What lexical sets tell us about conceptual categories

Elisabetta Jezek[1]
Patrick Hanks[2]

Abstract

It is common practice in computational linguistics to attempt to use selectional constraints and semantic type hierarchies as primary knowledge resources to perform word sense disambiguation (cf. Jurafsky and Martin 2000). The most widely adopted methodology is to start from a given ontology of types (e.g. Wordnet, cf. Miller and Fellbaum 2007) and try to use its implied conceptual categories to specify the combinatorial constraints on lexical items. Semantic Typing information about selectional preferences is then used to guide the induction of senses for both nouns and verbs in texts. Practical results have shown, however, that there are a number of problems with such an approach. For instance, as corpus-driven pattern analysis shows (cf. Hanks et al. 2007), the paradigmatic sets of words that populate specific argument slots within the same verb sense do not map neatly onto conceptual categories, as they often include words belonging to different types. Also, the internal composition of these sets changes from verb to verb, so that no stable generalization seems possible as to which lexemes belong to which semantic type (cf. Hanks and Jezek 2008). In this paper, we claim that these are not accidental facts related to the contingencies of a given ontology, but rather the result of an attempt to map distributional language behaviour onto semantic type systems that are not sufficiently grounded in real corpus data. We report the efforts done within the CPA project (cf. Hanks 2009) to build an ontology which satisfies such requirements and explore its advantages in terms of empirical validity over more speculative ontologies.

Keywords: ontology – semantic type – lexical set – coercion – corpus analysis – computational lexicography

[1] Department of Theoretical and Applied Linguistics, Pavia University, Italy: jezek@univpv.it
[2] Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic: patrick.w.hanks@gmail.com

1. Background and motivation

Corpus linguistics has shown, if nothing else, that word use is very highly patterned, although the boundaries of each pattern tend to be fuzzy and variable. A lexicographical task for the future, therefore, is to use corpus evidence to tease out the different patterns of use associated with each word in a language and to discover the relationship between meaning and patterns of use[3]. At the same time, rare and unusual boundary cases at the edges of patterns must be identified for what they are and not allowed to interfere with reliable accounts of the normal structure of clauses and use of words in a language.

The primary goal of the CPA (Corpus Pattern Analysis) project (cf. Hanks 2009) is to create a repository of all the normal patterns of use of all the normal verbs in English, called the Pattern Dictionary of English Verbs (PDEV). A pattern, in our sense, is a semantically motivated and recurrent piece of phraseology. Patterns for verbs consist of valencies plus semantic types of arguments within valencies – populated by lexical sets, e.g. paradigmatic sets of words occupying the same syntagmatic position. In example (1) below we provide a pattern for the English verb *attend* (corresponding to its sense 'be present at'):

(1) *attend*
[[Human]] attend [[Activity]][4]
Lexical set [[Activity]] = {meeting, conference, funeral, ceremony, course, school, seminar, dinner, reception, workshop, wedding, concert, premiere …}

Meanings of verbs in CPA are associated, not with verbs in isolation, but with verbs in different patterns—that is, in relation to nouns that co-occur with each verb in different clause roles: subjects, direct objects, and propositional objects or 'adverbials'. Nouns are therefore grouped together into lexical sets according to how they affect the meanings of verbs. The hope is (or was) that these lexical sets could be grouped into a hierarchical ontology in order to predict the meaning of verbs according to their context and to be able to model verbal polysemy in a way which is computationally tractable. In this paper we examine two of the problems that stand in the way of fulfilling this goal. The first is, as corpus-driven pattern analysis shows (cf. Hanks et al. 2007), that the paradigmatic sets of words that populate specific argument slots within the same verb sense do not map neatly onto conceptual categories, as they often include words belonging to different semantic types. The second is that the internal composition of these sets changes when one moves from verb to verb, so that no stable generalization seems possible as to which lexemes belong to which semantic type (cf. Hanks and Jezek 2008). We claim that these problems are not accidental facts related to the contingencies of a given ontology, but rather the result of an attempt to map distributional language behaviour onto semantic type systems that are not sufficiently grounded in real corpus data. We report the efforts done within the CPA project, based on the theory first outlined in Hanks (1994) and on work that was first reported in Pustejovsky, Hanks, and Rumshisky (2004), to build an ontology which satisfies such requirements and explore its advantages in terms of empirical validity over more speculative ontologies.

[3] Throughout this paper, we use "meaning" and "sense" as approximate synonyms. This is in line with the theoretical assumption that words in isolation have an abstract meaning potential rather than a (list of) clearly defined meaning(s), and that the precise contribution that a word makes to the interpretation of a complex linguistic expression is generated in context.

[4] The names of semantic types are conventionally written in double square brackets with capital initial letters.

2. What counts as an ontology?

The term *Ontology* is a fashionable term these days. It has at least three meanings. We start, therefore, by summarizing these different meanings and selecting the one in which we shall use the term in this paper.

In philosophical parlance, *ontology* is a mass noun denoting everything that exists (both physical and metaphysical), and hence the entire subject matter of philosophical enquiry. This concept of *ontology* goes back to Aristotle. It is not a sense that need concern us any further here.

