Idiomatic object usage and support verbs

Pasi Tapanainen, Jussi Piitulainen and **Timo Järvinen***
Research Unit for Multilingual Language Technology
P.O. Box 4, FIN-00014 University of Helsinki, Finland
http://www.ling.helsinki.fi/

1 Introduction

Every language contains complex expressions that are language-specific. The general problem when trying to build automated translation systems or human-readable dictionaries is to detect expressions that can be used idiomatically and then whether the expressions can be used idiomatically in a particular text, or whether a literal translation would be preferred. It follows from the definition of idiomatic expression that when a complex expression is used idiomatically, it contains at least one element which is semantically "out of context". In this paper, we discuss a method that finds idiomatic collocations in a text corpus. The method detects semantic asymmetry by taking advantage of differences in syntactic distributions.

We demonstrate the method using a specific linguistic phenomenon, verb-object collocations. The asymmetry between a verb and its object is the focus in our work, and it makes the approach different from the methods that use e.g. mutual information, which is a symmetric measure.

Our novel approach differs from mutual information and the so-called t-value measures that have been widely used for similar tasks, e.g., Church et al. (1994) and Breidt (1993) for German. The tasks where mutual information can be applied are very different in nature as we see in the short comparison at the end of this paper. The work reported in Grefenstette and Teufel (1995) for finding *empty support verbs* used in nominalisations is also related to the present work.

* Email: *Pasi.Tapanainen@ling.helsinki.fi*, *Jussi.Piitulainen@ling.helsinki.fi* and *Timo.Jarvinen@ling.helsinki.fi*, Parsers & demos: *http://www.conexor.fi*.

2 Semantic asymmetry

The linguistic hypothesis that syntactic relations, such as subject-verb and object-verb relations, are *semantically asymmetric* in a systematic way (Keenan, 1979) is well-known. McGlashan (1993, p. 213) discusses Keenan's principles concerning directionality of agreement relations and concludes that *semantic interpretation of functor categories varies with argument categories, but not vice versa.* He cites Keenan who argues that the meaning of a transitive verb depends on the object, for example the meaning of the verb *cut* seems to vary with the direct object:

- in *cut finger* "to make an incision on the surface of",
- in *cut cake* "to divide into portions",
- in *cut lawn* "to trim" and
- in *cut heroin* "diminish the potency".

This phenomenon is also called *semantic tailoring* (Allerton, 1982, p. 27).

There are two different types of asymmetric expressions even if they probably form a continuum: those in which the sense of the functor is *modified* or *selected* by a dependent element and those in which the functor is *semantically empty.* The former type is represented by the verb *cut* above: a distinct sense is selected according to the (type of) object. The latter type contains an object that forms a fixed collocation with a semantically empty verb. These pairings are usually language-specific and semantically unpredictable.

Obviously, the amount of tailoring varies considerably. At one end of the continuum is idiomatic usage. It is conceivable that even a highly idiomatic expression like *taking toll* can

be used non-idiomatically. There may be texts where the word *toll* is used non-idiomatically, as it also may occur from time to time in any text as, for instance, in *The Times* corpus: *The IRA could be profiting by charging a toll for cross-border smuggling.* But when it appears in a sentence like *Barcelona's fierce summer is taking its toll*, it is clearly a part of an idiomatic expression.

3 Distributed frequency of an object

As the discussion in the preceding chapter shows, we assume that when there is a verb-object collocation that can be used idiomatically, it is the object that is the more interesting element. The objects in idiomatic usages tend to have a distinctive distribution. If an object appears only with one verb (or few verbs) in a large corpus we expect that it has an idiomatic nature. The previous example of *take toll* is illustrative: if the word *toll* appears only with the verb *take* but nothing else is done with tolls, we may then assume that it is not the *toll* in the literary sense that the text is about.

The task is thus to collect verb-object collocations where the object appears in a corpus with few verbs; then study the collocations that are topmost in the decreasing order of frequency.

The restriction that the object is always attached to the same verb is too strict. When we applied it to ten million words of newspaper text, we found out that even the most frequent of such expressions, *make amends* and *take precedence*, appeared less than twenty times, and the expressions *have temerity*, *go berserk* and *go ex-dividend* were even less frequent. It was hard to obtain more collocations because their frequency went very low. Then expressions like *have appendix* were equivalently exposed with expressions like *run errand*.

Therefore, instead of taking the objects that occur with only one verb, we take all objects and distribute them over their verbs. This means that we are concerned with all occurrences of an object as a block, and give the block the score that is the frequency of the object divided by the number of different verbs that appear with the object.

The formula is now as follows. Let $\mathbf{o}$ be an object and let

$$\langle F_1, V_1, \mathbf{o}\rangle, \ldots, \langle F_n, V_n, \mathbf{o}\rangle$$

be triples where $F_j > 0$ is the frequency or the relative frequency of the collocation of $\mathbf{o}$ as an object of the verb V_j in a corpus. Then the score for the object $\mathbf{o}$ is the sum $\sum_{k=1}^{n} F_k/n$.

