The VNC-Tokens Dataset

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson

Univeristy of Toronto
Toronto, Canada
{pcook, afsaneh, suzanne}@cs.toronto.edu

Abstract

Idiomatic expressions formed from a verb and a noun in its direct object position are a productive cross-lingual class of multiword expressions, which can be used both idiomatically and as a literal combination. This paper presents the VNC-Tokens dataset, a resource of almost 3000 English verb–noun combination usages annotated as to whether they are literal or idiomatic. Previous research using this dataset is described, and other studies which could be evaluated more extensively using this resource are identified.

1. Verb–Noun Combinations

Identifying multiword expressions (MWEs) in text is essential for accurately performing natural language processing tasks (Sag et al., 2002). A broad class of MWEs with distinct semantic and syntactic properties is that of idiomatic expressions. A productive process of idiom creation across languages is to combine a high frequency verb and one or more of its arguments. In particular, many such idioms are formed from the combination of a verb and a noun in the direct object position (Cowie et al., 1983; Nunberg et al., 1994; Fellbaum, 2002), e.g., *give the sack*, *make a face*, and *see stars*. Given the richness and productivity of the class of idiomatic verb–noun combinations (VNCs), we choose to focus on these expressions.

It is a commonly held belief that expressions with an idiomatic interpretation are primarily used idiomatically, and that they lose their literal meanings over time. Nonetheless, it is still possible for a potentially-idiomatic combination to be used in a literal sense, as in: *She **made** a **face** on the snowman using a carrot and two buttons*. Contrast the above literal usage with the idiomatic use in: *The little girl **made** a funny **face** at her mother*. Interestingly, in our analysis of 60 VNCs, we found that approximately half of these expressions are attested fairly frequently in their literal sense in the British National Corpus (BNC).[1] Clearly, automatic methods are required for distinguishing between idiomatic and literal usages of such expressions, and indeed there have recently been several studies addressing this issue (Birke and Sarkar, 2006; Katz and Giesbrecht, 2006; Cook et al., 2007).

In order to conduct further research on VNCs at the token level, and to compare the effectiveness of the varying proposed methods for their treatment, an annotated corpus of VNC usages is required. Section 2 describes our dataset, VNC-Tokens, which consists of almost 3000 English sentences, each containing a VNC usage (token) annotated as to whether it is literal or idiomatic. Sections 3, 4, and 5 respectively describe previous research conducted using VNC-Tokens, other work on idioms which could make use of this dataset, and possible ways in which VNC-Tokens could be extended. We summarize the contributions of the VNC-Tokens resource in Section 6.

[1] `http://www.natcorp.ox.ac.uk`

2. The VNC-Tokens Dataset

The following subsections describe the selection of the expressions in VNC-Tokens, how usages of these expressions were found, and the annotation of the tokens.

2.1. Expressions

We began with the dataset used by Fazly and Stevenson (2006), which includes a list of VNCs. We eliminated from this list any expression whose frequency in the BNC 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 gave 60 candidate expressions.

Two expert judges, both native English-speaking authors of this paper, examined the candidate expressions and eliminated 7 of them. The idiomatic meaning of *blow one's (own) horn*, *get the bird*, and *pull one's hair (out)* were not familiar to one judge, and therefore could not be annotated with confidence.[2] For the expressions *catch one's breath*, *cut one's losses*, and *push one's luck* the annotators agreed that a literal interpretation was not possible, while they judged that *give a lift* does not have a clear idiomatic meaning. This gave a final set of 53 expressions.

2.2. Sentence Extraction

To identify usages of a VNC in text, we first parsed the BNC (Collins, 1999), and then looked for sentences containing the component verb and noun from one of our 53 VNCs in a direct object relation. For each expression, 100 sentences containing its usage were randomly selected, and for expressions with less than 100 usages, we extracted all sentences.

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.

