The VNC-Tokens Dataset

Paul Cook, Suzanne Stevenson, and Afsaneh Fazly

1 VNC-Tokens

Many English idioms are composed of a verb and a noun in its direct object position, e.g., *give the sack*, *make a face*, and *see stars*. We refer to such expressions as verb–noun combinations (VNCs). VNCs may be used with their idiomatic meaning, e.g., *The little girl* ***made*** *a funny* ***face*** *at her mother*, or as a literal combination, e.g., *She* ***made*** *a* ***face*** *on the snowman using a carrot and two buttons*. This document describes VNC-Tokens, a dataset of VNC tokens and their corresponding annotation as a literal or idiomatic usage. The data used by Cook et al. (2007) is roughly a subset of VNC-Tokens, and indeed some of the description of this dataset is taken directly from that paper.

2 Expressions

We begin with the dataset used by Fazly and Stevenson (2006), which includes a list of VNCs. We eliminate from this list any expression whose frequency in the British National Corpus (BNC, Burnard, 2000) is less than 20 or does not occur in at least one of two idiom dictionaries (Cowie et al., 1983; Seaton and Macaulay, 2002). This results in 60 candidate expressions.

Two expert judges, both native English-speaking authors of this paper, examined the candidate expressions and eliminated 7 of them. In some cases this was done either because an annotator was not familiar with the expression, or because the idiomatic and literal senses were not clear to them. In other cases, expressions were removed because the literal usage of the expression did not seem plausible to an annotator. Some of the expressions seemed to be mainly used as verb-particle constructions or light-verb constructions. Although such expressions may be, to varying degrees, idiomatic, they are

not the focus of this annotation project and were therefore discarded. This gave a final set of 53 expressions.

3 Sentence Extraction

For each expression, 100 sentences containing its usage were randomly selected from the BNC, automatically parsed using Collins's (1999) parser. For expressions with less than 100 usages, all usages were extracted.

This dataset was originally created using the BNC World edition for which licenses are no longer available. A number of files occurring in this version of the BNC are not part of the newer BNC XML edition. Therefore the 8 sentences extracted from these files have been eliminated from this release.

We observed that there were a number of duplicates in our selected sentences. To ensure consistency across the expressions, we therefore also extracted any sentence which contained the same text as any one of the sentences in our dataset. Thus, all expressions have all duplicates included for any originally selected sentence. The final dataset consists of 2984 VNC tokens, of which 2920 are unique occurrences.

4 Token Annotation

The two judges from Section 2 annotated each instance of our 53 expressions in the extracted sentences as one of literal, idiomatic, or unknown. When annotating, a judge had access to only the sentence in which a VNC usage occurred, and not the surrounding sentences. If this context was insufficient to determine the class of the expression, the judge assigned the unknown label.

Idiomaticity is not a binary property, rather it is known to fall on a continuum from completely semantically transparent, or literal, to entirely opaque, or idiomatic. The human annotators were required to pick the label, literal or idiomatic, that best fit the usage in their judgment; they were not to use the unknown label for intermediate cases.

Some situations towards either end of the literal–idiomatic continuum are worth noting. Many idioms are highly semantically opaque, as in *hit the roof*, where the idiomatic interpretation has at most a very indirect or metaphorical relation to its literal meaning. However, an idiomatic usage may be more

directly related to its literal meaning, as in *I was in a bad mood, and he kept pestering me, so we **had words***. This sentence was classified as idiomatic since the idiomatic meaning is much more salient than the literal meaning (i.e., *In contrast, the French, for example, **have** two **words** for citizenship* (taken from the BNC)). At the other end of the spectrum, figurative extensions of literal meanings were classified as literal if their overall meaning was judged to be fairly transparent, as in *You turn right when we **hit the road** at the end of this track* (also taken from the BNC).

This dataset was originally intended for use in Cook et al. (2007). The 53 selected expressions were divided into three sets: development, test, and skewed. Skewed contains expressions for which one of the literal or idiomatic meanings is very infrequent, while the expressions in development and test are more balanced across the senses.