The frequency of a given object is divided by the number of different verbs taking this given object. If the number of occurrences of a given object grows, the score increases. If the object appears with many different verbs, the score decreases. Thus the formula favours common objects that are used in a specific sense in a given corpus.

This scheme still needs some parameters. First, the distribution of the verbs is not taken into account. The score is the same in the case where an object occurs with three different verbs with the frequencies, say, 100, 100, and 100, and in the case where the frequencies of the three heads are 280, 10 and 10. In this case, we want to favour the latter object, because the verb-object relation seems to be more stable with a small number of exceptions. One way to do this is to sum up the squares of the frequencies instead of the frequencies themselves.

Second, it is not clear what the optimal penalty is for multiple verbs with a given object. This may be parametrised by scaling the denominator of the formula. Third, we introduce a threshold frequency for collocations so that only the collocations that occur frequently enough are used in the calculations. This last modification is crucial when an automatic parsing system is applied because it eliminates infrequent parsing errors.

The final formula for the distributed frequency $DF(\mathbf{o})$ of the object $\mathbf{o}$ in a corpus of n triples $\langle F_j, V_j, \mathbf{o}\rangle$ with $F_j > C$ is the sum

$$\sum_{k=1}^{n} \frac{F_k^a}{n^b}$$

where a, b and C are constants that may depend on the corpus and the parser.

4 The corpora and parsing

4.1 The syntactic parser

We used the Conexor Functional Dependency Grammar (FDG) by Tapanainen and Järvinen (1997) for finding the syntactic relations. The new version of the syntactic parser can be tested at `http://www.conexor.fi`.

4.2 Processing the corpora

We analysed the corpora with the syntactic parser and collected the verb-object collocations from the output. The verb may be in the infinitive, participle or finite form. A noun phrase in the object function is represented by its head. For instance, the sentence *I saw a big black cat* generates the pair ⟨*see, cat*⟩. A verb may also have an infinitive clause as its object. In such a case, the object is represented by the infinitive, with the infinitive marker if present. Naturally, transitive nonfinite verbs can have objects of their own. Therefore, for instance, the sentence *I want to visit Paris* generates two verb-objects pairs: ⟨*want, to visit*⟩ and ⟨*visit, Paris*⟩. The parser recognises also clauses, e.g. *that*-clauses, as objects.

We collect the verbs and head words of nominal objects from the parser's output. Other syntactic arguments are ignored. The output is normalised to the baseforms so that, for instance, the clause *He made only three real mistakes* produces the normalised pair: ⟨*make, mistake*⟩. The tokenisation in the lexical analysis produces some "compound nouns" like *vice+president*, which are glued together. We regard these compounds as single tokens.

The intricate borderline between an object, object adverbial and mere adverbial nominal is of little importance here, because the latter tend to be idiomatic anyway. More importantly, due to the use of a syntactic parser, the presence of other arguments, e.g. subject, predicative complement or indirect object, do not affect the result.

5 Experiments

In our experiment, we used some ten million words from a *The Times* newspaper corpus, taken from the *Bank of English* corpora (Järvinen, 1994). The overall quality of the result collocations is good. The verb-object collocations with highest distributed object frequencies seem to be very idiomatic (Table 1).

DF(o)	F(vo)	verb + object
37.50	73	take toll
28.00	28	go bust
25.00	25	make plain
24.83	60	mark anniversary
22.00	22	finish seventh
21.00	21	make inroad
21.00	21	do homework
21.00	21	have hesitation
20.40	93	give birth
19.50	28	have a=go
19.25	128	make mistake
18.00	18	go so=far=as
18.00	18	take precaution
17.50	76	look as=though
17.50	61	commit suicide
17.25	62	pay tribute
17.04	817	take place
17.00	17	make mockery
17.00	17	make headway
16.29	152	take wicket
16.17	319	cost £
16.00	16	have qualm
16.00	16	make pilgrimage
15.69	248	take advantage
15.57	84	make debut
15.00	15	have second=thought
14.57	190	do job
14.50	27	finish sixth
14.50	16	suffer heartattack
14.47	165	decide whether
14.14	110	have impact
14.12	329	have chance
14.00	133	give warn
14.00	14	have sexual=intercourse
14.00	14	take plunge
14.00	14	have misfortune
14.00	14	thank goodness
13.90	226	have nothing
13.63	131	make money
13.50	25	strike chord

Table 1: Verb-object collocations from The Times

The collocations seem to have different status in different corpora. Some collocations appear in every corpus in a relatively high position. For example, collocations like *take toll*, *give birth* and *make mistake* are common English expressions.