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

[2] *Pull one's hair (out)* is a verb-particle construction. Although such expressions may be, to varying degrees, idiomatic, they were not the focus of this study.

for any originally selected sentence. The final dataset consists of 2984 VNC tokens, of which 2920 are unique occurrences.

2.3. Token Annotation

Each instance of the 53 chosen expressions was annotated by the two judges as one of literal, idiomatic, or unknown. During annotation the judges were presented with the single sentence containing the VNC usage; sentences in the surrounding context were not included. If the judge was unable to determine the class of a token based on the sentence in which it occurs, the judge chose the unknown label.

The idiomaticity of an expression is not binary. Expressions may be more or less idiomatic, falling on a continuum ranging from completely literal expressions, i.e., *get the present*, to semantically opaque idioms, i.e., *get the sack* (which has the idiomatic interpretation of losing one's employment). For usages falling towards the middle of this continuum, the human annotators were instructed to choose the most appropriate label according to their judgement, as opposed to using the unknown label.

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. These preliminary annotations were used to divide the expressions into the three sets. 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.

A number of issues arose during reconciliation of disagreements that are worth noting, particularly with respect to usages that fall somewhat towards the middle of the literal–idiomatic continuum. For example, there are idiomatic usages of the expression *have word* that have a meaning that is somewhat related to its literal meaning, as in: *At the moment they only had the word of Nicola's husband for what had happened*.[3] The final annotation for this sentence was idiomatic since the idiomatic meaning was judged to be much more salient than the literal meaning, as in: *In contrast, the French, for example, have two words for citizenship*. Further towards the literal end of the continuum are certain usages of expressions such as *hit the road*. This expression may be used in a clear literal sense, as in: *Gina Coulstock, 18, stumbled, fell heavily and was knocked out when she hit the road*. It may also be used with the idiomatic meaning of departure, as in: *The marchers had hit the road before 0500 hours, and by midday they were limping back to Heumensoord*. However, this expression may also be used in a more intermediate sense, as in: *You turn right when we hit the road at the end of this track*. Such usages of *hit the road*, and similar usages of other expressions, were judged to be figurative extensions of literal meanings, and were therefore classified as literal.

The items in each of the development, test, and skewed sets, along with their number of usages in each sense, are given in Table 1. The observed agreement and unweighted Kappa score for each set, and over all sets, before the judges discussed their disagreements, is given in Table 2.[4]

3. Previous Research Using VNC-Tokens

The only research to date which has made use of VNC-Tokens is that of Cook et al. (2007). They perform an extensive token-based study of VNCs using an earlier version of the development and test subsets of VNC-Tokens for development and evaluation of their methods. Their study is based on the observation that the idiomatic meaning of a VNC tends to be expressed in a small number of preferred lexico-syntactic patterns, referred to as canonical forms (Riehemann, 2001). For example, while both the idiomatic and literal interpretations are available for the phrase *kicked the bucket*, only the literal meaning is possible for *kicked a bucket* and *kicked the buckets*.

Cook et al. hypothesize that idiomatic usages of a VNC will usually occur in one of that expression's canonical forms, while the literal meaning will be expressed in a wider variety of forms. Drawing on established unsupervised methods for determining the canonical forms of a VNC (Fazly and Stevenson, 2006), Cook et al. propose three unsupervised methods for distinguishing literal and idiomatic VNC usages that incorporate their hypothesis.

Their CFORM method relies solely on information about canonical forms, and simply classifies a usage of an expression as idiomatic if it occurs in one of that expression's canonical forms, and as literal otherwise. Their other two methods, $\text{DIFF}_{\text{I-CF, L-NCF}}$ and $\text{DIFF}_{\text{I-CF, L-COMP}}$, incorporate lexical co-occurrence information along with the syntactic information provided by canonical forms. In these methods, three co-occurrence vectors approximating each of the meaning of the target token to be classified, the literal meaning of the expression, and that expression's idiomatic meaning are formed. The vector representing the target is then compared using cosine to those for the literal and idiomatic meanings, and the target is assigned the

[3] All examples in this subsection are taken from the BNC and occur in VNC-Tokens.

