An empirical classification of verbs based on Semantic Types: the case of the 'poison' verbs

Jane Bradbury
RIILP
University of Wolverhampton
J.Bradbury3@wlv.ac.uk

Ismail El Maarouf
RIILP
University of Wolverhampton
i.el-maarouf@wlv.ac.uk

Abstract

This article proposes a new approach to verb classification based on Semantic Types selected in corpus-based verb patterns. This work draws on Hanks's theory of Norms and Exploitations (Hanks 2013) and applies Corpus Pattern Analysis to a subset of verbs from Levin's 'poison' class, including verbs such as *hang* and *stab*. These patterns are taken from the Pattern Dictionary of English Verbs, which aims at recording prototypical phraseological patterns for the most frequent verbs of English using the British National Corpus.

1 Introduction

This article proposes a new approach to verb classification based on Semantic Types (STs) of the verbs' arguments. This work draws on Hanks's theory of Norms and Exploitations (Hanks 2013) and applies Corpus Pattern Analysis to a subset of verbs from Levin's 'poison' class (class 42.2) [*asphyxiate, crucify, drown, hang, knife, poison, smother, stab, strangle, suffocate:* those verbs available in PDEV at the time of writing], (Levin 1993: 232-233). The patterns are taken from the Pattern Dictionary of English Verbs (PDEV[1]), which records prototypical patterns of use for English verbs in the British National Corpus[2].

This paper focuses on the patterns that relate to the 'literal' (i.e. killing-related) senses of the 'poison' verbs. According to Levin, they 'lexicalize a means component and it is this means that differentiates amongst them'. For example, the verb *poison* entails the notion that an attempt to kill is being made by means of a poisonous substance, whereas the verb *knife* entails the notion of a knife as a means of killing.

Levin argues that the meaning of the verb determines to a large extent its syntactic behaviour. She therefore undertakes the description of English verb classes that share both syntactic alternations and similar meaning. This claim bears comparison with empirical work in Corpus Linguistics, such as Sinclair's account of *yield* (Sinclair 1990: 53-65), where convincing evidence that sense and syntax are closely associated was found. Similar claims are made in Natural Language Processing, especially in distributional models of meaning (Grefenstette 1994; Bieman & Giesbrecht 2011) used in a large number of applications (Cohen & Widdows 2009).

Levin's verb classes have been integrated and extended into a lexical resource for Natural Language Processing (NLP), VerbNet[3] (Kipper et al. 2008) used in applications such as Semantic Role Labelling (Swier et al. 2004). The present paper proposes to create a semantic network from PDEV, by building strings of STs and linking them to verbs. One of the motivations behind this work is that PDEV contains useful information which is absent in NLP resources: while VerbNet analyses the interface between thematic roles (e.g. Agent, Patient) and selectional restrictions (e.g. [+ANIMATE], [+CONCRETE]), PDEV maps clause roles (e.g. Subject, Object) using STs (e.g. [[Human]], [[Location]]).

This paper describes the background and methodology for this work (section 2) and provides a detailed analysis of Levin's claims about the 'poison' verb class (section 3), before describing results obtained from using the semantic network (section 4).

[1] freely available at http://deb.fi.muni.cz/pdev/
[2] available at www.natcorp.ox.ac.uk/
[3] see http://verbs.colorado.edu/verb-index/index.php

2 Methodology

2.1 Background

Corpus Pattern Analysis (CPA) is a new technique for mapping meaning onto words in text (Hanks 2012). The focus of CPA is on analysing large corpora to identify the prototypical syntagmatic and collocational patterns with which words are associated. It has simultaneously given rise to a new theory of language in use, the Theory of Norms and Exploitations (TNE, see Hanks 2013), which can be compared with Pattern Grammar (Hunston and Francis 2000) and Construction Grammar (Goldberg 1995).

PDEV (in progress) aims to provide a well-founded corpus-driven account of verb meaning, using STs to stand as prototypes for collocational clusters occurring in each clause role. Current CPA practice has shown that the scientific concepts from WordNet[4], the most widely used semantic repository, do not map well onto words as they are actually used; this is partly because folk concepts, and not scientific concepts, form the foundation of meaning in natural language (Wierzbicka 1984). For this reason, a new shallow Ontology consisting of 225 STs has been developed for PDEV which contrasts with WordNet in the following key respects:

- WordNet contains many scientific concepts, whereas the PDEV Ontology is modeled on folk concepts, for example, WordNet has over 50 hyponyms for *Animate Being*, whereas PDEV has only 17 STs listed under [[Animate]];
- WordNet is intuition-based whereas the PDEV Ontology is 'corpus-driven' and built from the words upwards.

