoative; example (11)). The lexicographer has the possibility of splitting apart concordances of *burn+down* according to this criterion. This would result in pattern (a) and (b) in example (12).

(12) Pattern a: [[Anything]] be burned down
Pattern b: [[Anything]] burn down [[NO OBJ]]

Pattern (a) can be subsumed under Pattern (c), which encompasses instances in both active and passive voices.

(13) Pattern c: [[Human]] burn down [[Anything]]

This operation could however be considered suspicious if done automatically: the lexicographer must seek supporting evidence that the active/passive alternation does not make any difference: here the most abstract type [[Anything]] was put in subject position for the sake of presentation, but this position needs to be probed and specified (cf. *infra* Table 4).

Now, the same question has to be asked in the case of the causative / inchoative alternation: should Pattern (b) also be considered as a minor alternation of pattern (c)? This question points to the relevant level of grammatical generalisa-

tion. In fact this level may vary according to the kind of evidence offered in the corpus, such as:

- Preferred grammatical structures: frequent patterns should be recorded. If both agentive and inchoative patterns are frequently used, then there is reason to keep them separated. Otherwise, if one is only used in a few occurrences, it can be considered as a syntactic alternation of the other.
- If the alternation entails a change in verb meaning or if it is correlated with a change in pattern element, such as semantic types, they should probably be kept separated.

For example, the inchoative uses of the phrasal verb *burn+down* combine with [[Fire]] in subject position. In this case, there is a clear meaning shift: a fire which burns down means it is becoming less intense, while someone burning down a building means that he/she sets fire in order to destroy it. This correlation between meaning, semantic type and syntactic frame constitute a sound clues for the creation of a new pattern (Table 8), even if the pattern is not frequently observed: four occurrences (see Figure 10):

Table 8: Pattern 10 of the verb *burn*

10	<1%	pv **[[Fire]] burn [NO OBJ] {down}**
10		[[Fire]] becomes less intense

like lighting two candles at the same time; **they**'re going to	*burn*	down	differently').
by a duergar to sit by his fire and warm himself. As the **fire**	*burnt*	down	, the duergar
right of the fire and placed it amongst the embers. As **this**	*burnt*	down	, the DWARF
for saying less.' Then he spoke softly with him until the **fire**	*burned*	down	very low.

Figure 10: Concordances for Pattern 10

Frequency should not rule out patterns relevant to the lexicographer. One reason is that in the process of annotation, he/she may miss some instances. Another reason is that by modifying the level of generalisation, like changing a semantic type, he/she may progressively populate this pattern, while keeping its global coherence intact.

It is important to note that patterns record the most frequent syntactic use (active, passive, inchoative) but that this use may not be systematically instanti-

ated in the corresponding concordances: the syntactic structure in the pattern does not necessarily match the syntactic structure in the instance. One should keep this issue in mind in computational linguistics experiments.

4.3 Optionality and syntactic exploitations

Patterns are typical semantic structures abstracted from texts. Texts are language units in use and obey various discourse constraints (see for example Halliday and Hasan 1976). In consequence, pattern instances may be interrupted, scattered or appear incomplete. This fact should not prevent the lexicographer from tagging these instances, provided some principles are given to him.

The first issue is optionality, which Hanks (2013: 200) sums up, in the case of adverbials, as follows:

> Some adverbials are obligatory; others are optional; and to make matters worse, some obligatory adverbials can be elided!

An example of an obligatory adverbial for the verb *to burn* is the one used in pattern 19 (see Table 9):

Table 9: Pattern 19 of the verb *burn*

19	1%	**[[Human]] burn [[Information | Image]] {into | onto | on [[Physical Object]]}**
		[[Human]] prints [[Information | Image]] on the surface of a [[Physical Object]] by application of heat

(14) The pattern of 0s and 1s is `*burned*' into it once and for all on manufacture.

The absence of the prepositional phrase would make the clause ambiguous; in fact, all instances with the same meaning were found to include this prepositional phrase. This was not the case for Pattern 21 (see Table 10):

Table 10: Pattern 21 of the verb burn

21	2%	**[[Human]] burn [NO OBJ] with [[Emotion]]**
		[[Human]] feels [[Emotion]] very intensely

(15) best preached by someone who ***burns*** *with zeal* rather than sexual desire

(16) Bono was a different kind of frontman ***burning*** *with conviction*,

(17) They record him ***burning*** *with anger* against those who practised their faith

(18) He still ***burned*** to preach the Gospel to the poor and to help them.

