Half an Article

Corpus patterns and lexicalist syntax

Maarten Janssen

1. Introduction

Word classes play an important role in many aspects of the formal treatment of language. To name just a few: in classic Chomskian grammars, rewrite rules combine word classes into sentences. In modern day computational linguistics, almost all forms of treatment start out with Part-of-Speech tagging. In inflectional morphology, the type of inflection a word undergoes depends directly on the word class. And most non-specialized language resources, such as dictionaries, state for each word what kind of class it belongs to. Given the importance of word classes, it is crucial to be able to assign a word class to each word in the lexicon.

Word classes give an indication of the position a word can take in a sentence: the word *abbreviate* cannot appear in direct object position, because it is a verb, and not a noun. In this paper, the notion of a word class will be used more widely to also include sub-class indications, which further indicate the combinatorial properties of the word: we say there are classes of transitive and intransitive verbs because verbs of the first type take an internal argument and words of the second type do not. And by making a distinction between absolute adjectives and comparative adjectives, we can explain why *my car is *green than yours* is ungrammatical, whereas *my car is greener than yours* is fine. Or in the other direction: why the word *better* in the sentence *that is better* has to be a comparative adjective, whereas in *he is a better* is has to be a noun (someone who bets).

In its strictest interpretation, word classes are the driving force behind syntax: syntactic rules describe how word classes combine into sentences, without caring about the exact lexical items that are being used. Grammatical sentences are those sentences that consist of sequences of word classes that can be correctly generated by the grammar. When the wrong word class is used we consider a sentence ungrammatical, but when the wrong lexical items within a word class are being combined, we say the sentence is grammatical, yet semantically unwellformed. Ideally, word classes should be specific enough to allow the syntax to distinguish between ungrammatical sentences and semantically unwellformed sentences.

There are two large areas where the assignment of word-classes becomes problematic. On the one side, it can be difficult to assign a part-of-speech tag to a particular use of word: should nominalised adjectives as in *The small and the great are there* be seen as a noun, or as an adjective with a suppressed noun? Another well-described problem is the distinction between the past participle and the adjective: should an adjectival past particle as in *the nudged ball* be seen as an adjective, or (still) as a past participle? In both cases, there are two part-of-speech tags that would correctly describe the grammatical behaviour of the word, and the issue is how many distinct classes to assign to a single word.

On the other side, problems arise when it is not clear which part-of-speech tag to assign to a word in the first place, and it is this type of problem that this article focuses on. This article will describe two cases where the assignment of a (sub) class is problematic. The first is the example of a rather well described phenomenon: if we attempt to assign subclasses to verbs that fully capture their behaviour, we need increasingly fine-grained verb classes, up to the point where some classes might end up with only one member (Fillmore 1968, Pollard & Sag, 1994, Levin 1993, Koenig *et al.* 2001, etc.). For instance, the behaviour of the verb *fire* is so specific that there is at best a small set of verbs that behaves the same, meaning that if we want to assign *fire* a verb class that predicts or describes its use, it has to be a very specific class, assigned at best to a handful of verbs.

The second example is somewhat more exotic one. We will provide an analysis of the English word *half*, showing that there is no good way to properly assign it a discriminating part-of-speech tag. It is in a sense a determiner, a noun, an adjective, a quantifier and an adverb at the same time. Yet it behaves like no other of any of those classes.

This paper argues that for both *fire* and *half*, it is misleading or wrong to assign it a morphosyntactic (sub) class: that the assignment of a class suggests a similarity between the behaviour of those two words that does not exist in practice. The special behaviour of *fire* and *half* is nothing special: in fact, a large amount of words in the English language behaves exactly like itself and nothing else. Therefore, it is better to model the behaviour of words directly, without the intermediary use of word classes.

One of the theories to model the distributional behaviour of words without the use of word classes is Corpus Pattern Analysis (CPA). Chapter 3 describes how CPA can be used to describe the word *fire* in a way that captures its unique behaviour. Like comparable frameworks, CPA is only used to describe the behaviour of verbs, and therefore, it is not directly usable for the description of the behaviour of *half*. However, we will sketch how an extended version of CPA could be used to model the peculiar behaviour of *half* as well.

2. Unique words

In this section we will look into the word class of two specific words: the verb *fire*, and the word *half*. Since we are looking at the verb *fire* (as opposed to the noun *fire*), the assignment of a class is easy: it is a verb. However, we will describe in some more detail what class of verb it is. The reason for looking at verb classes is that the tentative solution suggested in the remainder of this paper (Corpus Pattern Analysis) is currently only applied to verbs. By looking at the word *half*, we will see how this framework could be extended to other classes of words as well.