[4] We expected the inter-annotator agreement scores would have been at least as high for the test subset as for the development subset, due to the discussion that took place after annotating the development expressions. However, as Table 2 shows, this is not so. The observed agreement for each development expression is above 80%, while for three test expressions this is not the case. For the expressions *have word* and *hold fire* the judges systematically disagreed on the label for one particular sense of each of these expressions. For the expression *make hit*, the low agreement may have been a result of the proportionally large number of questionable usages (see Table 1). Eliminating these three expressions gives an observed agreement and unweighted Kappa score of 89% and 0.83, respectively, for the remaining test expressions.

Subset	Expression	I	L	Q	Total
Dev.	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
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

Table 1: Number of tokens annotated idiomatic (I), literal (L), and unknown (Q), as well as the total number of tokens (Total), for each expression, grouped by subset of VNC-Tokens.

Set	Observed Agreement (%)	Kappa
Development	89	0.83
Test	78	0.65
Skewed	93	0.67
All	88	0.76

Table 2: Percent observed agreement and unweighted Kappa score for each set.

meaning of the more similar vector. In both DIFF methods, the co-occurrence vector for the idiomatic meaning is created by considering the words in a 5-word window on either side of all canonical form usages of that expression. In this way they obtain an unsupervised, but noisy, estimate of the idiomatic meaning. The two DIFF methods estimate the literal meaning of an expression in differing ways. $\text{DIFF}_{\text{I-CF, L-NCF}}$ approximates the literal meaning using non-canonical form usages in a similar manner to the estimate of the idiomatic meaning. $\text{DIFF}_{\text{I-CF, L-COMP}}$ assumes that a literal VNC usage is compositional, and averages the co-occurrence vectors for each of the component verb and noun in a VNC to estimate its literal meaning.

Cook et al. compare their methods to a baseline which classifies every token as idiomatic. They also compare against a slightly modified version of the supervised method proposed by Katz and Giesbrecht (2006), which classifies a token according to the gold-standard labels of the k nearest tokens according to cosine distance between their co-occurrence vectors. Cook et al. find all three of their unsupervised methods to outperform the baseline of 62% accuracy, with CFORM achieving the highest accuracy of 72%. The CFORM method performs as well as the supervised method with k set to 1; however, using the 5-nearest neighbours in a supervised setting achieves the best performance of 76% accuracy.

Fazly et al. (2008) extend the work of Cook et al. in several ways. Fazly et al. represent the context of a token as the full set of words from the sentence in which it occurs, in an effort to overcome data sparseness problems reported by Cook et al. Consequently, they compare tokens using a set-based similarity measure, Jaccard index. Fazly et al. examine the performance of their methods on all three subset of VNC-Tokens, and present a detailed analysis of their results. They too find CFORM to have the highest unsupervised performance on the test subset. However, their results on the previously-unused skewed subset indicate that their unsupervised method using context outperforms CFORM on expressions that are predominantly used idiomatically.

4. Related Work on Idioms

Two approaches to distinguishing between literal and non-literal tokens have recently been proposed that could be evaluated more extensively using the VNC-Tokens dataset. Katz and Giesbrecht (2006) perform a token-based study of the German expression *ins Wasser fallen* which when used literally means *to fall into water*, but which also has an idiomatic interpretation of *to fail to happen*. They propose a supervised method to distinguish between literal and idiomatic usages of this expression, which is quite similar to,

and in fact was the motivation for, the supervised 1-nearest neighbour method considered by Cook et al. (2007). The main difference between these two approaches is that Katz and Giesbrecht employ singular value decomposition to reduce the dimensionality of the co-occurrence vectors. They evaluate their method on 67 instances of *ins Wasser fallen* found in a corpus of text from a German newspaper, and report an accuracy of 72% on this task which has a baseline of 58%. One of the main shortcomings of this study is that it only presents results for one expression. The VNC-Tokens dataset addresses this by allowing for a more extensive evaluation, although not on German idioms.