From this traditional philosophical definition it was but a short step (in the 20th century) to using the term as a count noun to denote an ordered collection of the contentful terms of a language, both abstract and concrete. This is the sense in which the term *ontology* is applied for instance to WordNet. It is the basis of the sense in which we shall use the term in this paper, but, rather than engaging in armchair speculation about semantic relations and synonym set, we use corpus evidence to try to find out which words are collocates of each other and how different collocations are associated with different meanings. In other words, we focus specifically on identifying the empirical, corpus-driven analysis of paradigmatic sets of nouns that select a specific meaning of a verb when the two words (the noun and the verb) are used together in a sentence. We then explore how the lexical set may be grouped according to their semantic types.

A third sense of *ontology* has grown up among researchers working on the semantic web, by whom it is used to denote large sets of computationally tractable objects, including for example a particular person's name, an appointment between a named person and a named doctor, the date, time, and location of such an appointment, the doctor's appointment book, a set of names and addresses, the doctor's reference manuals, holiday guides, lists of hotels, documents used for reference or other purposes, medicines and other named business products, names of business enterprises, government departments and other institutions, and so on. This too is a sense that will not concern use here, beyond remarking that this new sense of *ontology* has nothing to do with words and meanings and has rich potential for confusion.

An ordered collection of content words does not necessarily have to be arranged as a hierarchy of concepts, but in practice this is what is done, with some plausibility. Terms in such a hierarchy are arranged as hyponyms of other terms, according to their semantic type: a *canary* is kind of *finch*, a *finch* is a kind of *bird*, a *bird* is a kind of *animate entity*, an *animate entity* is a kind of *physical object*, and so on. *Finch* may be classed as a co-hyponym (under *bird*) of *parrot, hawk*, *cuckoo, penguin,* etc. It must be noted that WordNet itself introduces a large number of scientifically motivated fine-grained subdivisions of the term *bird*, e.g. *passerine, carinate, ratite, gallinaceous*, and other terms that are not in ordinary usage and therefore not useful for collocational analysis. WordNet reflects scientific conceptualization, not ordinary usage.

An arrangement of terms of this sort is sometimes called an IS-A hierarchy, for obvious reasons. IS-A hierarchies work comparatively well when applied to natural-kind terms such as plants and animals and artefacts such as tools. Disturbing questions arise, however, when, as in the case of WordNet, an attempt is made to arrange <u>all</u> the terms of a language in an IS-A hierarchy. For example, abstract terms do not lend them themselves readily to hierarchical arrangement: is an *idea* a *concept* or is a *concept* an *idea*, or are they both co-hyponyms of

something else? There does not seem to be any obvious answer to such questions, which multiply alarmingly as more and more terms are examined.

3. Why bother with a hierarchical ontology at all?

Given the difficulties just alluded to, one might reasonably ask, why bother with a hierarchical ontology at all? And indeed, some computational linguists are now proposing that semantic types and hierarchical ontologies are unnecessary: semantic distinctions can be achieved, they say, by using very large corpora to group words into paradigm sets by computational cluster analysis. This, it seems to us, is throwing the baby out with the bathwater. The fact that a task encounters unexpected difficulties does not mean that is not worth doing.

There are two answers to the question just posed. The first is that some lexical sets are open-ended, with boundless potential for adding new members, so no amount of cluster analysis in corpora will provide a sound basis for identifying the set membership of these large sets. For example, many verbs have a strong preference in at least one of their senses for colligation in the subject or object clause role with terms denoting a human being. This preference, in some cases, contrasts with other meanings of the verb where the subject or object is not human. For example, it is the semantic type preference of the direct object that distinguishes toasting a person (= celebrate) from toasting a piece of bread (= cook under radiant heat). This distinction can only be expressed formally if the semantic type [[Human]] is available in contrast with other semantic types, e.g. (here) [[Food]]. An extensional definition of a semantic type such as [[Human]] (i.e. listing all the relevant lexical items) will inevitably fail to predict many of the lexical items that will quite normally occur in the direct object slot in relation to *toast* in this sense, and the same is true of all other verbs that have a sense that normally requires a human subject or object. The notion of listing extensionally all the names of all the people in the world—past, present, and future—and all the terms that have been or ever will be used to denote their status and roles (*parent, bride, judge, defendant, bricklayer, pianist, soloist, author,* etc.), is obviously absurd, but that is what would be needed if the semantic type [[Human]] were to be effectively replaced by a paradigmatic cluster of lexical items for identifying the ‘celebrate’ sense of the verb *toast*. Ontologies help prediction, and abandoning semantic types is abandoning a good principle.

On the other hand, an extensional definition of canonical members of the set of foodstuffs that are normally toasted is quite easy to compile by applying a tool such as the Word Sketch Engine (Kilgarriff et al. 2004) to one or more very large corpora. Although in principle open-ended, the list of canonical members of this set runs out fairly quickly: a list of about a dozen items (*bread, slice, sandwich, bun, bagel, baguette, crumpet, teacake, muffin, marshmallow, almond, walnut, ...*) identifies most of them and has strong predictive power: when *toast* is used in the sense ‘cook under radiant heat’, one of these words is very likely to be its direct object. Conversely, if one of these words is found in the direct object slot, the sense is almost certain to be ‘cook under radiant heat’ and not ‘celebrate’.

