DE GRUYTER MOUTON DOI 10.1515/phras-2015-0008 — YoP 2015; 6: 125–130

Koenraad Kuiper

On Pawley's conjecture

Abstract: This paper shows that Pawley's conjecture that the frequency of lexical items in text corpora is positively correlated with the number of phrasal lexical items which have those lexical items as heads of phrase is confirmed. Data for testing Pawley's conjecture are taken from two sources: Kilgarriff's lemmatized frequency lists from the BNC of the 6,318 words which appear more than 800 times (http://www.kilgarriff.co.uk) and the around 14,000 PLIs in the *Syntactically Annotated Idiom Dictionary* (Kuiper et al., 2003). Why this statistical fact should be the case is a matter for further research.

Keywords: syntactic head, PLI frequency, Pawley

1 Introduction

The vocabulary items in the lexicon of a mature native speaker can be classified according to their structural properties. This classification can be narrow so that it includes only words or broad so that it includes any lexicalized expression longer than a word. A broad classification would include items which are mono-morphemic such as *dog*, derived words such as *doggy*, compounds such as *dog biscuit*, and phrasal lexical items (PLIs) such as the idiom *BE in the doghouse*. PLIs are phrases consisting of words though sometimes these are bound words such as *kith* in *kith and kin* and *umbrage* in *take umbrage at*. The size of the phrasal vocabulary in comparison with the word vocabulary is contested with Jackendoff (1995) and Williams (1994) suggesting it is about the size of the single

Koenraad Kuiper, Linguistics Department University of Canterbury & Linguistics Department, University of Sydney, Unit 312 1 Marian St, Redfern, NSW 2016, Australia. E-mail: koenraad.kuiper@canterbury.ac.nz

word vocabulary or more, while Pawley and Syder (1983) and Mel'čuk (1995) suggest it is bigger by multiples of as much as an order of magnitude.

To justify this latter position Pawley proposes a number of thought experiments. One asks readers to count the number of PLIs in which individual words participate. This leads Pawley (1985: 93) to the view that '[t]he most common verbs such as *have*, *get*, *take*, *go*, *come* and *do* each figure in several hundred formulas. Verbs of the second and third rank in frequency, such as *know*, *see* and *think*, each figure in several score.' This can be termed Pawley's conjecture.

Pawley's conjecture has not been subject to empirical test. To do so requires it to be put into a more testable form, namely the form of a hypothesis. Pawley's hypothesis is that the number of PLIs containing a particular verb is positively correlated with the frequency of that verb. Increasing the testability a little further one might propose that the number of existing PLIs with a particular lexical head of phrase is positively correlated with the frequency of that head of phrase. That is the hypothesis which will be tested below by reference to three types of phrase, those headed by prepositions, verbs and nouns.

These heads have been selected because they provide a reasonable basis for testing Pawley's hypothesis. First, prepositions and verbs are both argument-taking predicates (Kearns, 2011) and tend to form PLIs with lexicalized complements. So we find cases like: *in despair, on time*, *for sale*, *in the nick of time*, *on the bones of NP's bum*, *by your leave*; *sit an exam*, *put NP's house in order*, *eat humble pie* and *stew in NP's own juice*. Furthermore, verbs participate in numerous verb-particle combinations, which can be seen as verbs with complements as (typically) NP gaps, e.g. *bring* NP *down*, *talk* NP *through*, *see* NP *off*.[1] Prepositions are a small closed set with relatively high frequency in text.

Second, verbs are an open set and productive heads of PLIs. Major English dictionaries of PLIs (Courteney, 1983; Cowie & Mackin, 1975; Cowie, Mackin, & McCaig, 1983; Long, 1979) contain more verb-headed PLIs than those headed by other lexical heads. Nouns are also, like verbs, a large open class set but they are not typically heads of predicate-argument structures. Thus the three heads: prepositions, verbs and nouns, provide a good basis to test Pawley's hypothesis.