The absence of an obligatory prepositional phrase (example 18) does not prevent us from understanding the meaning of the verb, though the nature of the emotion is fuzzy. Example (18) could be paraphrased as 'be excited or stimulated'.

Passive voice very often entails an elision of the agent, either because it is obvious in the context ('text-transitive'), or because it is part of a strategy from the writer. Such elisions are not considered as syntactically anomalous, though it would be interesting to draw a gradation of syntactic anomaly and multiply the number of possible subcategories to characterise such uses.

Finally, the lexicographer sometimes has to analyse the wider context (before or after the main sentence) to decide on a pattern. Using the wider context is a sign that a use may be anomalous, since all normal uses should contain sufficient clues. However, the lexicographer may use it to confirm a semantic type, especially in cases of pronominalisation. Syntactic anomalies are therefore quite rare, because they are often intended to provoke an effect. They are sometimes found in poetry, as illustrated by the following example:

(19) The more the kindled combat rises higher, The more with fury *burns* the blazing fire.

Syntactic anomalies also cover unusual idiom constructions, such as the following (related to Pattern 24; cf. *supra*):

(20) The brothers had money to *burn* and they were often in disguise

(21) unless you've got money to *burn* these expensive guitars are probably not the instruments to get you started.

5 Figurative language

CPA relies on the Theory of Norms and Exploitations (TNE), the main claim of which is that "a language consists of a constantly moving and developing double helix of rules governing linguistic behavior: normal uses and exploitations of normal use." (Hanks 2013: 215). This binary tension between norms and exploitation may remind us of Sinclair's distinction between the 'idiom principle' and the 'open choice principle', but these sets of terms are not interchangeable. What is more, Hanks is, to the best of my knowledge, the first to propose an

explanation for creative use of language as a deviation from a given norm on a given dimension. His theory provides principles to describe and identify these exploitations. And most importantly, this theory is applied and tested on large corpora.

5.1 Idiomatic expressions and fixity

One of the first elements which spring to mind concerning figurative language, is the status of idiomatic expressions. Research in corpus linguistics, through lexical analysis, has unveiled the pervasiveness of idiomatic constructs to the point of proposing that language comes as "a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair 1991: 110), also known as Sinclair's 'idiom principle'. Co-selection of lexical items is unveiled through the mechanism of collocation, at one end of the spectrum, and through 'idiomatic expressions', at the other end.

Idiomatic expressions have received several definitions, mainly based on the issue of 'compositionality', where the meaning of a sequence can fully be derived from the meaning of its parts. Idioms have been defined as non-compositional and are considered as fixed expressions (Villada Moirón 2005). They may be considered as anomalous, because they are fixed or "fossilized units, restricted collocations with specialized and idiosyncratic meanings, which lie *outside the general grammar* of the language" (Moon 1996: 245; my emphasis). CPA proposes to identify them and to list them along with the patterns by representing their structures with the same means as the ones used for norms, and by providing a corresponding implicature. Fixity is not a representational problem since patterns regularly involve fixed lexical items at specific positions. At the same time, the representational apparatus of CPA seems sufficient to represent possible variation in idiom instances. Idioms are therefore identified on the basis of their non-compositionality.

Four idioms have been identified for the verb *to burn* (see Table 11). They are generally listed at the end of the dictionary entry, which may be a sign of their specificity:

Table 11: Idiomatic patterns of the verb *burn*

22	1%	**idiom [[Human]] burn {REFLDET finger}**
		[[Human]] has failed in an attempt to do something and has suffered problems as a result

23	<1%	**idiom [[Human]] burn {the candle} {at {both ends}}**
		[[Human]] is doing too much and so may not have time to sleep at night
24	1%	**idiom [[Money]] burn {hole} {in {pocket}}**
		[[Human]] has [[Money]] and has a reckless desire to spend it
25	1%	**idiom [[Human]] keep the home fire burning**
		[[Human]] keeps the {home} in good order while other members of the family are away, especially at war