The second answer is that a hierarchical ontology enables relevant generalizations to be made and implicatures stated. For example, people *knock over* all sorts of physical objects: tables, chairs, tea cups, glasses of wine, pedestrians, lamps, buckets, vases, and plant pots (to name but a few). A hierarchical ontology enables the analyst to group together tables, chairs, lamps, vases, and plant pots as furniture and to distinguish the implicatures of knocking over

furniture from knocking over pedestrians (humans). The implicatures are only slightly different at this level, and could even be lumped together into a single sense. The implicatures of knocking over physical objects such as furniture, cups, glasses, and pedestrians have something in common that is not shared by the metaphorical sense, at a higher level of abstraction, of knocking over an abstract object such as a hypothesis.

4. Problems

Straightforward examples such as *toast* above have led to the expectation that semantic types such as [[Human]], [[Animate]], [[Artifact]], [[Physical Object]], [[Food]], [[Event]], [[Activity]], etc. can be used systematically for word sense disambiguation and that semantic typing information about selectional preferences can guide the induction of senses for both verbs and nouns in texts (Manning and Schütze 1999: 288).

Unfortunately, however, this expectation is often disappointed. Consider again *toast*. As we said, typically a person "toasts a [[Human]]" (=celebrate) or "toasts [[Food]]" (=cook under radiant heat) but in the list of the most salient object collocates of *toast* in the BNC (via Sketch Engine), no [[Human]] shows up[5]. Instead, we find nouns denoting relevant attributes of humans, such as *victory*, *success* and *health*:

(2) *toast*
Expectation: the direct object of *toast* will be either [[Human]] or [[Food]]
But the Word Sketch for *toast,* v., in BNC shows[6]:

- bread 20.59
- crumpet 19.31
- victory 15.48
- slice 12.71
- success 12.02
- future 8.47
- health 7.31

Similar remarks hold for the verb *attend*. Typically, a person 'attends' an [[Activity]] (meeting, lecture, funeral, coronation, etc.). However, there are many activities (e.g. route march, stroll, booze-up, suicide) that people do not 'attend', while some of the things that people do attend (e.g. school, church, clinic) are not [[Activity]], but rather [[Location]][7].

(3) *attend*
Direct Object:
a. [[Activity]]: meeting, wedding, funeral, mass, game, ball, event, service, premiere
b. [[Location]]: clinic, hospital, school, church, chapel

'About thirty-five close friends and relatives attended the *wedding.*'
'For this investigation the patient must attend the *clinic* in the early morning.'

[5] The settings we use in our corpus investigation are the following: minimal frequency 3, maximum number of items per grammatical relation 150.
[6] Numbers reported next to each lexical item represent salience scores (for information about the statistics used in the Sketch Engine, cf. http://trac.sketchengine.co.uk/wiki/SkE/DocsIndex).
[7] The data below is taken from the BNC corpus and presented adopting a layout proposed in Rumshisky et al. 2007.

'He no longer attends the *church.*'

The problems exemplified by *toast* and *attend* arise for different reasons. While with *toast* in (2) the focus is on the relevant attributes of the entity that is affected by the verb, rather than on the entity itself, in (3) the verb *attend* modulates the meaning of the nouns that fill the object position and causes words denoting [[Location]]s such as *school*, *church* and *clinic* to be reinterpreted as [[Activity]]s contextually.

The CPA (Corpus Pattern Analysis) project provides two steps for dealing with the phenomena illustrated by *toast* and *attend*:

i. Non-canonical lexical items are coerced into "honorary" membership of a lexical set in particular contexts, e.g. *school*, *church*, *clinic* are coerced into membership of the [[Activity]] set in the context of *attend*, but not, say, in the context of *arrange*.

ii. The ontology is not a rigid yes/no structure, but a statistically based structure of collocational preferences, which we call "shimmering lexical sets". Thus, each canonical member of a lexical set is recorded with statistical contextual information, like in (4):

(4) [[Activity]]:
... meeting <attend Obj. 431/3206; arrange Obj. 101/1747; ...>

In the following sections we discuss these issues in more detail and illustrate how steps i. and ii. are implemented in the project.

5. Encoding coercions in the Pattern Dictionary

Semantic coercion can be generally described as a modulation of the basic meaning of a word due to semantic requirements imposed by other words in a given context[8]. Coercion is a principled mechanism for accounting for the variety of interpretations that words exhibit in different contexts. In particular, coercion occurs when the meaning that a word exhibits in context is not inherent to the word itself (i.e. coded lexically) but rather the result of compositional processes induced by the linguistic co-text (i.e. by the semantics of the co-occurring words[9]. In a broad definition, coercion covers conventionalized sense modulations as those discussed in Pustejovsky 1995 and Copestake and Briscoe 1995, but also non-conventional (i.e. creative or dynamic) exploitations of conventionalized uses (i.e. novel uses of words), as described in Hanks (forthcoming).

