Shimmering Lexical Sets

Patrick Hanks
Masaryk University

Elisabetta Ježek
University of Pavia

For natural language processing and other applications, it has long seemed desirable to group words together according to their essential semantic type-[[Human]], [[Animate]], [[Artefact]], [[Physical Object]], [[Event]], etc.-and to arrange them into a hierarchy. Vast lexical and conceptual ontologies such as WordNet and BSO have been built on this foundation. Examples such as fire a [[Human]] (=dismiss from employment vs. fire a [[Weapon]] (=cause to discharge a projectile) have led to the expectation that semantic types such as [[Weapon]] and [[Human]] can be used systematically for word sense disambiguation. Unfortunately, this expectation is often unwarranted. For example, one attends an [[Event]]-a meeting, a lecture, a funeral, a coronation, etc., but there are many events-e.g. a thunderstorm, a suicide-that people do not attend, while some of the things that people do attend—e.g. a school, a church, a clinic-are not [[Event]]s, but rather [[Location]]s where specific events take place. The sense of attend is much the same in all these examples, unaffected by differences in the semantic type of the direct object. Nevertheless, the pattern [[Human]] attend [[Event]] is well established and intuitively canonical.

The CPA (Corpus Pattern Analysis) project at Masaryk University, Brno, provides two steps for dealing with this kind of inconvenient linguistic phenomenon:

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 [[Event]] set in the context of attend, but not, for example, in the context of arrange.

The ontology is not a rigid yes/no structure, but a statistically based structure of shimmering lexical sets.

Thus, each canonical member of a lexical set is recorded with statistical contextual information, like this: [[Event]]: ... meeting. Thus, the semantic ontology is a shimmering hierarchy populated with words which come in and drop out according to context, and whose relative frequency in those contexts is measured. A shimmering ontology of this kind preserves, albeit in a weakened form, the predictive benefits of hierarchical conceptual organization, while maintaining the empirical validity of natural-language description.

1. Introduction

Corpus linguistics has shown, if nothing else, that human linguistic behaviour—and in particular word use—is very highly patterned, although the boundaries of each pattern of use tend to be fuzzy and variable (a fact that has led to much confusion in speculative linguistic research based on invented examples: bizarre examples are conducive to bizarre theories). 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 usage. 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 and use of words in a language.

The primary goal of the CPA (Corpus Pattern Analysis) project at Masaryk University, Brno (*http://nlp.fi.muni.cz/projekty/cpa/*) is to provide an inventory of all the normal patterns of use of all the normal verbs in English. Meanings of verbs in CPA are associated, not with verbs in

isolation, but with verbs in relation to nouns and other clues: the nouns that co-occur with each verb in different clause roles: subjects, direct objects, and prepositional objects. 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 reflect the structure of linguistic meaning and predict the meanings of verbs according to their groups or semantic frames. In this paper we examine some of the problems the are obstacles to the fulfilment of such a goal.

2. Ontologies and lexical sets

Ontology is a vogue word in modern computational linguistic and lexical research. It is used with several different meanings, some of which are of quite recent origin. We start, therefore, by summarizing these different meanings in three main groups and 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 a list of all known concepts, and hence all the words and expressions in a particular language that denote concepts, both abstract and concrete. This is the sense in which the term *ontology* is applied to WordNet and other, similar ordered collections of the contentful terms of a language. 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.

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 *all* 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. Polysemy and disambiguation

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 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 to be effectively replaced by a paradigmatic cluster of lexical items for identifying the "celebrate" sense of the verb *toast*.

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.

In fact, very often semantic distinctions among the different senses of a verb depend crucially on the semantic type of the direct object, e.g. "fire [Human]" (=dismiss from employment) vs. "fire [Weapon]" (=cause to discharge a projectile), so we need these types for lexical analysis. Straightforward examples such as this have led to the expectation that semantic types such as [Weapon] and [Human] can be used systematically for word sense disambiguation and that

typing information on selectional preferences can guide the induction of senses for both verbs and nouns in texts (Manning and Schütze 1999: 288).