Table 11 shows that semantic types may well be used for representing idioms. Pattern 22, for example, generalises subjects to [[Human]] and sets the specifier of its lexicalised object (*finger*) as a reflexive determiner. Most interestingly, idioms are prone to the same syntactic alternations as regular patterns. As can be seen from Figure 11, all the forms of the verb *to burn* (*burning*, *burn*, *burnt*, *burned*) can be found, and the idiom can be passivised with *get*, which preserves the same subject ([[Human]]):

Opposition	*burn*	their fingers on microwave Sketch.
Since	*burning*	their fingers on 100% lending that flared into bad debt, lenders
Having	*burned*	their fingers two years ago, senior Ford executives
`We got our fingers	*burnt*	there,' he admits.

Figure 11: Concordances for Pattern 22

The meaning stays the same, despite these variations, and this meaning is clearly not literal: human beings only symbolically burn their fingers, that is, they have failed in an attempt to do something and have suffered problems as a result (Pattern 22, Table 11). Pattern 25 is probably the most fixed of all four idioms, as illustrated in Figure 12:

Hampstead or with one of her many friends, kept the home fire	*burning*
Women keep the home fires	*burning*
here's how to keep the home fires	*burning*

Figure 12: Concordances for Pattern 25

The way the pattern is built also allows elements to be introduced in between curly bracket groups. The introduction of the adverb *furiously* after *candle* (Figure 13) is a clue for idiom chunking:

My candle	*burn*	at both ends. It will not last the night.
work overtime,	*burn*	the candle at both ends, slog on when I'm asleep
Despite his injuries, he	*burned*	the candle furiously at both ends

Figure 13: Concordances for Pattern 23

Idioms are part of Figurative Language. In the case of *burn*, they symbolically evoke meaning through visual scenes. Figurative use has several places in TNE, because it is not led only by semantic considerations.

5.2 The tension between usage and figurative language

As hinted in Sub-section 5.1, corpus linguistics' and CPA's main claims is that language is highly patterned, to the point that patterns are more relevant than words, as meaning units. Idioms are highly lexicalised sorts of patterns but are mainly identified on the basis of their figurative uses. However, figurative uses are not only restricted to idioms, and idiomaticity is not constrained to idioms (cf. 'idiom principle'; Sinclair 1991). In fact, Hanks describes idioms as "frozen conventional metaphors" (Hanks 2013: 344). His main point concerning figurative language, or metaphor, is that most of its use is normal, or conventional.

> Anyone who undertakes corpus analysis cannot but be struck by the large number of metaphorical uses of everyday words. However, most of these uses are conventional—that is, they are secondary norms, which were once, no doubt, creative exploitations of a norm, but have now become established as secondary norms in their own right. Novel, creative metaphors are much rarer.
>
> (Hanks 2013: 221)

> [Metaphorical] Developments such as these are of great interest to historians of meaning change, but completely irrelevant to the meaning of these words in modern English. Meaning change is a slow- moving conveyor belt that, every now and again, assimilates a novel metaphor or some other innovative usage, conventionalizes it, and eventually (perhaps) changes it further or discards it in favor of some other, competing convention that has arisen.
>
> (Hanks 2013: 302)

This point also applies to the analysis drawn so far, since most conventional metaphors of the verb *to burn*, have already been discussed as patterns. For

example, Patterns 13 and 14, repeated here for convenience, have been discussed with regards to syntactic alternations. They belong to the same phrasal verb *burn+out*, which gave rise to four patterns in total (see Table 12):

Table 12: Patterns of the phrasal verb *burn out*

11	1%	**pv [[Fire | Light Source]] burn ([[Self]]) {out}**
		[[Fire | Light Source]] finishes burning or shining, having consumed all the fuel or other material available
12	1%	**pv [[Human]] burn [[Artefact | Vehicle]] {out}**
		[[Artefact | Vehicle]] stops working properly because [[Human]] set fire to it
13	3%	**pv [[Human]] burn [NO OBJ] {out}**
		[[Human]] is exhausted, typically because of hard work over a long period of time
14	1%	**pv [[Human]] {be | get} burned [NO OBJ] {out}**
		[[Human]] is or becomes exhausted, typically because of hard work over a long period of time

The first two patterns can be described as literal, while the last two are examples of physical/abstract alternation. This alternation, where *fire-burning* disappears, is correlated by a change in the semantic type of the *burned-out* entity. The metaphorical mapping could be the following: *burning* is the main activity of [[Fire]] and *burning out* means that this activity does not exist anymore. Similarly an [[Artefact]] or [[Vehicle]] would stop its main activity (glossed as 'working properly'; Pattern 12) and [[Human]] would stop its main activity (Pattern 13 and 14).