There are several different sorts of coercions. *Coercion of semantic type* (in short 'type coercion') has been extensively discussed within the Generative Lexicon (GL) framework (Pustejovsky 1995, 2006). Type coercion is an operation of type adjustment that occurs when

[8] Since the notion of "basic meaning" of a polysemous word is extremely controversial, the following points should be made: A) It is the meaning that is most 'terminological' and least 'phraseological'. B) The basic meaning never has any metaphorical semantic 'resonance' with another meaning. Conversely, quite often, metaphorical meanings have semantic resonance with a basic, literal meaning. C) If a word has both concrete and abstract meanings (e.g. "grasp an object" vs. "grasp an idea"), the basic meaning is a concrete one. D). Quite often the basic meaning is NOT the most frequent one.

[9] Note that this criterion (linguistic co-text) excludes that meaning shifts due to pure pragmatic processes be treated as coercions, although it is clear that it is often very difficult to draw a dividing line between what is coded (dictionary) and what belongs to commonsense knowledge (encyclopedia).

none of the selectional preferences of a predicator match the type of a noun that it combines with in a particular text. In this case, type coercion is invoked to explain how a mismatching verb-argument combination can be interpreted.

A paradigmatic example of type coercion is *event type coercion*. This occurs for instance when a verb that normally selects an [[Activity]] as direct object (e.g. *finish* ***drinking*** *something*) is used in combination with an artifactual entity (e.g. *finish* ***one's beer***). In this case, the verb induces a re-interpretation of the noun from [[Physical Object]] to [[Activity]], so that it can successfully fill the object argument slot[10]:

(5) *finish*
Direct Object:
[[Activity]]: journey, tour, treatment, survey, race, game, training, ironing
[[Physical Object]]: penicillin, sandwich, cigarette, cake, dessert, food
[[Beverage]]: drink, wine, beer, whisky, coke

'when they finished the wine, he stood up.'
'just finish the penicillin first.'

What is interesting here is that the activity reconstructed by the operation of coercion is not a general event, but a specific activity that is conventionally associated with the object: *finish the wine* means 'finish drinking the wine', *finish the penicillin* means 'finish ingesting the penicillin as a medicine', and so on. In other words, the reconstructed event is an event in which the object is typically involved[11]. Within the GL model, it is assumed that these events are coded as Qualia relation in the noun's meaning (Pustejovsky 1995: 85).

A different example of type coercion is provided by the verb *ring*. In its 'call by phone' sense, *ring* selects for an object type [[Human]] and coerces all the nouns appearing as its direct objects (e.g. [[Institutions]], [[Locations]]) to this type:

(6) *ring*
Direct Object
a. [[Human]]: mother, doctor, Gill, Chris, friend, neighbour, director
b. [[Institution]]: police, agency, club
c. [[Location]]: bank, hospital, office, reception, flat, house; Moscow, Chicago, London, place

'I rang the *house* a week later and talked to Mrs Gould.'
'The following morning Thompson rang the *police.*'
'McLeish had rung his own *flat* to collect messages.'
'I said *Chicago* had told me to ring *London*.'

Here, the reinterpretation of the objects is felicitous, because *house*, *flat*, *Chicago* etc. are places where people live or work. *Rivers*, *forests*, and *deserts* are [[Location]]s, but phrases like "*ring the river/Rhine/forest/jungle/desert/Sahara*" do not occur. Thus, there is subset of nouns with semantic type [[Location]] that occur as direct object of the verb *ring*, and a second subset that do not.

[10] For a collection of examples of coercion extracted from corpora, including some of those reported here, cf. Pustejovsky and Jezek (2008).
[11] There are some difficulties with this formulation: however, we will not discuss them here.

Yet another example of coercion is provided by the verb *open* in its 'make accessible' sense (selected type: [[Container]]):

(7) *open*
Direct Object
a. [[Container]]: drawer, bottle, cupboard, envelope, folder, tin, can, box, fridge, bag, cage, suitcase
b. [[Beverage]]: wine, champagne, beer

'I opened the wine carefully.'
'Just as he was about to open the beer, the doorbell rang'

In this case, when the verb co-occurs with artefactual liquids such as *wine*/*champagne*/*beer*, it induces a reinterpretation of these nouns from [[Beverage]] to [[Container]].

Interestingly, examples (5) and (6) show that the kind of coercion differs depending on how the missing piece of semantic information is retrieved: while the [[(drinking) Activity]] and the [[Human (living in)]] are available as a part of the meaning of *wine* and *house* respectively, [[Container]] is not part of the meaning of [[Beverage]] and is introduced contextually[12].

Let us now look at how these phenomena are dealt with in The *Pattern Dictionary of English Verbs* (PDEV) To start with, the lexicographer has two main options to encode apparent mismatches between the selecting and selected type such as those mentioned above:

- the lexicographer may split a pattern into two more fine-grained patterns, each corresponding to a different verb sense, for instance "[[Human]] attend [[Activity]]" (= be present at) vs. "[[Human]] attend [[Location]] (= go regularly to)", or

- the non-canonical lexical items of the set may be recorded as a coercion to the expected semantic type - thus, *school, church, clinic* are coerced into 'honorary' membership of the [[Activity]] set, but only in the context of *attend* and not, say, in the context of *arrange*.