1 Emonds (1976) suggests that the underlying form of many of these cases has the preposition (or particle to the right of the NP slot complement).

2 Methodology

2.1 Data

Data for testing Pawley's hypothesis are taken from two sources: the BNC's 10 million word corpus and Kilgarriff's lemmatized frequency lists from the BNC of the 6,318 words which appear more than 800 times (http://www.kilgarriff.co.uk) and the around 14,000 PLIs in the *Syntactically Annotated Idiom Dictionary* (Kuiper et al., 2003). This dictionary is the largest data set of PLIs containing syntactic annotation, which allows for the selection of heads of phrase from PLIs.

Limitations of the data should be mentioned at this juncture. The limit of 8 occurrences per 1 million words leaves many words off the Kilgarriff list. However, since the frequencies of these words is below statistical significance, this is a necessary limitation. The BNC is a balanced corpus and so the frequencies of its words will be influenced by the nature of the balance. It may be that PLIs are much more likely to be headed by words which have a high frequency in spoken English while the BNC contains 90% written text.

Both the BNC and the dictionaries from which the SAID data are taken are from British English. This ignores other varieties of English but does provide a uniform variety of English for the comparison.

The major shortcoming is the data from SAID. If Mel'čuk is correct and a native speaker knows around an order of magnitude more PLIs than single words, (PLIs include idioms, collocations, and clichés), then the 14,000 idiom PLIs in the SAID corpus cannot be in any way representative of such a much larger vocabulary of PLIs. Furthermore SAID is based on idiom dictionaries and idioms are a relatively small subset of PLIs (Mel'čuk, 2012). For example, the prepositionally headed PLI *beneath contempt* is not present nor is *among other things* possibly because they are almost compositional. The PLIs in SAID are also taken from the general vocabulary and do not come from the many specialized varieties of English which are represented in the BNC. For example, no PLIs in SAID come from oral communication among aircraft maintenance engineers (Newsome, 2006). However since SAID is the largest source of syntactically annotated PLIs it is the best source available, given the task.

2.2 Method

The BNC frequencies of each head of phrase were correlated with the number of PLIs with that head of phrase using a Pearson's correlation.

3 Results

Results are as follows:

Table 1: Correlations between the frequencies of heads of phrase and the number of PLIs with that head of phrase

	Prepositions	Verbs	Nouns
N =	67	1275	3335
Pearson's correlation (r)	.63	.60	.54
One tailed probability	0.0	0.0	0.0
Two tailed probability	0.00000001	0.0	0.0

Pawley's conjecture is confirmed in a statistical sense given that the correlations are all positive, reasonable high and statistically significant.

4 Discussion

The question now arises as to why Pawley's conjecture should be corroborated. It is not Pawley's 'law' and there are many examples of significant sets of PLIs with low frequency heads. For instance, *out* has a relatively low frequency as a preposition (rank 3064 in the Kilgarriff word frequency list) but there are four PLIs with it as head in SAID:

[PP[Pout][PP[Pof][NP[DETthe][Ncorner][PP[Pof][NP[POSS[NP[PRONone]]'s][Neye]]]]]]

[PP[Pout][PP[Pof][NP[DETthe][Nmouth(s)][PP[Pof][NP[NP[Nbabes]][CONJand][NP[Nsucklings]]]]]]]

[PP[Pout][PP[Pof][NP[Nhand]]]]

[PP[Pout][PP[Pof][NP[Nhours]]]]

The low frequency verb *pile* (rank 6117 in the Kilgarriff word frequency list) has twelve PLIs with it as head in SAID:

[VP[Vpile][NP[DETthe][Npressure]][PP[Pon]]]

[VP[Vpile][NP][PP[Pon]]]

[VP[Vpile][NP][PP[Ponto][NP]]]

[VP[Vpile][NP][PP[Pup]]]

[VP[Vpile][NP][PP[Pwith][NP]]]