Unfortunately, this expectation is often disappointed. Consider the verb *attend*.

Typically, a person "attends" an [Event] (meeting, lecture, funeral, coronation, etc.). However, there are many events (e.g. thunderstorm, suicide) that people do not "attend", while some of the things that people do attend (e.g. school, church, clinic) are not [Event]s, but rather [Location]s.[1]

(1) *attend*
Direct Object:
a. [Event]: 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".
"He no longer attends the church".

The sense of *attend* is much the same in all these examples. It involves the human subject being physically present at an event in a given location. The sense of the verb does not differ with the differences in semantic type of the direct object. The pattern "[Human] attend [Event]" is well established as a core use of this verb. It contrasts with another, rarer sense of *attend,* meaning "be accompanied by", as in 2.

(2) ...the public shock that attended the *corporate failures* at the end of the 1980s.

Here, the meaning is "accompany" or "be a result of". In this second sense, the subject of the verb is generally an abstract noun, not a human being. The direct object, however, is still an [Event]—a "corporate failure" is indisputably an [Event]—but a different kind of event from a meeting, wedding, or funeral. Corporate failure is not the sort of event that people go to a particular location to participate in.

Thus, the large set of lexical items with semantic type [Event] consists of at least two subsets in relation to the verb *attend*—those that are ceremonies in particular locations and those that are processes accompanying other events, processes, or states of affairs. A natural reaction to this finding would be to subcategorize the nouns of semantic type [Event] into subtypes, and this is indeed what ontologies tend to do. But this turns out to be a false step. It would be worth taking only if it can be shown that each of the two groups of lexical items occurs in several different contexts. There is no evidence that this is the case. Rather, it seems that a large number of different subsets of nouns are event-specific, i.e. they occur only in particular contexts. In other words, *every verb picks out its own set of typical direct objects from within a broad semantic class*. To show how this works, let us look at two of the most typical "event type" direct objects of *attend—meeting* and *wedding*—and see some of the verbs with which they typically do or do not co-occur.

VERB	Meeting	Wedding
attend	YES	YES
arrange	YES	YES
plan	YES	YES
organize	YES	YES
convene	YES	NO
summon	YES	NO
call	YES	NO

[1] The data below is taken from the BNC corpus and presented adopting a layout proposed in Rumshisky et al. 2007.

chair	YES	NO
celebrate	NO	YES
perform	NO	YES

Table 1. Typical verbal collocates of *meeting* and *wedding*

All of the verbs in Table 1 take events as direct objects, but they are different, overlapping subsets of events. The different subsets do not have enough generality to be classed as semantic subcategorizations. Instead, they should be seen as paradigmatic clusters within a single overall semantic type.

The CPA (Corpus Pattern Analysis) project provides two steps for dealing with such mismatches between sets and types and overlapping subsets:

1. Non-canonical lexical items are coerced into "honorary" membership of a semantic type in particular contexts.
2. The ontology is not a rigid yes/no structure, but a statistically based structure of collocational preferences, which we call "shimmering lexical sets".

In the following sections we explain how these two steps are implemented in the project.

4. Lexical sets and coercion

Semantic coercion can be generally described as a shift of the basic meaning of a word due to semantic requirements imposed by other words in a given context. Coercion is a principled mechanism for accounting for the variety of senses that words exhibit in different contexts. In particular, coercion applies when the sense that a word exhibits in context is not inherent to the word itself but rather the result of compositional processes. In a broad definition, coercion covers both conventionalized sense modulations, e.g. metonymic shifts of the kind animal/food, container/containee, mass/countable, etc. (as discussed in Pustejovsky 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 investigated within the Generative Lexicon (GL) framework (Pustejovsky 1995 2006). Type coercion is an operation of type adjustment that occurs when 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, frequently quoted in the literature, is *event type coercion*. This occurs for instance when a verb that normally selects an [Event] as direct object (e.g. *finish* drinking *something*) is used in combination with an artefactual entity (e.g. *finish* one's beer). In this case, the verb induces a re-interpretation of the noun, so that it can successfully fill the object argument slot[2]:

(3) *finish*
Direct Object:
a. [Event]: journey, tour, treatment, survey, race, game, training, ironing
b. [Physical Object]: penicillin, sandwich, cigarette, cake, dessert, food
c. [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 event activated by the operation of coercion is not a general event, but a specific event that is conventionally associated with the object: *finish the wine*

[2] For a collection of examples of coercion extracted from corpora, including some of those reported here, see Pustejovsky and Ježek (2008).