Two other verb patterns are conventional metaphors: Pattern 17, which has already been discussed, and which can be described in the same way as the previous patterns (*burning+off*; *eliminating [[Energy]]*), and Pattern 21. Pattern 21 is an interesting case of conventional metaphor, because it illustrates a regular type of metaphorical mapping (see Table 13):

Table 13: Pattern 21 of the verb *burn*

21	2%	**[[Human]] burn [NO OBJ] with [[Emotion]]**
		[[Human]] feels [[Emotion]] very intensely

The metaphor involves the comparison of [[Human]] emotions with [[Fire]]. The main effect or meaning of this pattern is to underline the intensity of emotions. This also happens with verbs like *boil*, *explode* or *fume*.

5.3 Figurative instances

Figurative uses may be spotted very early in the process of pattern creation. In case no relevant pattern yet appears to the lexicographer, they are first considered as figurative exploitations. Later in the process, the lexicographer goes through the figurative exploitations again, in case a use is so frequent as to count as a pattern. Therefore, some lines which at first appeared to be exploitations turn out to be regular patterns of use in the end.

Three types of exploitations can actually be identified through the CPA interface:

- s = where a syntactically anomalous construction is identified (cf. 3.3)
- a = where an anomalous argument is identified.
- f = where the use is figurative.

The interface does not enable the lexicographer to tag more than one exploitation for one instance: such categories are exclusive. However, it would sometimes be desirable to do so. As briefly noted earlier, it would also be desirable to specify which element in the pattern is responsible for the anomaly.

The boundary between figurative use and anomalous argument categories is fuzzy. This is so because, figurative uses, such as metaphors, often involve abnormal arguments. In the case where both can be tagged, figurative use is preferred.

Given two patterns (or more) sharing a significant amount of data, it is sometimes difficult to decide, when a figurative use is identified, of which pattern it is an exploitation. This is all the more difficult when patterns are syntactic alternations of one another.

(22) I had no idea if we had sustained damage to the undercarriage, although we had **three greens** *burning* bright

(23) Between these two ridges the fire of the sunset falls along the trough of the sea, dyeing it with an awful but glorious light, the intense and lurid **splendour** which *burns* like gold, and bathes like blood.

Example (22) is a specific use from the aeronautic domain: three greens are lights in the cockpit. This is uncommon to the regular user and should thus be considered as an abnormal use. Example (23) sounds more poetic (‘hypotypo-

sis') but the use is not really figurative. It is not the light which is the actual syntactic argument but its *splendour which burns like gold*. *Splendour* is however not a sort of [[Light Source]], nor of [[Fire]], and should therefore be considered as anomalous.

Figurative uses involve a change of meaning, like physical / abstract (see Lakoff and Johnson 1987 or Fauconnier 1997 for examples of metaphorical mappings), and most importantly, they do not require an anomalous argument.

(24) The earth he painted was impregnated earth. His skies boiled and *burned*,

(25) the evening sky begins to *burn* with a transparent intensity

(26) Love *burns* hot even on cold Sabbaths

(27) a devotion to art that `*burns* with a pure flame'

(28) the flame of hope *burns* brightly here,' he said

(29) First the fire of God's anger *burns* so fiercely among them

Examples (24) to (27) are instances of both anomalous arguments and figurative uses. In fact it is not even certain how *sky-burning* should be interpreted: is it intensely blue or is it filled with colours that makes it look like a fire? Example (26) resonates with the notion of heat from *burning*: *love* is compared to a home fire, i.e. a safe place. Example (27) may be an idiomatic use or a specific domain pattern (religion); there are too few instances to make a pattern. It is a metaphor which relies on the belief that fire has the power to purify the soul (hence *devotion*), as can be found in some instances of Pattern 6.

