Can Corpus Pattern Analysis Be Used in NLP?

Silvie Cinková[1], Martin Holub[1], Pavel Rychlý[2],
Lenka Smejkalová[1], and Jana Šindlerová[1]

[1] Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
[2] Masaryk University in Brno, Faculty of Informatics, Department of Information Technology

Abstract. Corpus Pattern Analysis (CPA) [1], coined and implemented by Hanks as the Pattern Dictionary of English Verbs (PDEV) [2], appears to be the only deliberate and consistent implementation of Sinclair's concept of Lexical Item [3]. In his theoretical inquiries [4] Hanks hypothesizes that the pattern repository produced by CPA can also support the word sense disambiguation task. Although more than 670 verb entries have already been compiled in PDEV, no systematic evaluation of this ambitious project has been reported yet.

Assuming that the Sinclairian concept of the Lexical Item is correct, we started to closely examine PDEV with its possible NLP application in mind. Our experiments presented in this paper have been performed on a pilot sample of English verbs to provide a first reliable view on whether humans can agree in assigning PDEV patterns to verbs in a corpus. As a conclusion we suggest procedures for future development of PDEV.

1 Corpus Pattern Analysis

1.1 What Is a Lexical Item?

John Sinclair, the Nestor of corpus linguistics, criticized the separation of grammar and lexicon in the sense that the grammar (in extreme cases) only describes the *form* of a lexical item with respect to its potential context, while the lexicon primarily describes the *meaning* comprised by its base form, regardless of the context. Not only are form and meaning tightly related, Sinclair argues [3, p. 59f.], they must even be identical, considering that most ambiguities are resolved by context in authentic language usage. Hence, a description of lexical items should take into account both aspects at the same time.

Instead of describing the paradigmatic properties of each lexical item by listing the potential senses of its lemma, he pleads for describing both the syntagmatic and the paradigmatic properties of each lexical item as patterns in which the given lexical item occurs [3, p. 69].

1.2 Pattern Dictionary of English Verbs (PDEV)

Hanks, Sinclair's collaborator on the first corpus-based dictionary ever, the Collins Cobuild English Language Dictionary [5], has proposed *Corpus Pattern Analysis*

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 67–74, 2010.

(CPA), a semi-formal lexical description method that consistently materializes Sinclair's concept of capturing meanings in patterns of language rather than lexical units in the token-centered lexicographic tradition.

The current CPA captures "normal", i.e. reasonably frequent, usages of a given verb by sorting them into *patterns*. Each pattern is formulated as a proposition in which the verb in question is lemmatized[1] and its relevant collocates are classified by means of two sets of semantic labels or listed as *lexical sets*, depending on whether the respective collocates can be listed (as a lexical set) or grouped together under the general heading of a *Semantic Type*. Each proposition is paraphrased by a sentence in which the relevant pattern arguments are labeled identically with the proposition part. This paraphrase embodies the *implicature* (or *meaning potential*, see [1]) activated by that particular pattern.

Each collocate that cannot be represented by a lexical set, is described by a Semantic Type. Semantic types are sometimes augmented by a *Semantic Role*. The Semantic Types are a finite set of labels hierarchically ordered in what Hanks calls a *shallow semantic ontology* [2]. The Semantic Types describe inherent properties of the collocates, such as *Human*, *Artifact*, *Stuff*, *Document*. The Semantic Roles describe the properties that are assigned to the word in a particular pattern or context.

CPA is implemented as PDEV, the Pattern Dictionary of English Verbs, built by Hanks and his collaborators [2]. It comprises two interlinked components: a list of patterns for each verb and a reference set of manually tagged sample data. Each verb in PDEV is linked to a reference sample of concordances, which contain the verb in question. The sample is randomly selected from the British National Corpus (BNC) [6], and its size is typically 250–500, depending on the semantic complexity of the verb.