If the lexical analyst discovers that a verb exhibits a particular submeaning (i.e. a meaning that is a specialization of a more general one) because of repeated contexts, the first option is taken (even if the more fine-grained distinction is not recorded in any standard existing dictionary). On the other hand, if a particular word or group of words is a unique outlier, coercion is the best option.

It is not always obvious which of these two options is best. While some cases are clear-cut, others are not. For example, the verb *visit* selects both for [[Location]] and [[Human]] as its direct objects. Unlike *ring*, these two uses of *visit* seem to map quite neatly onto two distinct senses and patterns (e.g. 'go to and spend some time in a place for tourism or other purpose' and 'go to and spend some time with a person, typically for social reasons'):

(8) *visit*
[[Human]] visit [[Location]]
He visited Paris in 1912

[12] For a formal account, see Pustejovsky (2006).

[[Human 1]] visit [[Human 2]]
She visited her father regularly

Prototypically, people visit places for touristic reasons; they visit other people for social reasons. It is perfectly possible to *visit* Paris for business, social or political purposes, but this is not the normal phraseology.

In other cases, it is impossible to establish where the boundary between senses should be drawn, or indeed whether a distinction should be made at all. Thus, intuitively, *storms abating* and *riots abating* seem to be very different kinds of events and could be entered in the dictionary as different senses of the verb *abate*, corresponding roughly to a more literal and a more metaphorical sense. On the other hand, they have a common factor in that the meaning of *abate* in both cases is 'become less intense' and both a storm and a riot is an event, so the two phrases can equally rationally be treated as different aspects of a single sense of the verb.

There are often good reasons to lump different uses together, especially if one does not want to lose the semantic links that exist between them. In the case of *attend,* for example, *schools*, *clinic*s, and *churches* are [[Location]]s where particular [[Activity]]s take place (a class, a treatment, a mass and so on). It is to these [[Activity]]s that we refer to when we say that we attend such-and-such a location. So, as with *ring*, it makes sense to coerce all these nouns to the type [[Activity]] in the context of *attend*.

Once a lexicographer has chosen either to split the pattern or to lump some uses together, he or she can then proceed to encode coercions in any of three different ways in the Pattern Dictionary. These are:

A. alternation of semantic types
B. semantic roles
C. exploitations

A. Alternations of semantic type are regular choices of types within an overall pattern in relation to a target lexical item. Two common alternations are for instance [[Human | Institution]] (very frequent in the subject of a large number of verbs that denote a cognitive action) and [[Human | Body Part]]:

(9) *hope* [[Human | Institution]] hope ...
Organizers are hoping to convince the council | The *government* hopes to cover 15% of running costs

(10) *bleed* [[Human | Body Part]] bleed
The *man* may possibly have been bleeding | my *hand* had stopped bleeding some time ago

Alternations can occur in subject position, as in (9) and (10), or in object position, as in (11):

(11) *calm*
Direct Object
a. [[Human]]: child, people, crowd, troops, daughter
b. [[Emotion]]: nerves, fears, anxiety

'He calmed the *crowd* and continued talking*'*
'Calm your *nerves* by deep breathing'

As noted in Hanks (forthcoming), one of the major functions of semantic type alternations such as those reported in (11) is to focus attention on the relevant attributes of the entity that is affected by the event and can be accounted for in terms of meronymy – the part-whole relationship.

B. The distinction between a semantic type and a semantic role can be defined as follows. The semantic type is an intrinsic attribute of a noun, while a semantic role has the attribute thrust upon it by the context. Thus, for instance, the verb *arrest* is found in the following pattern:

(12) [[Human 1 = Police Officer]] arrest [[Human 2 = Suspect]]
'but what does this man want? He can not *arrest* everybody'
'The Germans *arrested* him and his wife two years later'
'indeed yesterday we unfortunately failed to *arrest* some very prominent IRA men'

There is nothing in the intrinsic semantics of *The Germans* to say that they were acting as police officers. This is a role assigned by context – specifically, by the selectional preferences of the verb *arrest*, which expects a subject with such characteristics.

A more complex case is found in (13) and (14), where the semantic roles [[Projectile]] and [[Employee]] are not assigned to the subject by the verb in isolation, but by the VP complex ('fire from a cannon' and 'fire from her job' respectively):

(13) fire [[Projectile]]
'Jane was fired from a cannon yesterday'

(14) fire [[Employee]]
'*Jane was fired from her job yesterday'*

Sometimes, both semantic role and alternation apply in a single pattern. For example in (15) the role of [[Driver]] is assigned contextually to the type [[Human]] by the verb *accelerate*.

(15) [[{Human = Driver} | Vehicle]] accelerate
'She was pressed back against the seat as Fergus accelerated again'
'I watched the car accelerate down the road'

Similarly, in (16) the pattern alternates between a [[Plane]], the [[Human]] who pilots it (contextual role: Pilot), and the [[Humans]] who are transported by it (contextual role: Passengers):

(16) [[Plane | {Human = Passenger | Pilot}]] land ([Advl[Location]])
'UN planes had already started landing at the airport'
'the pilot decided to land on the beach'
'we landed at Cardiff'

C. An exploitation is a deliberately creative or unusual use of a norm – an established pattern of word use. While alternations are regular choices of elements within an overall pattern,

exploitations are dynamic and creative choices. Newly created metaphors are exploitations, but there are many other kinds of exploitation, too, some of them quite unremarkable, as in (17).