Examples (28) and (29) both contain a [[Fire]] collocate in subject position. However, these subjects are modified by a prepositional phrase which contains the semantic head: *hope* in (28) and *anger* in (29). Therefore, they are not literal uses, but exploitations. The meaning of the patterns is similar to *spreading* (28), or *abating* (29).

The last example illustrates a figurative use of a conventional metaphor. The only case found for the verb *to burn* is an exploitation of Pattern 17. Most of the time the lexicographer spots an anomaly, he/she needs to analyse the wider context in order to annotate an instance as a figurative use.

(30) **Anderson**, often eager to jump around and *burn* off some **frustration**, cut the little bald patch on the top of his head, and cursed himself.

In this case, *frustration* is an anomalous lexical item: its use as subject of the verb *burn* is not conventional. *Frustration* is metaphorically analysable as a bad kind of fat that can be burnt off by engaging into some activity. The meaning of the verb here is also more abstract, which is why the lexicographer did not annotate it as an instance of anomalous argument.

6 Conclusion and perspectives

During the process of pattern creation, the CPA lexicographer is constantly testing the elasticity of his or her pattern categories. Every line may contain clues to start a new collocational set, category, or pattern. This article has hardly touched upon real problems which spring everywhere when using corpora. However it will serve as useful premises for lexicographers interested in CPA.

CPA proposes a general methodology based on principles set out in TNE (Hanks 2013). Putting this theory into practice is a major challenge for a lexicographer. This is why all must be done to formalise as much as possible the procedure and identify difficult areas or fuzzy categories. This article has gone through three important components in which subtle decisions are made by the lexicographer in pattern construction and annotation: Semantic Types, Syntactic Constructions, and Figurative Language.

I hope to have shown, throughout the article, that the task posed to the CPA lexicographer is huge: he/she must be aware of multiple possible orientations in his/her entry and could theoretically craft it in very different ways. This is where computational linguistics could bring some support. Various tools can be imagined: tools for consistency checking, inside the pattern, or between the corpus and the pattern. Preprocessing tools are also much awaited by lexicographers. In fact the actual editing environment does not rely on state-of-the-art Natural Language Processing tools such as syntactic or semantic parsers. Statistics could also help in clustering or cluster analysis. Finally, visualisation is an important component of the analysis and comfort of the lexicographer and should not be neglected. Such are the possible perspectives to be explored in the near future, thanks to the DVC project.

References

Dictionaries

COLLINS COBUILD student's dictionary (1993)
MACMILLAN English dictionary for advanced users (2005)

Other

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4): 243–257.

Cinková, Silvie, Martin Holub and Vincent Kríž. 2012. Managing uncertainty in semantic tagging. In *Proceedings of EACL*, 840–850.

Fauconnier, Gilles. 1997. *Mappings in thought and language*. Cambridge (UK): Cambridge University Press.

Halliday, M.A.K. and Christian Matthiessen. 2004. *An introduction to functional grammar.* London: Arnold.

Halliday, M. A. K. and Ruqaiya Hasan. 1976. *Cohesion in English.* London: Longman.

Hanks, Patrick. 2004. Corpus pattern analysis. In *Euralex Proceedings*, 87–97.

Hanks, Patrick. 2013. *Lexical analysis*. Cambridge (Ma): MIT Press.

Hanks, Patrick and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée* 10 (2): 63–82.

Hanks, Patrick and Elisabetta Ježek. 2008. Shimmering lexical sets. In *Euralex Proceedings*, 391–402.

Hunston, Susan and Gill Francis. 2000. *Pattern grammar.* Amsterdam: John Benjamins.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrž and David Tugwell. 2004. The Sketch Engine. In *Euralex Proceedings*, 105–116.

Lakoff, George and Mark Johnson. 1980. *Metaphors we live by.* Chicago: University of Chicago Press.

Moon, Rosamund. 1996. Data, description, and idioms in corpus lexicography. In *Euralex Proceedings*, 245–256.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, John. 1998. The lexical item. In E. Weigand (ed.). C*ontrastive lexical semantics*, 1–24. Amsterdam: John Benjamins.

Sinclair, John. 2004. *Trust the text: Language, corpus and discourse*. London: Routledge.