2.1. Firing verb classes

Verbs do not form a homogeneous class: not all verbs are interchangeable in grammatical sentences. Transitive verbs require a direct object, whereas intransitive verbs cannot have one. Furthermore, there are verbs that require a PP argument, verbs that need a reflexive clitic, etc. There is a large amount of literature on the description of different classes of verbs, many of them proposing very fine-grained classifications of verbs.

One of the most extensive works on the classification of is the study by Levin (1993), who gives a detailed classification of over 3000 English verbs, and which distinguishes over 200 distinct verb classes. She lists *fire* as a verb of class 17.1, which is the class of *throw* verbs. This because like *throw*, *toss,* and *fling*, the verb *fire* is a verb that involves motion of an object from a source to a target, both of which can be optionally realized, by an agent by means of expulsion.

However, the verb *fire* does behave differently from *toss* and *throw* in number of ways. Firstly, you toss a ball *to* somebody, but you fire a bullet *at* somebody. And secondly, *throw* can be used ditransitively, but *fire* cannot:

(1) John tossed Mary the ball
(2) *The police fired the thieves the bullet

If *fire* is a verb of the same general class as *throw*, it is a specific type of member, probably due to the fact that you expect Mary to have the ball if you toss it to her, but you do not expect a thief to hold/own the bullet when you fired it at him. The suggestion that the "projected possession" plays a role is supported by the fact that you tend to throws balls *to* people, but when you throw rocks *at* people, you typically intend to hurt them (rather than provide them with rocks). And you do throw someone a ball, but you do not throw someone a rock (as a weapon).

Because of the lack of "projected possession", *fire* behaves more like those *throw* verbs (in the Levin classification) where the direct object is a "target", and not a "goal", such as for instance *shoot* or *catapult*: you shoot arrows at people, and catapult stones at buildings. However, the verb *fire* does not behave exactly like *catapult* or *shoot* either. Let us just consider two differences.

One of the characteristics of *fire* is that the instrument can be realized as a direct object: you cannot only fire a bullet (from a gun), but you can also fire a gun. Since *catapult* is a verb that has the instrument built into the verb, you cannot really catapult a catapult; you cannot even catapult stones from a catapult, you can only fire stones from a catapult. The verb *shoot*, however, does allow the realization of the instrument as the direct object: you can shoot an arrow from a bow, or you can shoot a bow. However, it is common to talk about shooting bullets from guns, but shooting guns seems to be more marginal. And similarly, you can fire arrows from bows, but firing bows is less common.

Also in other ways, *shoot* and *fire* do not behave the same: you can shoot a person, but you cannot really fire a person (at least not in the sense of firing a bullet at him). And if you fire a warning shot, you probably fire a projectile from a weapon, but the direct object is neither the projectile nor the weapon. With shooting, such a construction does not seem possible.

Summing up, the verb *fire* behaves different even from the verbs that are supposed to be its closest relatives. Therefore, if we want to maintain that word classes can distinguish ungrammatical from grammatical sentences, we have two choices: either we have to assign the verb *fire* a verb class that is so restrictive that it only contains the verb *fire*, or we have to maintain that sentences like (2) are not ungrammatical, but just semantically off.

However, in terms of meaning, there does not seem to a good reason why you cannot use *fire* ditransitively, or why you cannot fire people (in the sense of shooting them). All the necessary semantic elements are in place, and these sentences in principle convey perfectly fine ideas, it is just that firing people or firing people bullet are not the correct way to express those ideas. Therefore, the tentative conclusion is that *fire* does not strongly belong to any verb class: assigning it a verb class can highlight some similarities with other verbs, but does not explain/model the full behaviour of the verb. Furthermore, although the verb *fire* is hence a unique verb, it is not unique in being unique: on close inspection, many if not most verbs have their own peculiar behaviour.

2.2. Half a word class

Let us now look at the distributional behaviour of the word *half*: what kind of word is it to start with? It can clearly be used as a noun in sentences like *the first half was rather boring,* but that is not the use of interest here. The word *half* is more frequently, and more interestingly, used in construction like (3).

(3) I ate half of the apple

The majority of dictionaries actually list the use of *half* in (3) as a noun, as is stated explicitly in the definition of *half* in the online version of the Merriam Webster:

> [1]**half** *n* **1 a** : either of two equal parts that compose something; *also* : a part approximately equal to one of these <half the distance> <the larger half of the fortune> **b** : half an hour —used in designation of time

However, it is clear the *half* is at least not a typical noun: there are very few nouns that can be used instead of *half* in (3). In that sentence, half occurs as a quantifier, and goes together with words like *some, most,* and *none*, but also with longer phrasal elements like *too much, a little, a really tiny portion*, and *some certain amount*.