(17) It is not a natural use of one's land to *cultivate* weeds in bulk.

The preferred semantic type of the direct object of the verb *cultivate* is [[Plant]], so at first sight it looks as if this preference has been satisfied. But what people normally cultivate is a subset of the set of all plants – flowers and vegetables, mainly. *Weeds* are outliers, and must be classified here as an exploitation. This preserves the homogeneity of the set of *cultivated* plants.

A more striking example of exploitation is (18), a sentence form the *Guardian Weekly,* cited by Copestake and Briscoe (1995).

(18) [Chester] serves not just country folk, but farming, suburban, and city folk too. You'll see *Armani drifting* into the Grosvenor Hotel's exclusive (but exquisite) Arkle Restaurant and *C&A giggling* out of its streetfront brasserie next door.
Guardian Weekly, 13 November 1993 (cited by Copestake and Briscoe 1995).

The two phrases in italics are exploitations. In the first place, *Armani* and *C&A* are literally the names of clothing suppliers (designer, manufacturer, or seller), who, as such, neither drift nor giggle. Copestake and Briscoe comment: "*Armani* and *C&A* are presumably intended to be interpreted along the lines of *people wearing clothes from Armani | C&A*." We would add that these names are being exploited as typifications of expensive vs. cheap clothing; it would not negate the meaningfulness of the sentence if it were discovered that the people in question were actually wearing clothes from Gucci and Marks & Spencers.

The distinction between alternations and exploitations as outlined here is complicated by the fact that an exploitation may be picked up by other users of the language and become established as a norm in its own right. Today's exploitation may be tomorrow's secondary convention.

6. Shimmering sets

Lexical sets are not stable paradigmatic structures. Another salient characteristic of lexical sets, besides the fact that they don't map neatly onto semantic types, is that their membership has a loose semantic unity. The lexical sets populating a node in the ontology (i.e. a semantic type) tend to shimmer – that is, the membership of the lexical set changes from verb to verb: some words drop out while other come in, just as predicated by Wittgenstein (family resemblances). Different verbs select different prototypical members of a semantic type even if the rest of the set remains the same. For example, two verbs, *wash* and *amputate*, both typically select [[Body Part]] as their direct object. One can wash one's {leg | arm | foot etc.} or one can have one's {leg | arm | foot etc.} amputated. But prototypically, you wash your {face | hands | hair} but you don't have your {face | hands | hair} amputated.

(19) *wash*
Direct Object
[[Body Part]]: hand, hair, face, foot, mouth

(20) *amputate*
Direct Object
[[Body Part]]: leg, limb, arm, hand, finger

Likewise, *put on*, *wear,* and *hitch up* all apply to [[Garments]]. But while pretty well the entire set of [[Garment]]s occurs as direct object of *put on* and *wear*, in the case of *hitch up* this is not true. What do people hitch up? Typically: trousers, pants, skirt. Maybe socks and stockings, nightdress, pyjamas. Outliers are: brassiere, bosoms. But you don't hitch up your hat, shirt, shoes, boots or gloves.

(21) *wear*
Direct Object
[[Garment]]: suit, dress, hat, clothes, uniform, jeans, glove, jacket, trousers, helmet, shirt, coat, t-shirt, shoe, gown, sweater, outfit, boot, apron, scarf, tie, bra, pyjamas, stocking,

(22) *hitch up*
Direct Object
[[Garment]]: dress, skirt, stocking, bikini top, coat, pants, trousers, underpants

If we now look at verbs that describe typical actions we do with [[Document]]s (e.g. *read*, *publish*, *send*, *translate*) the following picture arises:

(23) What is a [[Document]]?

read {book, newspaper, bible, article, letter, poem, novel, text, page, passage, story, comics script, poetry, report, page, label, verse, manual}
publish {report, book, newspaper, article, pamphlet, edition, booklet, result, poem, document, leaflet, newsletter, volume, treatise, catalogue, findings, guide, novel, handbook, list}
send {message, letter, telegram, copy, postcard, cheque, parcel, fax, card, document, invoice, mail, memo, report}
translate {bible, text, instructions, abstract, treatise, book, document, extract, poem, menu, term, novel, message, letter}

Although, from a conceptual point of view, [[Document]] is a well-defined type, its linguistic membership varies according to context. This is because we don't perform exactly the same sort of operation with the objects that represent this type. For example, *translating* is a typical activity that people do with [[Document]]s such as books, poems, and novels, but there are other [[Document]]s such as newspapers that we typically don't translate (although in principle we could) but rather do other things with them. A *newspaper* is typically read, while a *message* is typically sent, a *report* is typically published, and so on (see Jezek and Lenci 2007 for a full discussion of the distributional behaviour of words denoting [[Documents]]).