We perceive Hanks' patterns as a means of discrimination of Sinclairian lexical items, which, in their own right, imply what is usually referred to as "meaning". To the best of our knowledge, PDEV is the first real and conscious implementation of Sinclair's principles concerning the lexical item and the way it should be described. This fact makes PDEV unique, yet there are certainly a number of other projects that formally describe semantic distinctions of verb uses, with different theoretical foundations, e.g. [7,8].

1.3 PDEV as a Source for NLP?

Hanks' approach to the lexical description of verbs is novel and linguistically sound at the same time. It has gained a world-wide reputation, judging by the more than 600[2] topic-related citations for Hanks, as well as the numerous keynote speeches Hanks has been invited to give on this subject since the first significant mention of CPA in [1]. CPA is intuitively plausible, and its formal encoding appears promising for various applications in NLP – the more so because Hanks has been continuously linking his lexicon to other well-known lexical sources, such as FrameNet [7] or the Erlangen Valency Bank [9].

[1] Exception: passivization.

[2] Harzing's Publish or Perish since Hanks, 1994 (recorded as 1993), quoted 2010-03-24.

However, the “qualified judgment” on the hypothesized NLP usability of CPA pronounced by a number of language experts has not yet been experimentally tested.

With our experiments we are taking a first step towards providing a reliable assessment whether or not the current PDEV is suitable for NLP application. In this short paper we report on an on-going pilot study, in which we examine the consistency of PDEV, which we regard as the basic prerequisite for its NLP-usability. Should we identify problematic issues, we suggest (and plan to implement) improvements based on a pilot sample in the next step.

2 Current Status of PDEV Development

2.1 Platform of PDEV development

The development of PDEV is supported by two interconnected applications. The first one used for pattern editing is based on the “Dictionary Editor and Browser” tool (DEB), a dictionary-making database platform developed at the Masaryk University in Brno (MU), Czech Republic [10]. This platform enables the lexicographic processing of XML-encoded data through a user-friendly web-based graphical user interface integrated as an add-on in the Firefox-Mozilla web browser. The data is stored on DEB servers located at MU. PDEV is one of the numerous applications of DEB. It incorporates a tailored interface for pattern creating, ontology browsing and editing. The second application, used for concordance tagging, is a modified version of the Sketch Engine [11].

2.2 Current PDEV Statistics

PDEV has been developed on basis of verb occurrences in the BNC50 corpus, a 50-million-word part of the BNC. BNC50 contains almost 5,800 verb types occurring in 8 million verb tokens. However, about 41% of all verb tokens represent auxiliary (‘will’, ‘do’, ‘have’, ‘be’) or modal (‘shall’, ‘can’, ‘must’, etc.) verbs that are not analysed in the PDEV project at all. The number of lexical verb types in BNC50 is 5,757 and the total number of the corresponding tokens is 4,673,003. Table 1 illustrates the well known fact that rare words do not significantly contribute to corpus coverage. Verbs with frequency higher than 27 cover the corpus up to 99.5%.

Currently (March 2010) the number of verbs compiled in PDEV is 678, 11.8% of all lexical verb types in BNC50. The number of corresponding tokens in BNC50 is 495,724, which cover 10.6% of all BNC50 lexical verb tokens.

Table 1. The coverage of BNC50 verb tokens. For example, 918 most frequent verbs, each of which occurs at least 610 times in BNC50, cover more than 90% of all BNC50 lexical verb tokens.

min. frequency	54,872	8,723	610	246	136	90	48	28	1
verb types	7	120	918	1,519	2,030	2,452	3,151	3,780	5,757
BNC50 coverage	11%	50%	90%	95%	97%	98%	99%	99.5%	100%

The number of all patterns created for those compiled verbs is 2,572. While the average number of patterns per compiled verb type is 3.79, a more interesting value, the expected number of patterns per token is 9.72 (more frequent verbs often have also more patterns). The correlation between verb frequency and the number of patterns is shown in Fig. 1.