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[3]. Within the GL model, it is assumed that these events are coded as qualia relations in the noun's semantics (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 to this type:

(4) *ring*
Direct Object
[Human]: mother, doctor, Gill, Chris, friend, neighbour, director
[Institution]: police, agency, club
[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 (see the implausibility of *"ring the garden").

Lastly, an example of coercion to [Container] is provided by the verb *open* in its "make accessible" sense:

(5) *open*
Direct Object
[Container]: drawer, bottle, cupboard, envelope, folder, tin, can, box, fridge, bag, cage, suitcase
[Beverage]: wine, champagne, beer
"I opened the wine carefully".
"Just as he was about to open the beer, the doorbell rang".

Examples (3-5) show that the kind of coercion differs depending on how the missing piece of semantic information is retrieved: while the [(drinking) Event] 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[4].

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

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

b) 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 [Event] 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 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 option 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

[3] There are some difficulties with this formulation: however, we won't discuss them here.

[4] For a more detailed and formal account, see Pustejovsky (2006).

objects. Unlike *ring*, these two uses of *visit* seem to map quite neatly onto two distinct patterns (e.g. "go to and spend some time in a place for tourism, business or other purpose" and "go to and spend some time with a person, typically for social reasons"):

(6) *visit*
[Human] visit [Location]
He visited Paris in 1912
[Human 1] visit [Human 2]
She visited her father regularly

In other cases, however, 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 event 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", 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 connections that exist between them. In the case of *attend,* for example, *schools*, *clinic*s, and *churches* are [Location]s where particular [Event]s take place (a class, a treatment, a mass and so on). It is to these [Events]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 [Event] 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. 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. Two common alternations are for instance [Human | Institution] and [Human | Body Part]:

(7) *think* [Human | Institution] think...
Alice thinks that | the *government* thinks that
negotiate [Human | Institution] negotiate...
The *chairman* negotiated a new deal | the *company* negotiated a new deal
comment [Human | Institution] comment
The *president* wouldn't comment | The *White House* wouldn't comment

(8) *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 (7) and (8), or in object position, as in (9):

(9) *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".

Notice that the occurrence of two alternating types—e.g.[Human] and [Body Part]—in the direct object position with a given verb is not necessarily always an alternation. Instead, the different types may each pick out a distinct verb sense. For instance, depending on the type of noun filling the subject and object argument slots, *irritate* can mean either "cause to feel angry and annoyed" or "cause inflammation or discomfort to":

(10) *irritate* [Anything] irritate [Human]]
Television news irritates me
[Stuff] irritate [Body Part]
Ozone irritates the eyes

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 *sentence* is found in the following pattern:

(11) [Human1 = Judge] sentence [Human 2 = Convict] {to [Punishment]}
"When the Duke sentences Lucio..."

There is nothing in the intrinsic semantics of *duke* to say that he is acting as a judge, nor in the semantics of *Lucio* to say that he has been convicted of a crime. These are roles assigned by context—specifically, by the selectional preferences of the verb *sentence*, which expects a subject and object with such characteristics.

A similar case is that of *arrest*. In (12) the antecedents of *he*, *the Germans,* and *we* respectively are assigned the role *Police Officer* in context.

(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 in 1942"
"indeed yesterday we unfortunately failed to arrest some very prominent IRA men"

Sometimes, both semantic role and alternation apply in a single pattern:

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

In (13) the role of [Driver] is assigned contextually to the type [Human] by the verb *accelerate*.