Since *half* is a lexical element rather than a phraseme, you would expect it to be of the same class as *some* and *most* (both of which are typed as adjectives in the Merriam Webster), forming a class of (say) quantifiers. But although all these three items are usable in the construction in (3), they behave quite differently in other constructions. First and foremost, *half* is not a "classical quantifier", and cannot be used in construction like: *all/some/*half apples are red*.

On the other hand, there are several constructions in which *half* can be used, but most of the other "quantifiers" cannot. Together with *all* – the word *half* is the only word for which the *of* can be dropped: *all/*some/half the apple*. The question is whether this is even the same construction as in (3), since there are so few words with which this is possible. The construction is more common with a plural NP: *twice/double the apples*.

To make matters worse, *half* and *all* do not really behave the same in this construction: with *all* you have to use a definite NP, but with *half* you can also use an indefinite NP: **all/half an apple*. And *half* can even be used as an adverb with the same meaning, whereas none of the other ones can: *I half read this book,* and *This book is half red*.

In short, *half* belongs to a class of words that has exactly one element: there is no other word in the English language with exactly the same distributional behaviour. This is not a specific property of *half*, the exact combinatorial properties of similar words like *all*, *most*, *many*, *double*, etc. differ from each other in similarly subtle ways, as shown in table 1.

	half	all	some	most	many	several
~ the book	+	+	-	-	-	-
~ a book	+	-	-	-	+	-
~ these books	+	+	-	-	-	-
a ~ book	+	-	+	-	-	-
~ books	?	+	+	+	+	+
~ of this book	+	+	+	+	+	-
~ of a book	+	?	?	?	-	-
~ of these books	+	+	+	+	+	+
~ read a book	+	-	-	-	-	-

Table 1. Distributional Behaviour of some Quantifiers

2.3 Semantic Restrictions

In order to explain the difference in distributional between the different quantifiers in table 1 by the (partial) use of word classes, we have to assume that either (a) they all belong to a different (sub) class, or (b) that there are additional motivations why they behave differently apart from their word class.

Saying that there is a special word class for *half* does not appear a very attractive option. It would mean that all combinatorial behaviour of *half* has to be explained by special rules involving the *half* class, which would be potentially reasonable if it were the only "unique" word, but it is not: you would also need special classes for the other words in table 1, and many others as well.

We can also say that the incorrect sentences in table 1 are grammatical, but semantically blocked. For instance, *several* attributes something to a group of object, which cannot be interpreted when talking about only one book. But there are two reasons why this solution is unsatisfactory. On the one hand, it is difficult to explain semantically why you can say *I read half a book* to indicate you read it half, but not *I read *all/*most a book* to say that you got almost or completely through it. And on the other hand, semantically off sentences like the classical *green ideas sleep furiously* can typically be made acceptable in the right context, but *I read some these books* seems much more ungrammatical and cannot be reinterpreted to make it correct.

Another explanation for the differences in table 1 is to say that although all the words belong to the same word class, they have unique combinations of features that explain their different behaviours. However, a feature system that gets all the facts in table 1 right has to be quite rich, and as far as I know, no existing framework can deal with these phenomena. Moreover, even if it would turn out to be possible to get the feature system dealing with this problem, it would not really answer the question of what the word class of *half* actually is: how it is possible that *half* can be argued to belong to most of the word classes without really being polysemous.

3. Corpus Pattern Analysis

Corpus Pattern Analysis (CPA) is a research method introduced by Patrick Hanks (2004) to extract semantic (and grammatical) behaviour of words from corpus in terms of so-called *corpus patterns*. Corpus patterns are descriptions of the combinatorial characteristics of words, embedded in a theory called Norms and Exploitations. Different from other frameworks for describing combinatorial behaviour, such as the work by Mel'cuk *et al.* (1984), CPA is intended to work not on introspection, but purely on the basis of corpus analysis. One of the main goals of CPA is the construction of a Pattern Dictionary of English Verbs (PDEV).

Corpus patterns define the behaviour of words (verbs) in terms of what are basically subcategorization lists: they specify how many arguments a verbs takes, and what the semantic types of those arguments are. An example of a corpus pattern is given in (4), slightly simplified from Hanks & Pustejovsky (2006). This pattern states that the verb *fire* can be used in sentences like *John fired his gun at the tree*, where the first argument is a person, the second argument is a type of firearm, and the third (optional) argument is a physical object.