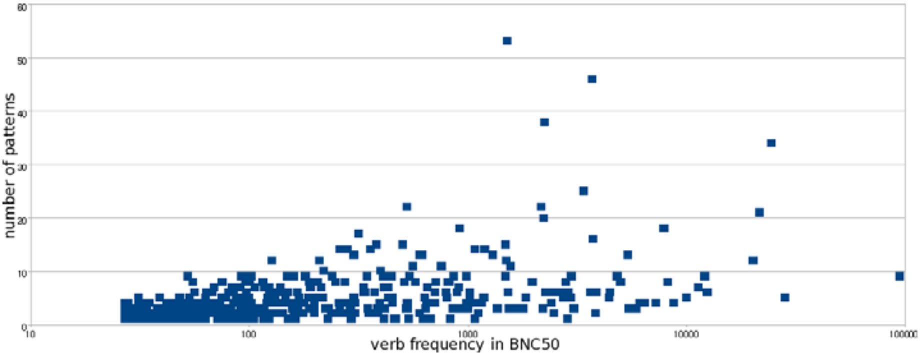

Fig. 1. The number of patterns of 502 PDEV verbs (with frequency at least 28) and their frequency in BNC50

3 First Evaluation of PDEV

3.1 Evaluation Method

PDEV with its tagged reference samples can be regarded as a manually created gold standard data set for machine-learning experiments. So far, the lexicon has mainly been built by Hanks. In terms of annotation, the entire data available has been annotated by one single annotator. Moreover, the author of the patterns and the data annotator are the same person. Our first question was therefore: are humans who did not create the entries themselves able to agree in pattern assignment? A reasonable degree of interannotator agreement is a prerequisite for any further automatic processing.

This assumption has two aspects, which we want to keep apart: creating the lexicon and annotating the data. Here we focus only on the consistency in tagging the data according to already existing patterns. We regard the mutual agreement of independently working annotators as a measure of quality of each given lexical entry.

As with any linguistically rich annotation, the annotators must be clearly instructed and trained before the interannotator agreement can be measured. The authors of this paper, who acted as annotators, have only learned details of the annotation procedure on-the-fly while discussing the patterns as well as own data findings with Hanks, watching him work and having ocassional hands-on experience with creating a new entry for more than one year. No detailed annotation guidelines were available at that point. We expected this fact to lower our inter-annotator agreement. While tagging, we kept each a record of difficult decisions for future reference when a regular annotation guide for new annotators is being created, and we analyzed our records along with the annotated data when all samples were finished. Hanks performed the same annotation with us, and his sample annotation served as a reference in case of doubt.

3.2 Experiments

For the first experiments we chose a pilot sample of 30 verbs selected from the set of complete compiled PDEV verbs. To measure the inter-annotator agreement we used the standard kappa function (Cohen's kappa for annotator pairs, and Fleiss' kappa for more than two annotators). Just for example, results for some of the pilot verbs are shown in Table 2.

Table 2. An example of 6 pilot verbs selected for validation and the interannotator agrement (IAA) measured on random samples selected from two different corpora. PEDT verbs were annotated by only 2 people.

Verbs	Verb Features			IAA on PEDT	IAA on BNC50	
	patterns	perplexity	BNC50 freq.	kappa	annotators	kappa
tell	21	3.80	21,550	0.66	2	0.27
lead	12	3.97	20,180	0.83	3	0.78
call	34	6.68	24,439	0.72	2	0.68
argue	7	1.73	11,362	0.93	3	1
claim	6	3.14	12,517	0.87	4	0.72
fire	15	7.96	1,488	0.71	2	0.42

To compare the inter-annotator agreement on different corpora, we used randomly selected concordance samples both from the BNC50 and from the Prague English Dependency Treebank (PEDT) [12]. PEDT consists of Wall Street Journal articles. The results show that a domain-restricted corpus sometimes implies better inter-annotator agreement. On the other hand, we are aware that patterns designed to fit the BNC50 corpus do not necessarily fit another corpus (the more so a strongly domain-restricted one).