For each verb in PDEV, a sample of ~250 lines is analysed and phraseological norms, or patterns, identified and then recorded using STs. For example, in the account of pattern 1 of *strangle*, below, the STs [[Human 1]] and [[Human 2]] are used to indicate that, prototypically, it is a human who performs the action of strangling and they typically perform this act upon another human.

(1) [[Human 1]] strangle [[Human 2]]

Where relevant, information about adverbial phrases is recorded as part of the pattern. For example, pattern 1 of *drown*, below, records that this use of *drown* frequently selects an adverbial phrase indicating in which [[Watercourse]] or what type of [[Liquid]] the drowning occurred.

(2) [[Human | Animal]] drown [NO OBJ] (in [[Watercourse]] | in [[Liquid]])

These STs form the basis of the semantic network, which generates semantic strings from patterns and link them to verbs.

2.2 A semantic network for verbs

PDEV allows for a new kind of verb classification, by clustering verbs according to the STs with which they combine across patterns. To explore this method further, PDEV patterns have been simplified to semantic strings, i.e. combinations of types in various pattern positions. More specifically, semantic strings are the result of the following two changes to the PDEV patterns:

- only STs and lexical sets of subjects, objects, adverbials, adverbial functions, and prepositions are kept and concatenated;
- since patterns allow for several alternative STs in the same clause role and for these clause roles to be optional, all combinations are generated.

Based on the analysis of more than 3500 patterns available in the PDEV, the current version of the network totals over 5064 different semantic strings, with 955 of them linking more than one verb (covering over 71% of patterns). This allows the identification of both the different strings a verb combines with and the verbs clustered around each semantic string. For example, the semantic string '[[Human]] verb [[Human]]', accounting for the transitive use of verbs such as *corner* and *sacrifice,* is the largest cluster of the network, with 188 verbs. Lastly, the network offers the possibility of computing the similarity between verbs, using their shared strings, and applying standard distributional methods.

3 Levin's hypotheses

3.1 Instrumental Phrases

Levin hypothesizes that few of the 'poison' verbs 'will select instrumental phrases (IPs), but that where this is the case, the instrumental phrase is a "cognate"'. Table 1 lists the proportion of tokens combining with IPs for each PDEV pattern, focusing on the patterns which relate to the 'literal' (ie killing-related) senses and with a non-instrumental subject, i.e. [[Human]], [[Institution]] or [[Animate]].

[4] see http://wordnet.princeton.edu/

It is notable that only *crucify* and *knife* – verbs where the means is lexicalized unambiguously – generated zero returns. Elsewhere, contrary to Levin's hypothesis, the selection of instrumental phrases is not infrequent. This could be explained by the broadness of the set of instruments lexicalized by verbs such as *stab* or *hang* (see below). It is interesting to compare the three patterns of *hang* illustrated below: '[[Human 1 | Institution]] hang [[Human 2]]' rarely selects an instrumental phrase, presumably because in this context, where the event described is a formally decreed execution, the instrument used (i.e. *rope*, *gallows*) is unambiguous. However, '[[Human]] hang [NO OBJ] ([Adv[Location]])' and '[[Human]] hang [Self]' both describe 'unofficial' acts where the instrument used cannot be taken for granted, and in both these patterns the verb selects an IP with relatively high frequency.

Levin's hypothesis, that where an instrumental phrase is used it will be a 'cognate', holds, if 'cognate' is taken to mean 'an object with similar physical properties to the object prototypically used to commit the act in question'.

The instrumental phrases for *stab* include cognates such as *carving knife*, *sheath knife*, and *butcher's knife*, along with the less conventional *screwdriver* and *pencil*. For the previously-mentioned 'unofficial' senses of *hang*, the instrumental phrases include, *rope, string, a belt,* and *blanket torn into strips*. The broad and open-ended nature of these lexical sets (i.e. 'anything sharp and pointed', or 'anything long, thin, flexible and rope-like') suggests that where the means lexicalized by a verb is ambiguous, it is not unusual for an instrumental phrase to be selected.

3.2 STs as subjects

Levin hypothesizes that few of the 'poison' verbs 'allow instrumental subjects'.