(4) [[Person]] **fire** [[Firearm]] (at [[PhysObj]])

Given that CPA is a corpus driven framework, the pattern in (4) is not meant to explain sentences like *John fired his gun at the tree*, since it is a sentence I constructed to explain the possible uses of (4). Rather, it is supposed to model the behaviour of the verb *fire* in actual corpus example of the verb, such as given in table 2, taken directly from PDEV:

and some buildings were burned. Guns were	**fired**	at the police, causing injuries to several
thought to be coming from a ship in distress,	**firing**	its guns to attract attention, so the people

touch of unreality for me to be able to	**fire**	my machine gun at everything I saw without
more interested than afraid. Some guns were	**fired**	. I wondered whether the flak was accurate
constables in the `strong-room': `the prisoners	**fired**	several pistols loaded with powder and

Table 2. Corpus examples of the pattern [[Person]] fire [[Firearm]] (at [[PhysObject]])

Some formal models of argument structure, such as for instance the theta system by Reinhart (2000) attempt to capture the intrinsic behaviour of verbs by assigning a single argument structure to a verb intended to capture (and explain) all possible uses of the verb. In CPA, on the contrary, verbs get assigned multiple patterns and the different patterns assigned to a verb are intended to explain the different uses of the verb. For instance, apart from the pattern in (4), the verb *fire* also has the pattern (5) assigned to it.

(5) [[Person]] **fire** [[Projectile]] (off) (from [[Firearm]])

The distinction between patterns (4) and (5) is intended to explain why there is a difference in interpretation between the following sentences:

(6) John fired his gun (at the tree)
(7) John fired a bullet (from his gun)

In fact, the distinction between (4) and (5) is intended to take the place of the distinction between two different word senses dictionaries traditionally postulate to account for sentences (6) and (7). The theory of CPA, alongside with, for instance, the dot object theory of Pustejovsky (1995), attempts to break into the traditional notion of a sense enumerative lexicon.

The patterns in (4) and (5) are not the only patterns available for *fire*: there are also other patterns, as the pattern in (8). In total PDEV lists 16 different patterns for *fire*.

(8) [[Person]] **fire** [[Person]] (from [[Job]])

One of the important characteristics of CPA is that the attribution of corpus patterns to a verb is not done by means of introspection. The work by Levin mentioned before is heavily corpus motivated: Levin hardly ever uses a made-up example, but always looks for real corpus examples to justify the syntactic behaviour of verb. CPA is corpus driven in a more direct fashion: the corpus patterns are intended to emerge from the corpus data and not merely verified by the corpus. The patterns are defined by looking through a closed list of corpus examples of a given verb, Hanks typically uses 250 concordances from the British National Corpus, and then looking through those examples one by one for regular patterns in the usage of the verb. That is to say, where the verb classes of Levin are based on the various ways in which a verb can be used, the patterns in CPA are descriptions of how the verb *is* used. Uses that are expected to be grammatical but are not evidenced by the corpus are ignored, and on the flip side, all uses of the verb have to be accounted for, even if they are metaphorical or playful.

Since CPA treats a closed set of corpus examples complete, the result is statistically significant: for each verb, CPA not only renders a list of all the patterns in which a verb can be used, but also the relative frequency of those patterns. For instance for the verb *fire*, PDEV shows that of the three patterns mentioned here, pattern (5) is the most common one, being used in 31% of the anatoted phrases, whereas patterns (4) and (8) account for 7% and 11% respectively.

3.1 Patterns and Surface Syntax

Corpus patterns represent sequences of words that can be recognized in a corpus: the pattern in (8) captures the fact that in an English corpus, we find sentences with contain an NP denoting a person, followed by a form of the verb *fire*, followed by a second NP denoting a person. In that sense, corpus pattern can be compared to the pattern of a theory called *lexicosyntactic patterns extraction* (LSPE). One of the most well known applications of LSPE is the extraction of hyperonymy relations from corpora (Hearst, 1992). The idea is that a pattern as the one in (9) and (10) can be used to extract hyperonymy relations from corpora.

(9) N_1 and other N_2
(10) N_1 like N_2

Sentences like (11) and (12) implicitly state that a trout is a type of fish, without being explicitly definitional phrases. And the relation between *trout* and *fish* can be extracted from such sentences by only focussing on sequences of words matching (9) and (10).

(11) This area contains some of our favorite recipes for Salmon, *Trout and other fish*.
(12) They are also called fish eagles because sea eagles eats largely on *fish like trouts* and salmon.

In LSPE, no deviation from the patterns is allowed. That is to say, even a sequence like *trout and other large fish* does not match the pattern in (9), since *large* is not a noun; *fish* of course is a noun, but it is not the first word following *other*. On the other hand, every sentence that has the sequence of words in pattern (9) will be considered a match, meaning that under pattern (9), the sentence in (13) will be interpreted as stating that a *net* is a type of fish, which is clearly a wrong conclusion.