The primary annotator annotated all the tokens in each subset of the data. The secondary annotator then annotated the sentences in the development set. The judges then discussed tokens on which they disagreed to achieve a consensus annotation. They also discussed the annotation process at length to improve the quality and consistency of their annotations. The primary judge then re-examined their own annotations for the test set to ensure consistency, while the secondary judge annotated these items. Again, disagreements were discussed to come to consensus annotations as well as to refine the annotation process. Consensus annotations were then determined for the skewed set in the same manner as for the test set.

The items in each of the development, test, and skewed sets, along with their number of usages in each sense, are given in Appendix A. The observed agreement and unweighted kappa score for each set, and over all sets, before the judges discussed their disagreements, is given in Appendix B.

5 File Format

Each line of the file VNC-Tokens describes a particular VNC token and is of the following form:

ANNOTATION VERB_NOUN FILENAME SENTENCE-NUM

ANNOTATION is the consensus annotation for this token and is one of I, L, or Q, corresponding to the idiomatic, literal, and unknown labels, respectively.

VERB_NOUN is the verb and noun which form this VNC. FILENAME and SENTENCE-NUM give the file name and sentence number, respectively, in the BNC XML edition where this token occurs.

VNC-Tokens is sorted by VNC (VERB_NOUN), then file name (FILENAME), and finally sentence number (SENTENCE-NUM).

A Number of Tokens in each Class for each Expression by Set

Set	Expression	#I	#L	#Q	Total
Development	blow trumpet	19	10	11	40
	find foot	48	5	12	65
	get nod	23	3	2	28
	hit road	25	7	17	49
	hit roof	11	7	11	29
	kick heel	31	8	7	46
	lose head	21	19	21	61
	make face	27	14	67	108
	make pile	8	17	3	28
	pull leg	11	40	22	73
	pull plug	45	20	15	80
	pull weight	27	6	17	50
	see star	5	56	9	70
	take heart	61	20	6	87
	Total	362	232	220	814
Test	blow top	23	5	0	28
	blow whistle	27	51	3	81
	cut figure	36	7	1	44
	get sack	43	7	29	79
	get wind	13	16	4	33
	have word	80	11	8	99
	hit wall	7	56	4	67
	hold fire	7	16	8	31
	lose thread	18	2	6	26
	make hay	9	8	11	28
	make hit	5	9	12	26
	make mark	72	13	12	97
	make scene	30	20	15	65
	pull punch	18	4	10	32
	Total	388	225	123	736

Set	Expression	#I	#L	#Q	Total
Skewed	blow smoke	0	52	3	55
	bring luck	24	0	0	24
	catch attention	100	0	0	100
	catch death	22	1	0	23
	catch imagination	45	0	0	45
	get drift	19	0	11	30
	give notice	95	0	6	101
	give sack	15	3	9	27
	have fling	21	0	0	21
	have future	100	0	0	100
	have misfortune	78	0	0	78
	hold fort	22	0	3	25
	hold horse	2	20	4	26
	hold sway	100	0	1	101
	keep tab	54	1	7	62
	kick habit	40	0	3	43
	lay waste	32	0	1	33
	lose cool	28	0	3	31
	lose heart	51	0	1	52
	lose temper	104	0	0	104
	make fortune	100	0	0	100
	move goalpost	13	2	8	23
	set fire	98	0	3	101
	take root	83	15	1	99
	touch nerve	24	0	6	30
	Total	1270	94	70	1434
All	Total	2020	551	413	2984

B Interannotator Agreement on each Set

Set	Observed Agreement (%)	Unweighted Kappa Score
Development	83	0.74
Test	77	0.63
Skewed	88	0.55
All	84	0.69

References

Lou Burnard. 2000. *The British National Corpus Users Reference Guide*. Oxford University Computing Services.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48. Prague, Czech Republic.

Anthony P. Cowie, Ronald Mackin, and Isabel R. McCaig. 1983. *Oxford Dictionary of Current Idiomatic English*, volume 2. Oxford University Press.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 337–344. Trento, Italy.

Maggie Seaton and Alison Macaulay, editors. 2002. *Collins COBUILD Idioms Dictionary*. HarperCollins Publishers, second edition.