Villada Moirón, Begoña. 2005. Data-driven identification of fixed expressions and their modifiability. University of Groningen (PhD Thesis).

Appendix

CPA entry for the verb *to burn*.

No.	%	Pattern / Implicature
1	19%	**[[Human]] burn [[Physical Object | Building]]**
		[[Human]] sets fire to [[Physical Object | Building]] in order to destroy it
2	3%	**[[Physical Object | Location]] burn [NO OBJ]**
		[[Physical Object | Location]] is in flames or is severely damaged because of fire
3	10%	**[[Human | Institution | Machine]] burn [[Stuff = Fuel]]**
		[[Human | Institution | Machine]] causes [[Stuff= Fuel]] to be consumed by fire or heat, typically in order to produce energy
4	3%	**[[Stuff]] burn [NO OBJ]**
		[[Stuff]] is consumed little by little by fire and so produces energy
5	<1%	**{incense} burn [NO OBJ]**
		{incense} is set on fire in order to produce a pleasant smell, especially in religious ceremonies
6	13%	**[[Human 1]] burn [[Human 2 | Flag | Picture]]**
		[[Human 1]] deliberately and publicly sets fire to [[Human 2 | Flag | Picture]] as a punishment, retaliation, protest, or religious excess
7	4%	**[[Physical Object]] burn ([[Human | Body Part]])**
		[[Physical Object]] causes damage, pain to [[Human | Body Part]]
8	5%	**[[Fire | Light Source]] burn [NO OBJ]**
		[[Fire | Light Source]] is in a state of combustion, producing intense heat or light
9	8%	**pv [[Human]] burn [[Building | Location]] {{down} | to {the ground}}**
		[[Human]] sets fire to [[Building | Location]] and completely destroy it
10	<1%	**pv [[Fire]] burn [NO OBJ] {down}**
		[[Fire]] becomes less intense
11	1%	**pv [[Fire | Light Source]] burn ([[Self]]) {out}**
		[[Fire | Light Source]] finishes burning or shining, having consumed all the fuel or other material available
12	3%	**pv [[Human]] burn [[Artefact | Vehicle]] {out}**
		[[Artefact | Vehicle]] stops working properly because it was damaged
13	1%	**pv [[Human]] burn [NO OBJ] {out}**
		[[Human]] is exhausted, typically because of hard work over a long period of time
14	1%	**pv [[Human]] {be | get} burned [NO OBJ] {out}**
		[[Human]] is or becomes exhausted, typically because of hard work over a long period of time
15	1%	**pv [[Fire]] burn [[Stuff | Physical Object]] {away}**
		[[Fire]] removes and destroys [[Stuff | Physical Object]] from [[Surface]]
16	1%	**pv [[Stuff | Physical Object]] be burned [NO OBJ] {off}**
		[[Stuff | Physical Object]] is removed from [[Surface]] by burning
17	2%	**pv [[Animate]] burn [[Energy | Food]] ({up | off})**
		[[Animate]] 's [[Body]] makes use of [[Energy | Food]], typically by doing exercise
18	<1%	**[[Human]] burn [[Food]]**
		[[Human]] spoils [[Food]] by overheating it or by cooking it for too long
19	1%	**[[Human]] burn [[Information | Image]] {into | onto | on [[Physical Object]]}**
		[[Human]] prints [[Information | Image]] on the surface of a [[Physical Object]] by application of heat
20	1%	**[[Air | Fire | Light Source]] burn {hole} {in [[Physical Object]]}**
		[[Air | Fire | Light Source]] destroys matter and causes a {hole} in [[Physical Object]]
21	2%	**[[Human]] burn [NO OBJ] with [[Emotion]]**
		[[Human]] feels [[Emotion]] very intensely
22	1%	**idiom [[Human]] burn {REFLDET finger}**
		[[Human]] has failed in an attempt to do something and has suffered problems as a result
23	<1%	**idiom [[Human]] burn {the candle} {at {both ends}}**
		[[Human]] is doing too much and so may not have time to sleep at night
24	1%	**idiom [[Money]] burn {hole} {in {pocket}}**
		[[Human]] has [[Money]] and has a reckless desire to spend it
25	1%	**idiom [[Human]] keep the home fire burning**
		[[Human]] keeps the {home} in good order while other members of the family are away, especially at war