(13) We manufacture and repair trawling nets, drag nets, gill nets, casting *nets and other fish* netting equipment.

Although CPA is a corpus driven approach, the patterns are not direct reflections of the corpus like the patterns of LSPE. The patterns of CPA do directly reflect a sequence of words. Firstly, the slots in the patterns are referential expression, which can be complex strings. So in (13), *nets* could never be a filler for a slot, but *casting nets* could. Secondly, there can be material between the subject and the verbs without affecting the pattern, such as for instance in (14), where *the passenger* is still the subject of *fire*.

(14) … the car stopped, the passenger got out and fired a Kalashnikov rifle at the police car.

Thirdly, the subject and internal object can be shifted around by means of for instance focus fronting, and the sentence will still match the pattern even when the internal object is placed to the left of the verb. And last but not least, the pattern is intended to even match phrases where even the syntactic roles are affected: a sentence can be used in an impersonal construction without a subject, or even in a passive construction with the subject and the object reversed while still exemplifying the same pattern. So (15) is an example of pattern (4) with an unexpressed first argument.

(15) he was in police custody when the gun was fired and the unfortunate …

In that respect, corpus pattern are not pattern in a computational sense, but underlying subcategorization frames described from a lexicalist perspective. They are more comparable with the kind of subcategorization that is described in HPSG (Pollard & Sag, 1994), but then with semantic restrictions on the arguments. They also resemble the characterization of verbs

in terms of thematic roles with selectional restrictions on the arguments as for instance in the entry in VerbNet corresponding to the pattern for *fire* given in (8) is shown in figure 1.

Class Fire-10.10		
AGENT [+ANIMATE | +ORGANIZATION], THEME [+ANIMATE | +ORGANIZATION], SOURCE [+ORGANIZATION], PREDICATE		
Members: can, dismiss, drop, expel, fire, force_out, oust, remove, sack, send_away, unseat		
Frames:		
Example	Syntax	Semantics
I fired two secretaries	Agent V Theme	CAUSE(AGENT, E) LOCATION(START(E), THEME, ?SOURCE) NOT(LOCATION(END(E), THEME, ?SOURCE))
I fired two secretaries from the company	Agent V Theme {from} Source	CAUSE(AGENT, E) LOCATION(START(E), THEME, SOURCE) NOT(LOCATION(END(E), THEME, SOURCE))

Figure 1. Simplified VerbNet entry for *fire* 10.10

VerbNet itself is, in turn, based on the classification of English verbs by Levin (1993). The frames in figure (1) closely resemble the patterns in CPA. However, there are differences. On the one hand, several patterns that would be considered the same in CPA are split out in VerbNet, such as the two patterns for *fire* given above. And VerbNet has a level that CPA, at least in its current form, is missing: frames are grouped into classes, which contain patterns that express the same structure in different ways. In the system of Levin, such uses are related to each other by means of verb alternations.

3.2. Exploitations

The corpus patterns of CPA provide a way to model the semantic selection in the subcategorization of verbs, and as such, are comparable to the observations already made by Pesetsky (1982) that verbs do not only select syntactically (c-selection), but also semantically (s-selection) and lexically (l-selection). As is well known, despite the intuitive appeal of s-selection, one of its major drawbacks is that it is not a rigid type of selection like c-selection. The subject of *drive* has to be an NP for the sentence to be grammatical. And typically, it will be a +human subject. But there is a great many ways to get a non-human subject for *drive*, so any sentence violating s-selection is marked at best, but certainly not incorrect or unusable.

To account for deviations from the semantic subcategorization defined by the patterns, CPA uses a notion of *exploitations*. The idea behind exploitations is that the pattern in (4) only describes a *typical* use of the verb *fire*. However, it is possible to use the pattern in (4) with some deviations, for instance when the subject is metonymically referring to a person, or where there is only a metaphorical firing going on, or where the internal object is not really a firearm, but merely something from which projectiles can be ejected. In those kinds of cases, we say that the sentence is exploiting the pattern in (4) creatively. Some example of exploitations of patterns (4), (5), and (8) is given in (15) and (16), all taken from PDEV.

(15) … seems to have `gone off on top doh' and fired all his big guns before has…
(16) The first election shots had been fired.

Both of these examples are cases in which metaphoric reference is made to firing of guns, but no actual firing is taking place.

4. Corpus Pattern Grammar

The corpus patterns of CPA are a way to do away with verb classes: by modelling the syntactic behaviour of verbs directly, there is no need to attribute specific verb classes to a verb: *fire* is simply a verb, lexically specified as to the kind of syntactic behaviour it displays. There are likely to be natural classes of verbs in the sense that many verbs are will end up syntactically behaving in the same way. However, in CPA such classification of verbs would be post factum: it is only after attributing corpus patterns to a verb based on the evidence from the corpus uses that one can compare different verbs to see if the same pattern is uses for a wider class of verbs.