Birke and Sarkar (2006) propose a minimally-supervised method for distinguishing between literal and non-literal usages of verbs. Their algorithm relies on seed sets of literal and non-literal usages of verbs that are automatically obtained from readily-available lexical resources. The class of a target verb token is then determined using the similarity between the context of that token and each of the seed sets. Although the annotations in VNC-Tokens are for the combination of a verb and its direct object, it may still be an appropriate resource for evaluating this algorithm. For many expressions in VNC-Tokens, such as *blow the whistle* and *move the goalposts*, the verb is used in a non-literal sense when the VNC is idiomatic, and in a literal sense when the VNC is literal. For other expressions, such as *get the nod* and *make a pile*, this may not be the case depending on the definitions of literal and idiomatic employed—the verb may be contributing a literal meaning even when the VNC it forms with its direct object is idiomatic. Nevertheless, some of the expressions in VNC-Tokens would be appropriate, and would allow for a more extensive evaluation of Birke and Sarkar's algorithm.

Hashimoto et al. (2006) build an unsupervised classifier that exploits manually-encoded lexical knowledge to distinguish between literal and non-literal usages of Japanese idioms, which they evaluate on a relatively small dataset of 309 tokens. However, since their classifier draws on specific properties of Japanese idioms, it is not clear that a more extensive evaluation of their method could be conducted using the English expressions in VNC-Tokens.

5. Future Extensions to VNC-Tokens

While annotating the items in VNC-Tokens, the human judges had access to only the sentence in which a VNC usage occurs (see Section 2.3). This limitation of the annotation process resulted in 413 tokens being assigned the unknown label. Had the annotators had access to more of the surrounding context of each token, far fewer items would have been labelled unknown. As future work, we intend to re-visit those tokens annotated as unknown, and attempt to label them as idiomatic or literal by examining a broader context of their usage.

VNC-Tokens currently consists of at most 100 usages of each of 53 expressions (see Section 2.2). For expressions which occur more than 100 times in the BNC, 100 tokens were randomly selected. VNC-Tokens could be expanded by including additional tokens for these expressions. This would require human effort to annotate the new tokens, but would not be an arduous task as the judges are already familiar with the expressions and the issues involved in their annotation. To expand VNC-Tokens by adding new expressions would be a substantially larger effort. This would require re-running the extraction software and then having human judges annotate the new tokens. Annotating instances of a novel expression would likely be more difficult than annotating new instances of an expression already in VNC-Tokens, as the specific properties of the newly-added expressions may give rise to new annotation issues.

6. Summary

This paper describes the VNC-Tokens dataset, a resource which facilitates research on potentially-idiomatic verb–noun combinations, a productive and common cross-lingual class of MWE. We have described one study which used VNC-Tokens for evaluation, and have shown how two similar studies could also be evaluated more extensively using this resource. Finally, we have identified several ways in which this resource could be expanded in the future.

7. References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-2006*, pages 329–336, Trento, Italy.

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 EACL-2006*, pages 337–344, Trento, Italy.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2008. Unsupervised type and token identification of idiomatic expressions. Submitted to *Computational Linguistics*.

Christiane Fellbaum. 2002. VP idioms in the lexicon: Topics for research using a very large corpus. In S. Busemann, ed., *Proc. of the KONVENS 2002 Conference*, Saarbruecken, Germany.

Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. 2006. Detecting Japanese idioms with a linguistically rich dictionary. *Language Resources and Evaluation*, 40(3–4):243–252.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/Coling Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Susanne Riehemann. 2001. *A Constructional Approach to Idioms and Word Formation*. Ph.D. thesis, Stanford.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico.

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