Let us finally look at the typical verbal collocates of nouns denoting [[Road Vehicles]], such as *cars*, *taxis*, *ambulances* and *trains*:

(24) What do we typically do with [[Road Vehicle]]s?

car {park, drive, steal, hire, buy, sell, stop, damage, change, own, smash, wash, crash, wreck, repair, assemble, clean, hit, overtake, take, lock, leave, destroy}
taxi {hail, hire, phone, afford, order, share, stop, drive, catch, own, get, take, call, leave, find}
ambulance {escort, drive, call, phone, send, get, require, need, provide}
train {board, derail, catch, miss, drive, change, hear, leave, take, get}

Here, the situation is similar to (21): a car is typically *driven* or *parked*, a train is *caught*, *boarded* or *missed*, an ambulance is *called*. Also, we typically *take* a taxi or a train to a destination, but it would be very odd to talk of someone *taking* an ambulance to hospital. *Take* is something you do with vehicles that run a public service. Moreover, for cars, we typically predicate some activity involving causing damage (*damage*, *smash*, *crash*, *wreck*, *hit*, *destroy*), whereas the typical co-occurrence of verbs like *require*, *need*, *provide* with ambulance shows how its purpose of use is central to its meaning.

This last data on lexical sets lets us see how complex the interplay is between the conceptual system and the selectional constraints that verbs impose on their arguments. Verb selectional behaviour is closely connected with the structure of our conceptual system, but selectional constraints do not always map neatly onto semantic types. When we predicate of a class of objects, we cut that class out of our conceptual system in a way which is partially independent from its categorial status. Consider verbs such as *close*, *lift* or *throw*: the classes of objects that can be closed, lifted or thrown do not constitute a semantic type per se but become cognitively and linguistically relevant as a class by virtue of the predicative context in which they appear.

It follows that the assumption that selectional constraints can be captured in terms of semantic types only is too strong. On the other hand, the empirical analysis of the lexical sets that populate specific argument slots in relation to target verbs can improve our understanding of the phenomenon of argument selection and provide better recognition cues for sense disambiguation purposes.

For that which concerns the CPA project, this salient characteristic of lexical sets, i.e. the fact that they shimmer according to what we predicate of them, has forced us to rethink the way a linguistic ontology should be structured. In this perspective, a node in the ontology (i.e. a semantic type) is not to be thought of as an address for 'all and only' the lexical items that belong to that node. Rather, it is an address for lexical items that typically belong to that node. The ontology is thus best conceived, not as a rigid yes/no structure, but as a statistically based structure of shimmering lexical sets. Each canonical member of a lexical set is recorded with statistical contextual information, like this:

(25) [[Activity]]:

... *meeting* <attend __ 663/5355[13]; hold __ 953/24798; arrange __ 200/3581; adjourn __ 36/424, organize __ 32/2055 …>
… *conference* <attend __ 267/5355; hold __ 382/24798; organize __ 81/2055; arrange __ 29/3581 …>
… *lecture* <attend __ 75/5355; deliver __ 65/3949; give 226/75759; organize __ 5/2055; hold __ 12/24798; arrange __ 5/3581 …>

[13] The first number (633) is the total number of occurrences of *meeting* with *attend* while the second (5355) is the total number of occurrences of *attend* in our reference corpus (British National Corpus).

… *concert* <stage __ 18/1157; attend __ 29/5355; play __ 27/17832; organize __ 13/2055; hold __ 21/24798; arrange __ 6/3581 …>

(26) [[Document]]:

... *book* <read __ 772/9037; write __ 933/13015; publish __ 416/7230, borrow __ 43/1358 …>
… *novel* <write __ 182/13015; read __ 88/9037; publish __ 45/7230; set __ 19/14528 …>
… *article* <write __ 263/13015; publish __ 174/7230; read __ 156/9037; contribute __ 28/1313 …>
… *letter* <write __ 1032/13015; send __ 540/12011; receive __ 544/18381; post __ 77/451 …>

(27) [[Road Vehicle]]:

... *car* <park __ 392/836, drive __ 490/4331, hire __ 85/1212, take __ 291/106749>
… *taxi* <hail __ 22/339, hire __ 7/1212, drive __ 13/4331, catch __ 9/6681, take __ 105/106749, call __ 23/28922>
… *ambulance* <drive __ 17/4331, call __ 64/28922>
… *train* <board __ 41/443, catch __ 154/6681, drive __ 17/4331, take __ 162/106749>

We hasten to add that these statistical details are mainly for computational use, not for humans. Humans are analogical engines who prefer to make analogies on the basis of broad generalizations, rather than to be bombarded with statistical details, which most of us process only subconsciously, if at all.

The examples discussed above demonstrate how the semantic ontology is a shimmering hierarchy populated partly with a fairly stable set of central and typical lexical items and partly with words which come in and drop out according to context. The relative frequency of such words in such contexts can be measured (and compared) in corpora. A shimmering ontology of this kind preserves, albeit in a weakened form, the predictive benefits of hierarchical conceptual organization, while maintaining empirical validity of natural-language description. Also, it is a structure which is representative of both typing and collocational information.