To make CPA usable for a description beyond verbs, it is necessary to extend to framework to a slightly more grammatically oriented system, which we might call *Corpus Pattern Grammar* (CPG). Despite the fact that CPA places itself explicitly against the "syntacticocentric" tradition, the theory can be viewed as a grammatical framework. This section will provide a course sketch of how CPG could be used to characterize the word *half*. Before the sketch of the framework itself, the next section will first demonstrate its appeal as a grammatical framework.

4.1 Exploitations and Coercion

One of the appealing things behind reinterpretations in the form of exploitations is that it can be used to account for creative or untypical use of language. Take, for instance, a semantically atypical sentence like (17).

(17) They fired the clown from the cannon.

With a description in corpus patterns, the semantic oddness of (17) is a result of the fact that (17) does not match any of the patterns for *fire*. Although the subject is a person (or rather, a group of people), the sentence does not match any of the patterns (4), (5) or (8): a clown is neither a firearm nor a projectile, and a cannon is a firearm, but not a job. Therefore, the sentence is not accepted as a normal use. In order to make it interpretable, it has to be forced into one of the existing patterns.

The oddness of sentence (17) is reminiscent of the classical example by Nunberg (1979), who has ham sandwiches do things they are not typically expected to:

(18) The ham sandwich is sitting at table 7.

There are typically two ways to deal with the sentence in (18): the solution proposed for instance by Nunberg himself is that the sentence gets re-interpreted at the pragmatic level. The other option is to attribute it to coercion mechanisms within the syntax itself, as proposed for instance by Jackendoff (2002). In the proposal of Jackendoff, the argument *ham sandwich* gets reinterpreted as a *person who ordered a ham sandwich* in what he calls "enriched composition". The basic intuition behind this is that the subject in (18) is expected to be agentive, which *ham sandwich* cannot be. It is this mismatch that triggers the reinterpretation.

In CPG, the reinterpretation of (18) would follow the second type of solution, and is driven by the verb: the verb *sit* has a corpus pattern like in (17).

(17) [[Human | Animal]] **sit** [NO OBJ] ([Adv[Location]])

Because the word *ham sandwich* is not of the semantic type [[Human | Animal]], the pattern in (17) cannot be (directly) applied to sentence (18). In order to make it interpretable, we have to interpret (18) as an exploitation of pattern (17), and coerce the word *ham sandwich* into the

right semantic type. That is to say, we have to reinterpret *ham sandwich* as a human or animal. This of course does not by itself say *how* to interpret a ham sandwich as a person, so the result will not be a person ordering a ham sandwich, but just any interpretation of a ham sandwich as a person (or animal). So the reinterpretation in CPG is weaker than what Jackendoff proposed, but does allow for (18) to refer to someone who is just called a ham sandwich, looks like a ham sandwich, or even is an actual ham sandwich in a fairy tale.

With the same type of mechanisms, we can (re)interpret (17) as well. Since there are various patterns available, there are also various ways to interpret (17). Here is an informal description of three possible interpretations of (17) according to the available patterns.

(8) [[Person]] **fire** [[Person]] (from [[Job]])

The first option is to attempt to interpret (17) with pattern (8). Since a canon is not a job, it is not directly interpretable, so in order to interpret it we have to coerce *cannon* into a type of job. Imagine for instance a group of circus artists who start a company to market themselves. And between themselves, they refer to this company as *the cannon*. In those circumstances, sentence (17) can be used when they have to let go of the clown.

Since the direct object in (8) is optional, there is another way of interpreting (17) as a use of pattern (8), which does not even require coercion: *from the cannon* could be a modifier of *the clown*. Imagine, for instance, a circus director who has too many clowns in his employment, and has to let go of one of them. If he fires the clown that at the time is sitting on top of/hiding behind the cannon, one could use (18) as a specification of which of the clowns he fired.

(4) [[Person]] **fire** [[Projectile]] (off) (from [[Firearm]])

The most obvious reinterpretation of (17) is, however, to coerce it into pattern (4). Since a cannon is a type of firearm, the only mismatching word is the word *clown*. And since it is a known circus act to shoot clowns from cannons as if they were projectiles, the interpretation that (17) is likely to get under normal circumstances is one in which *clown* is taken to be a type of projectile.