(14) [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"

In (14) the pattern alternates between a [Plane], the [Human] who drives it (contextual role: Pilot), and the [Humans] who are transported by it (contextual role: Passengers).

c) An exploitation is a deliberately creative or unusual use of an established pattern of word use—a norm. 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 (15).

(15) 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 (16), a sentence form the *Guardian Weekly,* cited by Copestake and Briscoe (1995).

(16) [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 boldface 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.

5. Shimmering lexical sets

Lexical sets are not stable paradigmatic structures. Another salient characteristic of lexical sets, besides the fact that they cut across semantic types, is that their membership has a loose semantic unity. The lexical set populating a node in the ontology (e.g. 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.

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

(18) 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.

(19) *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,

(20) hitch up
Direct Object
[Garment]: dress, skirt, stoking, 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:

(21) 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 fuller discussion).

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

(22) 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. 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 the purpose is central to this noun's meaning.

This salient characteristic of lexical sets, i.e. the fact that they shimmer according to what we predicate of them, forces 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:

(23) [Event]:

... *meeting* <attend __ 663/5355[5]; 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 …>

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

(24) [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 …>

(25) [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.

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

The examples discussed above demonstrate how the semantic ontology is a shimmering hierarchy populated with words which come in and drop out according to context, and whose relative frequency in those 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.

6. Concluding remarks and future work

Words with a common central 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 and collocational information. 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 supported in part by grant T100300419 of the Academy of Sciences of the Czech Republic and grant 2C06009 of the Czech Ministry of Education (National Research Program II).

References

Copestake, A.; Briscoe, T. (1995). "Semi-productive Polysemy and Sense Extension". *Journal of Semantics* 12 (1). 15-67.

Hanks, P. (2006). "The organization of the Lexicon: Semantic Types and Lexical Sets". In *Euralex Proceedings 2006*. 1165-1168.

Hanks, P. (forthcoming). *Lexical Analysis: Norms and Exploitations.* Cambridge: MIT Press.

Hanks, P.; Pustejovsky J. (2005). "A Pattern Dictionary for Natural Language Processing". *Revue française de linguistique appliquée* 10 (2). 63-82.

Hanks, P.; Pala, K.; Rychlý, P. (2007). "Towards an empirically well-founded ontology for NLP". In *Proceedings of GL 2007, Fourth International Workshop on Generative Approaches to the Lexicon.* Paris, May 10-11, 2007.

Ježek, E.; Lenci, A. (2007). "When GL meets the corpus: a data-driven investigation of semantic types and coercion phenomena". In *Proceedings of GL 2007, Fourth International Workshop on Generative Approaches to the Lexicon*. Paris, May 10-11, 2007.

Kilgarriff, A. et al. (2004). "The Sketch Engine". In *Proceeding of Euralex 2004*. Lorient, France.

Manning, C. D.; Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

Miller, G. A.; Fellbaum, C. (2007). "WordNet then and now". *Language Resources and Evaluation* 41 (2). 209-214.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge: MIT Press.

Pustejovsky, J. (2006). "Type Theory and Lexical Decomposition". *Journal of Cognitive Science* 6. 39-76.

Pustejovsky, J. et al. (2006). "Towards a Generative Lexical Resource: The Brandeis Semantic Ontology". In *Proceedings of LREC 2006*. Genoa, Italy.

Pustejovsky, J.; Hanks, P.; Rumshisky, A. (2004). "Automated Induction of Sense in Context". In *Proceedings of COLING 2004*. Geneva, Switzerland. 924-931.

Pustejovsky, J.; Ježek, E. (2008). "Semantic Coercion in Language: Beyond Distributional Analysis". In Lenci, A. (ed.) *Distributional Models of the Lexicon in Linguistics and Cognitive Science*, special issue of *Italian Journal of Linguistics* (to appear).

Rumshisky, A.; Grinberg, V.; Pustejovsky J. (2007). "Detecting selectional behaviour of complex types in text". In *Proceedings of GL 2007, Fourth International Workshop on Generative Approaches to the Lexicon.* Paris, May 10-11 2007.