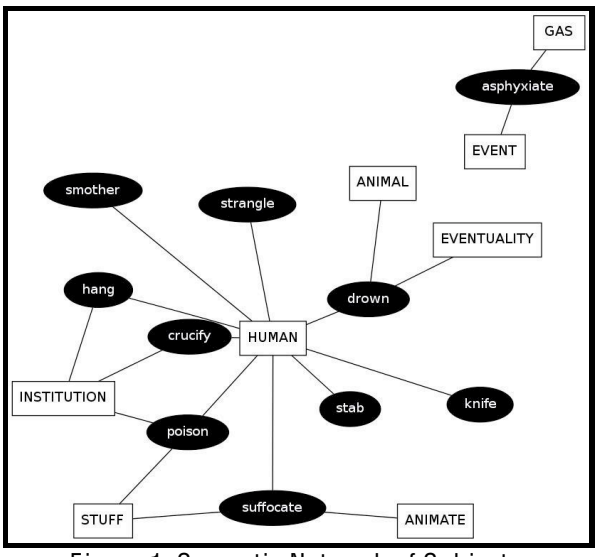

Figure 1. Semantic Network of Subjects

Verb pattern	no of lines with IPs
[[Human 1]] **crucify** [[Human 2]]	0/40
[[Human | Animal]] **drown** [NO OBJ] (in [[Watercourse]] | in [[Liquid]])	25/76*
[[Human]] **drown** [Self]	2/6*
[[Human 1 | Eventuality]] **drown** [[Human 2 | Animal]]	10/51*
[[Human 1 | Institution]] **hang** [[Human 2]]	3/45
[[Human]] **hang** [NO OBJ] ([Adv[Location]])	20/39
[[Human]] **hang** [Self]	12/55
[[Human 1]] **knife** [[Human 2]] (in [[Body Part]] | to {death})	0/22
[[Human 1 | Institution]] **poison** [[Human 2 | Animate]]	5/69
[[Human | Institution | {Stuff = Toxic}]] **poison** [[Location | Watercourse]]	9/31
[[Human 1]] **smother** [[Human 2]]	1/6
[[Human 1]] **stab** [[Human 2]] (in [[Body Part]] | through [[Body Part]]) (to {death})	30/198
[[Human 1]] **strangle** [[Human 2]]	9/84
[[Human | Animate]] **suffocate** [NO OBJ]	10/30**
[[Stuff | Human 1 | Animate 1]] suffocate [[Human 2 | Animate 2]]	6/23**

*This includes references to location where it is [[Watercourse]] or [[Liquid]].
**This includes references to events which caused the suffocation.

Table 1. Instrumental phrases

In order to investigate this aspect further, STs in Subject position have been extracted from the semantic network (Figure 1). As can be seen, both Instrument and Agent roles occur as subjects: the subjects are [[Gas]], [[Stuff]], [[Event]] and [[Eventuality]]. In conformity with Levin's claim, [[Stuff]] accounts for relatively few subjects for *suffocate* (25%) and *poison* (12%). In contrast, in our sample *asphyxiate* was only observed to select [[Gas]] and [[Event]] as subjects. The network also reveals subjects that are neither Agents nor Instruments, represented by the ST [[Eventuality]] (example 3).

(3) A rock-fall into Shimbara Bay caused three surges which *drowned* 15,000 people.

Figure 1 also shows how verbs can be grouped according to STs: all verbs except *asphyxiate* select [[Human]] in subject position. In addition, *poison*, *hang* and *crucify* all select [[Human]] and [[Institution]] as subjects.

4 ST-based classification

4.1 Literal senses of 'poison' verbs

The semantic network records a total of 161 semantic strings for the 'poison' verbs, only 28 of which are related to 'killing', and 9 strings cluster two or more verbs. The largest cluster is around the string '[[Human]] verb [[Human]]', which is selected by all verbs with the exception of *asphyxiate* (see 3.2). No strings provide evidence that *asphyxiate* belongs to the 'poison' class, as opposed to e.g. *poison* and *suffocate*, which share three strings.

The network includes strings of '[[Human]] verb [NO OBJ]' for *hang, drown,* and *suffocate*, which are inchoative alternations of the transitive/causative pattern. Levin has these alternations as a separate class ['suffocate' verbs] which only includes *asphyxiate, choke, drown, stifle* and *suffocate*. Levin does not list *hang* as having an inchoative use in the 'killing' sense, but evidence is found in the corpus (40 examples out of 500) as in *He was sentenced to hang*.