7. Concluding remarks and future work

Words with a common basic meaning are grouped together in ontologies according to their semantic type. Corpus-driven pattern analysis groups words together in lexical sets according to their syntagmatic behaviour. Syntagmatic lexical sets are not the same as sets of synonyms and hyponyms in traditional conceptual ontologies, but there is enough overlap for the relationship to be interesting and worth exploring. A context-dependent ontology like the one currently under development in the CPA project aims to represent the interactions between ontological information and information coming from the analysis of the syntagmatic dimension. Also, it measures the typicality of a given word as a member of a particular semantic type, and allows us to characterize it as a canonical member of the type or an outlier. No empirically well-founded ontology exists that groups words together into paradigmatic sets according to their syntagmatic behaviour (as opposed to their place in a conceptual

hierarchy). A linguistic ontology, if it is to qualify as truly 'linguistic', should account for combinatorial constraints on lexical items as well as their place in a conceptual hierarchy.

Acknowledgement

This work was funded in part by the Czech Ministry of Education (MSM 0021620838) and the Czech Science Foundation (P406/2010/0875) as part of a corpus-driven investigation of lexical patterns at the Institute of Formal and Applied Linguistics of the Charles University in Prague.

Bibliography

COPESTAKE Ann and **BRISCOE** Ted, "Semi-productive Polysemy and Sense Extension", *Journal of Semantics* Vol. 12 (1), 1995: 15-67.

HANKS Patrick, "Linguistic norms and pragmatic explanations, or why lexicographers need prototype theory and vice versa", *in* KIEFER F., KISS G., and PAJZS J. (eds.), *Papers in Computational Lexicography: Complex '94*. Research Institute for Linguistics, Hungarian Academy of Sciences, 1994.

HANKS Patrick, Corpus Pattern Analysis. CPA Project Page, 2009. Retrieved May 15, 2009, from http://nlp.fi.muni.cz/projects/cpa/.

HANKS Patrick Forthcoming, *Lexical Analysis: Norms and Exploitations.* Cambridge, MA: The MIT Press.

HANKS Patrick and **PUSTEJOVSKY** James, "A Pattern Dictionary for Natural Language Processing", *Revue Française de linguistique appliquée* Vol. 10 (2), 2005: 63-82.

HANKS Patrick, **PALA** Karel and **RYCHLY** Pavel, "Towards an empirically well-founded ontology for NLP", *Proceedings of the 4th International Workshop on Generative Approaches to the Lexicon,* Paris, May 10-11 2007.

JEZEK Elisabetta and **LENCI** Alessandro, "When GL meets the corpus: a data-driven investigation of semantic types and coercion phenomena", *Proceedings of GL 2007, Fourth International Workshop on Generative Approaches to the Lexicon.* Paris, May 10-11, 2007.

JURAFSKY Daniel and **MARTIN** James H., *Speech and Language Processing*: *An introduction to Natural Language Processing, computational linguistics and speech recognition*, New Jersey, Prentice Hall, 2000.

KILGARRIFF Adam, **RYCHLÝ** Pavel, **SMRŽ** Pavel and **TUGWELL** David, "The Sketch Engine", *Euralex Proceedings*, Lorient, France, 2004.

MANNING Christopher D. and **SCHÜTZE** Hinrich, *Foundations of Statistical Natural Language Processing*, Cambridge MA, The MIT Press, 1999.

MILLER George A. and **FELLBAUM** Christiane, "WordNet then and now", *Language Resources & Evaluation* 41, 2007: 209–214.

PUSTEJOVSKY James, *The Generative Lexicon*, Cambridge MA, The MIT Press, 1995.

PUSTEJOVSKY James, "Type Theory and Lexical Decomposition", *Journal of Cognitive Science* 6, 2006: 39-76.

PUSTEJOVSKY James, **HAVASI** Catherine, **LITTMAN** Jessica, **RUMSHISKY** Anna and **VERHAGEN** Marc, "Towards a Generative Lexical Resource: The Brandeis Semantic Ontology", *Proceedings of the International Conference on Language Resources and Evaluation* (*LREC 2006*), Genoa, Italy, 2006: 1702-1705.

PUSTEJOVSKY James, **HANKS** Patrick and **RUMSHISKY** Anna, "Automated Induction of Sense in Context", *Proceedings of the 20th International Conference on Computational Linguistics* (COLING-04) Geneva, Switzerland, 2004: 924-931.

PUSTEJOVSKY James and **JEZEK** Elisabetta, "Semantic Coercion in Language: Beyond Distributional Analysis", *Distributional Models of the Lexicon in Linguistics and Cognitive Science*, special issue of *Italian Journal of Linguistics*, 20.1, 2008: 175-208.

RUMSHISKY Anna, **GRINBERG** Victor and **PUSTEJOVSKY** James, "Detecting selectional behaviour of complex types in text", *Proceedings of the 4th International Workshop on Generative Approaches to the Lexicon*, Paris, May 10-11, 2007.

RUMSHISKY Anna, **HANKS** Patrick, **HAVASI** Catherine and **PUSTEJOVSKY** James, "Constructing a Corpus-based Ontology using Model Bias", *Proceedings of the 19th International FLAIRS Conference*, Melbourne Beach, Florida, 2006.