3.3 Disagreement Analysis

We have identified the following types of disagreements:

1. *Vagueness of instructions on context consideration.* Whether, to what extent, and how wide the context is to be taken into account is not yet clearly defined in the theoretical foundations of CPA.
2. *Markable-Unmarkable.* Sometimes it is difficult to decide whether a form is a markable or an unmarkable; e.g. in participial forms.

3. *Elliptical usages.* Elliptical usages are problematic because of their inherent ambiguity. There are two types of ambiguous ellipses: a) the context does not enable a distinction to be made between two potentially relevant patterns, of which one requires zero realization of an argument and the other allows optional omission; or b) there are two patterns with different implicatures, both allowing optional omission of an argument. In such cases, it is not possible to say what the collocation would look like if the ellipsis was restored; so the text meaning can only be determined by examination of the wider context, which is beyond the scope of CPA.
4. *Collocate matches several Semantic Types.* In a few cases, the context fits more than one pattern by its implicature, and one of the pattern-relevant collocates of the verb has several inherent semantic features, of which each allows the collocate to match a different Semantic Type in different patterns.
5. *Insufficient competence in English.* Sometimes, the non-native annotators misunderstood a concordance.
6. *Missing pattern.* The concordances had been taken mainly from BNC, but some tasks contained only concordances from PEDT, which we regard as a domain-restricted corpus. Some usages frequent in PEDT were not explicitly captured by the patterns based on BNC, and the annotators tagged them as exploitations of various different patterns, in which they disagreed. Some (rare) suggestions to add a pattern arose also from the BNC annotation.
7. *Implicatures too fine-grained.* The random sample showed in some cases that the context often does not allow for disambiguation of very fine-grained distinctions between implicatures activated by different patterns.
8. *Semantic Types too fine-grained.* Some (in fact quite numerous) concordances did not match a pattern because the collocates in question did not match the Semantic Type prescribed by the pattern, although intuitively it seemed to fit well with that pattern, too.

3.4 Discussion

The results of the pilot project measuring inter-annotator agreement were not entirely impressive. However, *only a few cases pointed at pattern inadequacy*. Our findings are not too different from the recently published analysis of annotation of polysemous predicates [13].

The annotation procedure had not become routine yet. Many errors were simply oversights: it happened e.g. that an annotator consistently confused one pattern number for another throughout one entire sample, a few concordances were misunderstood and the annotators also sometimes forgot about the fact that one single implicature is split into different patterns when a collocate is typically realized both as a noun or a verb clause (nouns are described by Semantic Types, while verb clauses by syntax), and he/she kept assigning only the one with the Semantic Type to it.

The most frequent type of frame inadequacy that we encountered is easily amended by adding a Semantic Type. We had decided to strictly classify all concordances as "unmarkable", whenever a concordance intuitively perceived as typical and norm-conforming (i.e. not an "exploitation") contained a collocate that was not included in the Semantic Types. This happened quite often, since some of the entries analysed had been finished long ago and were compiled with an outdated set of Semantic Types.

Missing/competing patterns, however, were rare, which is a good sign.

4 Prospects and Conclusions

The pilot annotation experiments were conducted to uncover potential problems prior to any large-scale annotation. The experiments helped us specify which issues must be particularly looked into when an annotation manual is being written. Once disagreements caused by points 1, 2, 3 and 5 (Section 3.3) have been eliminated by a better instruction specification and, hopefully, by hiring native speakers, the annotation will provide valuable feedback for pattern building. We suggest the following procedure for the "validation" of lexical entries in PDEV:

1. The initial patterns will be created by Hanks as usual.
2. When declared ready for validation, they will be given to annotators, along with the tagged reference data.
3. The annotators will tag a new randomly selected BNC sample and keep notes on potentially missing patterns, incomprehensible context, etc.
4. The interannotator agreement will be measured, disagreements identified and discussed with Hanks.
5. Based on the disagreement analysis, the patterns will be revised and/or the annotation instructions enhanced.
6. The revised entries will be returned to the annotators, along with a (different) data sample to be tagged.
7. The entire process will be repeated until the inter-annotator agreement (at least in the most relevant points, such as 4, 6 and 7 listed in Section 3.3) has risen to an acceptable level.
8. Each revised lexical entry will be declared as "validated" and ready for machine-learning experiments.