Reflexive object uses such as in '[[Human]] verb [SELF]' for *drown* and *hang* have not been identified by Levin (see Obligatorily Reflexive Objects class), but must be accounted for.

Strings that include adverbials are relevant to *knife* and *stab*, which share '[[Human]] verb [[Human]] {to death}' (resultative), and '[[Human]] verb [[Human]] {in [[Body Part]]}'. These adverbial phrases serve to clarify some of the semantic ambiguity that these verbs entail, i.e. whether or not the action resulted in death, and the body part affected; verbs such as *asphyxiate* and *strangle* entail no such ambiguity and are not observed to select these adverbial patterns.

4.2 Extended meanings of 'poison' verbs

The semantic network identifies similarities beyond those previously discussed where the focus has been on strings entailing the notion of 'killing'. Semantic strings extend to non-[[Human]] patients as exemplified by '[[Human]] verb [[Physical Object]]' (*smother*, *hang*). Here, the verb does not entail 'killing', e.g. *hanging* a lamp or *smothering* burning clothes with blankets.

Moreover, some strings entail metaphorical meanings. For example, *drown* and *smother* share the '[[Sound]] verb [[Sound]]' string; both patterns can be interpreted literally as one [[Sound]] being so loud that another [[Sound]] cannot be heard. *Strangle* and *suffocate* share the string '[[Anything]] verb [[Eventuality]]'; both conveying the notion that an [[Eventuality]] can be hindered or brought to an undesired end by [[Anything]]. The network thus helps to unveil the fact that the similarities between verbs can hold on several dimensions of meaning: whilst the 'poison' verbs also select strings which express a means of killing, some of them share other strings which are not covered in Levin's book.

5 Conclusion and perspectives

This paper has proposed a new approach to verb classification based on strings of STs (selected from a well-founded ontology) extracted from a semantic network based on PDEV patterns. Focusing on the 'poison' verbs has enabled the identification of key differences between this resource and Levin's account. The paper has studied the hypothesis that the 'poison' verbs lexicalize a means, through claims made on syntactic and semantic constraints on prepositional phrases and subjects. The analysis has revealed that this class must be revised in light of corpus evidence, and that sub-groupings can be made.

This work will be extended to:

- systematically explore the network with NLP techniques (e.g. distributional methods) to rank the similarity between verbs;
- investigate degrees of ambiguity in lexicalization focusing on instrumental phrases and instrumental subjects;
- explore metaphorical class extension.

Acknowledgements

We would like to thank Patrick Hanks and anonymous reviewers for their comments on an earlier draft. This work was supported by AHRC grant [DVC, AH/J005940/1, 2012-2015].

References

Chris Biemann and Eugenie Giesbrecht. 2011. *Proceedings of the Workshop on Distributional Semantics and Compositionality*. Portland: ACL. http://www.aclweb.org/anthology/W11-13

Lou Burnard. 1995. *Users' reference guide to the British National Corpus.* Oxford: Oxford University Computing Services.

Trevor Cohen, Dominic Widdows. 2009. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, Volume 42 (2). pp 390-405.

Christiane Fellbaum (1998). *WordNet: An Electronic Lexical Database*. Cambridge MA: MIT Press.

Gill Francis and Susan Hunston. 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English (Studies in Corpus Linguistics).* Amsterdam: John Benjamins.

Adele Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure (Cognitive Theory of Language and Culture Series)*. Chicago: University of Chicago Press.

Gregory Grefenstette. 1994. *Explorations in automatic thesaurus discovery*. Boston: KAP.

Patrick Hanks. 2012. How people use words to makes meanings: Semantic Types meet Valencies. In Alex Boulton and James Thomas (eds.) *Input, Process and Product: Developments in Teaching and Language Corpora.* Masaryk, Czech Republic: Masaryk University Press, pp 54-69.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations.* MIT Press, Cambridge MA.

Karin Kipper, Anna Korhonen, Neville Ryant and Martha Palmer. 2008. A Large-Scale Classification of English Verbs. In *Journal of Language Resources and Evaluation*. 42(1), pp 21-40.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago CHI : University of Chicago Press.

Robert S. Swier and Suzanne Stevenson. 2004. Unsupervised Semantic Role Labelling. In Proceedings of *EMNLP 2004*. pp. 95-102.

Anna Wierzbicka. 1984. "Apples" are not a "Kind of Fruit": The Semantics of Human Categorization. *American Ethnologist*, Vol 11(2). pp 313-328.