As shown, using corpus patterns as a basis for a grammatical model that allows for coercion correctly predicts the possible interpretations of "semantically odd" sentences. It (correctly) predicts that in (17), *clown* is interpreted as a type of projectile, and in (18), *ham sandwich* is interpreted as a type of human or animal. However, it does not specify in and by itself why it is possible to reinterpret *ham sandwich* as a person, nor what the relation between the person and the ham sandwich is, nor does it pose any restrictions on the type of reinterpretations that can be made. For a discussion of the type of coercion one would expect, as well as a model on how to interpret the coerced entities, see for instance the work by Pustejovsky & Jezek (2004).

4.2. Grammatical subcategorization

The patterns of CPA specify the necessary and optional arguments of a verb in semantic terms. The entities in double square brackets stand for referential expressions, hence in principe NPs. However, the framework does allow some other types of specifications as well, as in the examples below, all taken from PDEV.

Firstly, it is possible to indicate obligatory or optional words, such as *on* and *eyes* in pattern (19), or even words of a group of words, such as REFLDET in (17), which indicates any reflexive determiner, meaning that pattern (19) represents sentences such as (20), where *your*

is the reflexive determiner, and should be seen as a shorthand for a set of words: {my|your|his|her|our|their|its}.

(19) [[Human]] **feast** {REFLDET eyes} {on [[Physical Object|Stuff]]}
(20) If you are just happy to dream, feast your eyes on this selection of motoring…

Secondly, it is in principle possible in CPA to add grammatical restrictions to the semantic classes. For instance, in pattern (21), the things that clog have to be either plural or mass. Although this is in part a grammatical restriction, it is not purely syntactic: the argument of *with* in (21) has to be an NP with a plural or mass *referent* of the type [[Physical Object]].

(21) [[Location | Artifact | Body Part]] **clog** [NO OBJ]
{(up)} {(with [[Physical Object = PLURAL | MASS]])}

Apart from element in double square brackets, patterns can also contain single square bracket elements, as for instance in (22). The argument [that-CLAUSE] is not a referential expression, but rather a categorical (syntactic) type. In that way, a CPA pattern combines elements of the traditional notions (Pesetsky 1982) of c-selection (categorial type, [that-CLAUSE]), s-selection (semantic type, [[Location]]), and l-selection (lexical element, with).

(22) [[Human | Institution | Document]] **claim** {[that-CLAUSE | QUOTE]}

It is possible to combine single and double square brackets, as can for instance be seen in pattern (23). The [NO OBJ] is there to indicate explicitly that *sit* in this pattern is an intransitive verb, and the only argument it can take is an *adverbial phrase* denoting a location, such as *at the table*, or *upstairs*.

(23) [[Human | Animal]] **sit** [NO OBJ] ([Adv[Location]])

In a sense, all argument roles are syntactically typed, except that in the "normal" case, the syntactic typing is left out, but in principle, one should interpret [[Human]] to be a shorthand for [NP[Human]].

With the categorical typing of arguments, it is possible to define patterns for words other than verbs. For instance, for an adjective like *savoury*, one can define that it normally takes a noun denoting an edible substance as its argument, as in (24). Of course, this pattern does not capture the entire distributional behaviour of *savoury,* it does not, for instance, specify whether *savoury* can be used predicatively, or whether it is a gradable adjective or not. However, it does provide an important part of the selectional restrictions.

(24) **savoury** [N[Foodstuff]]

When taking CPA beyond the scope of verbs, it becomes necessary to add real syntactic restrictions to the categories. For this, we introduce the following notation: [N+plural] indicates a noun in grammatical plural form. To exemplify the difference with the indications within the semantic part of the pattern, consider (25). The Dutch adjective *zwanger* (pregnant) selects a noun which has has a female referent of type [[Animal]] or [[Human]].

(25) **zwanger** [N[Animal|Human = FEMALE]]

Because the selection for females is within the s-selection part, the word *zwanger* is said to combine also with words that are grammatically masculine but semantically neuter, such as *meisje* (little girl). If *zwanger* would select for a grammatically female term independently of its referent, we would have to indicate [N+female[Animal|Human]] instead. So the indication

+female with the syntactic type is a grammatical feature, and the indication =FEMALE with the semantic argument is a semantic feature.

4.4. Half (of) a pattern

Now let us finally turn to the description of *half*. The basic pattern for *half* is rather permissive: it optionally l-selects for the word *of*, and then c-selects for any NP, as in (26).

(26) **half** {of} [NP]

The fact that there is no restriction on the NP argument of *half* distinguishes it from *all*, which only takes a definite NP, as in (27), and furthermore has a use as a quantifier that *half* does not have, as in (28).

(27) **all** {of} [NP+definite]
(28) **all** [N+plural]

If we consider for instance the use of the word *several*, it cannot drop the word *of*, and furthermore only takes a plural definite NP, as in (29), and it has a pattern similar to (28) as well.