Some other collocations are corpus specific. An experiment with the *Wall Street Journal* corpus contains collocations like *name vice+precident* and *file lawsuit* that are rare in the British corpora. These expressions could be categorised as cultural or area specific. They are

F	MI (scaled)	t-value (scaled)	Verb + object
15	9.47	3.87	wreak havoc
12	8.62	3.46	armour carrier
11	8.48	3.32	grasp nettle
14	8.42	3.74	firm 1p
12	8.30	3.46	bury Edmund
13	8.21	3.60	weather storm
21	8.18	4.58	bid farewell
12	8.17	3.46	strut stuff
18	8.10	4.24	breathe sigh
10	8.10	3.16	suck toe
13	8.05	3.60	incur wrath
12	8.03	3.46	invade Kuwait
11	7.92	3.31	protest innocence
17	7.91	4.12	hole putt
13	7.91	3.60	poke fun
11	7.80	3.31	tighten belt
12	7.72	3.46	stem tide
11	7.72	3.31	heal wound

Table 2: Collocations according to mutual information filtered with t-value of 3

position	verb + object
124	finish seventh
157	mark anniversary
478	go bust
770	do homework
862	give birth
1009	make inroad
1033	take toll
1225	make mistake
1244	make plain
1942	have hesitation
2155	have a=go

Table 3: The order of top collocations according to mutual information

likely to appear again in other issues of WSJ or in other American newspapers.

6 Mutual information

Mutual information between a verb and its object was also computed for comparison with our method. The collocations from The Times with the highest mutual information and high t-value are listed in Table 2. See Church et al. (1994) for further information. We selected the t-value so that it does not filter out the collocations of Table 1. Mutual information is computed from a list of verb-object collocations.

frequency	verb + object
329	have chance
302	have it
274	have time
256	have effect
247	have right
229	have problem
226	have nothing
210	have little
203	have idea
186	have power
164	have what
155	have much
142	have child
139	have experience
138	have some
135	have reason
132	have one
123	have advantage
122	have intention
119	have plan

Table 4: What do we have? – Top-20

The first impression, when comparing Tables 1 and 2, is that the collocations in the latter are somewhat more marginal though clearly semantically motivated. The second observation is that the top collocations contain mostly rare words and parsing errors made by the underlying syntactic parser; three out of the top five pairs are parsing errors.

We tested how the top ten pairs of Table 1 are rated by mutual information. The result is in Table 3 where the *position* denotes the position when sorted according to mutual information and filtered by the t-value. The t-value is selected so that it does not filter out the top pairs in Table 1. Without filtering, the positions are in range between 32 640 and 158 091. The result shows clearly how different the nature of mutual information is. Here it seems to favour pairs that we would like to rule out and vice versa.

frequency	verb + object
21	have hesitation
28	have a=go
16	have qualm
15	have second=thought
110	have impact
329	have chance
14	have sexual=intercourse
14	have misfortune
226	have nothing
135	have reason
117	have choice
274	have time
41	have regard
28	have no=doubt
256	have effect
18	have bedroom
17	have regret
10	have penchant
10	have pedigree
10	have clout

Table 5: The collocations of the verb *have* sorted according to the DF function

7 Frequency

In a related piece of work, Hindle (1994) used a parser to study what can be done with a given noun or what kind of objects a given verb may get. If we collect the most frequent objects for the verb *have*, we are answering the question: *"What do we usually have?"* (see Table 4). The distributed frequency of the object gives a different flavour to the task: if we collect the collocations in the order of the distributed frequency of the object, we are answering the question: *"What do we only have?"* (see Table 5).

8 Conclusion

This paper was concerned with the semantic asymmetry which appears as syntactic asymmetry in the output of a syntactic parser. This asymmetry is quantified by the presented distributed frequency function. The function can be used to collect and sort the collocations so that the (verb-object) collocations where the asymmetry between the elements is the largest come first. Because the semantic asymmetry is related to the idiomaticity of the expressions, we have obtained a fully automated method to find idiomatic expressions from large corpora.

References

D. J. Allerton. 1982. *Valency and the English Verb.* London: Academic Press.

Elisabeth Breidt. 1993. Extraction of V-N-collocations from text corpora: A feasibility study for German. *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 74–83, June.

Kenneth Ward Church, William Gale, Patrick Hanks, Donald Hindle, and Rosamund Moon. 1994. Lexical substitutability. In B.T.S. Atkins and A Zampolli, editors, *Computational Approaches to the Lexicon*, pages 153–177. Oxford: Clarendon Press.

Gregory Grefenstette and Simone Teufel. 1995. Corpus-based method for automatic identification of support verbs for nominalizations. *Proceedings of the 7th Conference of the European Chapter of the ACL*, March 27-31.

Donald Hindle. 1994. A parser for text corpora. In B.T.S. Atkins and A Zampolli, editors, *Computational Approaches to the Lexicon*, pages 103–151. Oxford: Clarendon Press.

Järvinen, Timo. 1994. Annotating 200 Million Words: The Bank of English Project. *COLING 94. The 15th International Conference on Computational Linguistics Proceedings.* pages 565–568. Kyoto: Coling94 Organizing Committee.

Edward L. Keenan. 1979. On surface form and logical form. *Studies in the Linguistic Sciences*, (8):163–203. Reprinted in Edward L. Keenan (1987). *Universal Grammar: fifteen essays.* London: Croom Helm. 375-428.

Scott McGlashan. 1993. Heads and lexical semantics. In Greville G. Corbett, Norman M. Fraser, and Scott McGlashan, editors, *Heads in Grammatical Theory*, pages 204–230. Cambridge: CUP.

Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71, Washington, D.C.: ACL.