The PDEV patterns seem to be a very promising way of describing verbs as Sinclairian Lexical Items. From the machine-learning view, the pattern inventory and the tagged reference data attached represent two complementary sources, which enable the combination of rule-based and statistical approaches to automatic verb disambiguation. The current PDEV needs increased standardization and regular evaluation of new entries by iterated multiple annotation and interannotator agreement measuring. Our pilot study is a first step towards a systematic validation and building-up PDEV as an NLP-applicable lexical source.

Acknowledgments

This project was kindly supported by EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), the Czech Science Foundation GA-P406/2010/0875 and the Czech Ministry of Education MSM 0021620838.

This work has been partly supported by the Ministry of Education of CR within the National Research Programme II project 2C06009 and by the Czech Science Foundation under the projects P401/10/0792, 102/09/1842 and 407/07/0679.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement number 248,307.

We would like to thank Patrick Hanks for his comments on an earlier draft of this paper. We also thank three anonymous reviewers for their valuable feedback.

References

1. Hanks, P.: Linguistic Norms and Pragmatic Exploitations, Or Why Lexicographers need Prototype Theory, and Vice Versa. In: Kiefer, F., Kiss, G., Pajzs, J. (eds.) Papers in Computational Lexicography: Complex 1994. Hungarian Academy of Sciences, Budapest (1994)
2. Hanks, P., Pustejovsky, J.: A Pattern Dictionary for Natural Language Processing. Revue Francaise de linguistique appliquée 10(2) (2005)
3. Sinclair, J.: The Lexical Item. In: Hanks, P. (ed.) Lexicology: Critical Concepts in Linguistics, 6 vols. Routledge, London; First published on Weigand, E. (ed.): Contrastive Lexical Semantics Amsterdam, pp. 1–24. John Benjamins, Amsterdam (1998, 2008)
4. Hanks, P.: The Lexicographical Legacy of John Sinclair. International Journal of Lexicography 21(3) (2008)
5. Sinclair, J., Hanks, P., et al.: The Collins Cobuild English Language Dictionary. HarperCollins, New York (1987)
6. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on Behalf of the BNC Consortium (2007)
7. Ruppenhofer, J., Baker, C.F., Fillmore, C.J.: The FrameNet Database and Software Tools. In: Braasch, A., Povlsen, C. (eds.) Proceedings of the Tenth Euralex International Congress, Copenhagen, Denmark, vol. I, pp. 371–375 (2002)
8. Palmer, M., Kingsbury, P., Gildea, D.: The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics 31(1), 71–106 (2005)
9. Herbst, T., et al.: A Valency Dictionary of English: a Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives. De Gruyter, Berlin (2004)
10. Horák, A., Rambousek, A.: Server for Dictionary Editor and Browser (DEB) Platform (2008), `http://deb.fi.muni.cz/`
11. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. In: Proceedings of the Eleventh EURALEX International Congress, pp. 105–116. Universite de Bretagne-Sud, Lorient (2004)
12. Cinková, S., Toman, J., Hajič, J., Čermáková, K., Klimeš, V., Mladová, L., Šindlerová, J., Tomšů, K., Žabokrtský, Z.: Tectogrammatical Annotation of the Wall Street Journal. In: The Prague Bulletin of Mathematical Linguistics, vol. (92). Charles University in Prague (2009)
13. Rumshisky, A., Batiukova, O.: Polysemy in Verbs: Systematic Relations between Senses and their Effect on Annotation. In: COLING Workshop on Human Judgement in Computational Linguistics (HJCL 2008), Manchester (2008)