(29) **several** of [NP+plural+definite]

Furthermore, contrary to *several* or *all*, the word *half* can also be used as a modifier of a wide range of things: a noun (a half book), an adjective (half green), an adjective (half jokingly), or a verb (he half read this book).

(30) **half** [N|Adj|Adv|V]

The fact that with the pattern in (30), the word *half* can be used in front of a noun means that it can be used in at least some positions where adjectives typically can appear as well, but the pattern is much more permissive than that of an adjective. On the same footing, one could call the use of *half* in (26) either a noun or a predeterminer if so desired, but there is no need to do so: the patterns in (26) and (30) characterize the distributional behaviour of *half* without the need to (misleadingly) refer to it as a noun, an adjective, a (pre)determiner, a adverb, a quantifier, or any other specific class.

5. Conclusions

As I hope to have show in this article, the word *half* is a unique word in the English language: there is no other word quite like it. The same holds for the verb *fire*. However, although they do not share their behaviour with any other English word, they do share their uniqueness with many other words: it is quite common for a word to be unique.

Given that the word *half* is unlikely any other word, it seems to be a wrong and misleading question to ask which word class it belongs to. By describing the distributional behaviour of the word *half* directly using corpus patterns, as shown in section 4.4, there is also no need to assign it a word class. This is not intended as an attempt to get rid of word classes altogether: it is difficult to even imagine how to describe the behaviour of *half* without referring to the large open classes of nouns, verbs, and adjectives. However, it does mean that the small, closed "classes" of words are not really neatly order into classes. Therefore, the use of word classes might best be restricted to the large open classes.

The more grammatically oriented extension to CPA sketched here is only a grammar in a weak sense of the word: it allows the application of CPA to a wider range of words, and not be a system restricted to verbs. In order to extend this to a full grammar that could be used for computational purposes, one would have to adopt some strategies from HPSG. It would be necessary to have an implementation of (feature) agreement. Also, it would be necessary to provide the "output" of a pattern with a syntactic type: verb patterns describe sentences, but not all other patterns do. In order to form sentences, the other classes would have to combine with verb pattern to create sentences.

It should be noted that, contrary to the philosophy of CPA, the pattern in section 4.4 are not the result of an extensive corpus study. This is in part due to a problem with numbers: the British National Corpus has 97 occurrences of the verb *abbreviate*, and according to PDEV, there are 3 patterns for *abbreviate*, the least frequent of which has only 3 examples. The word *half*, on the other hand, has 29.863 occurrences. If one were to look only at the first 250 of those, it is likely that some patterns that are much more common than the most common pattern for *abbreviate* nevertheless will not appear in that sample.

References

Fillmore, C. (1968). "The Case for Case". In Bach and Harms (Ed.): *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston, 1-88.

Hanks, Patrick. (1994). "Linguistic Norms and Pragmatic Exploitations, or Why Lexicographers need Prototype Theory and Vice Versa" *in* F. Kiefer, G. Kiss, and J. Pajzs (eds.), *Papers in Computational Lexicography: Complex '94*, Budapest, Hungary.

Hanks, Patrick. (2004). "The Syntagmatics of Metaphor and Idiom". *International Journal of Lexicography*, vol. 17, 254-274

Hanks, Patrick & James Pustejovsky (2005). "A Pattern Dictionary for Natural Language Processing". *Revue Française de Langue Appliquée*, vol. 10:2.

Hearst, M. A. (1992). "Automatic Acquisition of Hyponyms from Large Text Corpora". In: *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pp. 539–545. Nantes, France.

Jackendoff, Ray. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford: Oxford University Press.

Kipper-Schuler, Karin. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania.

Koenig, Jean-Pierre & Anthony R. Davis. (2001). Sublexical Modality and the Structure of Lexical Semantics. *Linguistics and Philosophy*, 24:71–124.

Levin, Beth. (1993). *English Verb Classes and Alternation: a Preliminary Investigation*. Chicago: The University of Chicago Press.

Mel'cuk, I. et al 1984, 1988, 1992. Dictionnaire Explicatif et Combinatoire du Franfais Contemporain: Recherches Lexico-Semantiques I, II, III. Montreal: Les Presses de l'Universite de Montreal.

Pesetsky, D. (1982). *Paths and categories,* PhD thesis, MIT.

Pollard Carl & Ivan A. Sag (1994): *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Pustejovsky, James & Elizabete Jezek. (2004). "Semantic Coercion in Language: Beyond Distributional Analysis." *Italian Journal of Linguistics,* vol. 20:181-214.

Steedman, M. (1987), 'Combinatory Grammars and parasitic gaps'. *Natural Language & Linguistic Theory* 5, 403–439.