[VP[Vpile][NP[PRONit]][PP[Pon]]]

[VP[Vpile][PP[Pin]]]

[VP[Vpile][PP[Pinto][NP]]]

[VP[Vpile][PP[Pon]][NP[DETthe][Nagony]]]

[VP[Vpile][PP[Pon]][NP[Nruns]]]

[VP[Vpile][PP[Pout]]]

[VP[Vpile][PP[Pout][PP[Pof][NP]]]]

On the other hand one reason why the correlations between heads of phrase, and PLIs having that head of phrase, are relatively high is that, for most of the lower frequency heads there are no PLIs at all in SAID.

We now enter the realm of speculation. If Pawley's conjecture is correct then we can suppose that since high frequency heads are semantically central vocabulary, they are learned earlier and by more speakers than low frequency words. Thus they may be more prone to have PLIs formed with them as heads. This is manifest in some Papuan languages with very small inventories of verb roots in the order of a few dozen (Pawley, 2006).

5 Conclusion

It appears (with the many caveats arising from the shortcomings of the data sets available for the study of PLIs) that Pawley's conjecture that high frequency words have more PLIs with them as heads than lower frequency words is corroborated. Why that should be the case is a subject for further investigation.

References

Courteney, Rosemary. 1983. *Longman dictionary of phrasal verbs*. Harlow: Longman.

Cowie, Anthony P. & Ronald Mackin (eds.). 1975. *Oxford dictionary of current idiomatic English: Verbs with prepositions and particles.* Oxford: Oxford University Press.

Cowie, Anthony P., Mackin, Ronald & Isabel R. McCaig 1983. *Oxford dictionary of current idiomatic English: Phrase, clause and sentence idioms*. Oxford: Oxford University Press.

Emonds, Joseph. 1976. *A transformation approach to English syntax*. NY: Academic Press.
Jackendoff, Ray S. 1995. The boundaries of the lexicon. In Martin Everaert, Erik-Jan van der Linden, André Schenk & Rob Schroeder (eds.), *Idioms: Structural and psychological perspectives,* 133–165. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Kearns, Kate. 2011. *Semantics* (2nd ed.). Basingstoke: Palgrave Macmillan.
Kilgarriff, Adam 1995. BNC database and word frequency lists. Retrieved 1 December 2013, from http://www.kilgarriff.co.uk/BNC_lists/lemma.al
Kuiper, Koenraad, Heather McCann, Therese Aitchison & Kees van der Veer. 2003. *SAID: A syntactically annotated idiom database*. Philadelphia: Linguistics Data Consortium, University of Pennsylvania
Long, Thomas Hill. 1979. *Longman dictionary of English idioms*. Harlow: Longman.
Mel'čuk, Igor. 1995. Phrasemes in language and phraseology in linguistics. In Martin Everaert, Erik-Jan van der Linden, André Schenk & Rob Schroeder eds., *Idioms: Structural and psychological pespectives,* 167–232. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Mel'čuk, Igor. 2012. Phraseology in the language, in the dictionary, and in the computer. In Koenraad Kuiper (ed.), *Yearbook of phraseology* (Vol. 3, 31–56). Berlin: de Gruyter.
Newsome, Georgina. 2006. *Estimating the size of the phrasal E lexicon of aircraft engineers*. Christchurch: University of Canterbury MA thesis.
Pawley, Andrew. 1985. On speech formulas and linguistic competence. *Lenguas Modernas, 12*. 84–104.
Pawley, Andrew. 2006. *Where have all the verbs gone? Remarks on the organisation of languages with small, closed verb classes.* Paper presented at the 11th Binnenial Rice Univerity Linguistics Symposium.
Pawley, Andrew & Frances Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack Richards & Richard Schmidt eds., *Language and communication,* 191–226. London: Longman.
Williams, Edwin. 1994. Remarks on lexical knowledge. *Lingua, 92*(1). 7–34.