In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp. 1127-30.

A Multilingual Database of Idioms

Aline Villavicencio[*], **Timothy Baldwin**[†], **Benjamin Waldron**[*]

[*] University of Cambridge Computer Laboratory,
William Gates Building, JJ Thomson Avenue,
Cambridge, CB3 0FD, UK
{av208,bmw20}@cl.cam.ac.uk
[†]CSLI, Ventura Hall, Stanford University
Stanford, CA 94305-4115, USA
tbaldwin@csli.stanford.edu

Abstract

This paper presents a possible architecture for a multilingual database of idioms. We discuss the challenges that idioms present to the creation of such a database and propose a possible encoding that maximises the amount of information that can be stored for different languages. Such a resource provides important information for linguistic, computational linguistic and psycholinguistic use, and allows for the comparison of different phenomena in different languages. This can provide the basis for a better understanding of regularities in idioms across languages.

1. Introduction

This work is concerned with enabling the creation of a multilingual database of idioms. Idioms are often defined as *a group of words which have a different meaning when used together from the one it would have if the meaning of each word were taken individually* (Collins, 2000).[1] They comprise expressions like *spill the beans*, *kick the bucket* and *pull strings*, that are usually employed in everyday language to precisely express ideas and concepts that cannot be compressed into a single word. Even though some idioms are fixed, and do not present internal variation, such as *ad hoc*, there is also a large proportion of idioms that allow different degrees of internal variability, and with a variable number of elements. For example, the idiom *spill the beans* allows internal modification (*spill mountains of beans*), passivization (*The beans were spilled on the latest edition of the report*), topicalization (*The beans, the opposition spilled*), and so on.

As we can see, idioms are a highly heterogeneous kind of multiword expression, ranging from (semi-)fixed cases (e.g. *kick the bucket*) which only allow morphological inflection, to more flexible ones (e.g. *spill the beans*) which can undergo different types of syntactic variation and modification (Nunberg et al., 1994). Moreover, for the later case, the type of syntactic variation that these idioms allow is highly unpredictable (Riehemann, 2001). Even if these works focus their discussion on idioms in English, the same phenomena can also be found in idioms in other languages.

Such variation tends to be a challenge for their successful (computational) linguistic treatment (Sag et al., 2002). In linguistics, for example, they have been often used as evidence for or against the properties of grammatical theories (e.g. *must "syntactic theory" include transformational operations?* from Nunberg et al. (1994)). In computational linguistics, for applications such as machine translation, appropriate understanding/treatment of idioms is necessary for these systems to be able to deal with natural languages, and avoid the generation of unnatural or nonsensical sentences in the target language. There are even cases where a pair of corresponding idioms in two different languages may share the same properties (e.g. *the other side of the coin* in English and its literal translation in Portuguese *o outro lado da moeda*, which is also a noun phrase idiom) But exactly how much variation do these idioms have? What is the proportion of idioms that are fixed in a given language? And what proportion have equivalents in other languages?

Having access to a multilingual database of cases and being able to analyse them can give us some insight into the nature of idioms, and into what is required of a proper treatment of idioms crosslingually. In this work we propose an encoding that supports the collection of idioms in several languages, and the mapping of equivalent parts.

2. Idioms across Languages

Idioms are commonly thought of as metaphors that have became fixed or fossilized over time. While in some cases the metaphor is transparent and can be easily understood even by non-native speakers (e.g. *kill two birds with one stone* as *achieve two things at the same time*), in other cases the metaphor is opaque and if the idiom is not known by the hearer, it can lead to misinterpretations (e.g. *kick the bucket* as *die*).

Some of these metaphors can be found in idioms across languages, and in some cases, in very similar idioms. For instance, one idiom that can be found in both English and Portuguese that shows full lexical, syntactic and semantic correspondence is *in the red*, which is *no vermelho* in Portuguese, where *no* is the contraction of *in* + *the* and *vermelho* means *red*, and both idioms are prepositional phrases (PPs) and have the same meaning. However, there is a large range of variation to be found in idiom pairs across languages, and some idioms do not have such a direct mapping, and may differ in one or more ways and/or may allow different forms of modification/variation. For example,

[1] However, as Nunberg et al. (1994) remark *attempts to provide categorical, single-criterion definitions of idioms are always to some degree misleading and after the fact*.

some idiom pairs are syntactically and semantically but not lexically equivalent. One example is *in the black* and its Portuguese counterpart *no azul* (*in the blue*), where both are PP idioms and the only difference is in the choice of colour (*blue* instead of *black*), or alternatively *bring the curtain down on* and its counterpart *botar um ponto final em* (*put the final dot in*) that are both verbal constructions. There are also idioms that are semantically equivalent, but realised using different constructions across languages. For example, *in a corner* and *encurralado* (meaning *cornered*) are semantically equivalent but realised by different constructions – a PP in English and an adjective in Portuguese). Finally, some idioms have multiple idiomatic equivalents in a second language, while others have none, and this information is also of importance (see Tanaka and Baldwin (2003) for a discussion of English and Japanese compound nouns in the context of a machine translation task).

The challenge is then to define a database design which is capable of encoding all the variation found in these phenomena as well as the correspondences between them in a common format. We propose a database design that can be used for such a task, allowing the maximum amount of information to be stored about an idiom and its counterparts in different languages.

3. A Possible Architecture

A typical session starts with the user entering some identification information, specifying his/her native language and then choosing a source language to be mapped to the target language (by default the user's native language). All idioms from the source language are then made available to the user, who can browse through them, and enter the idiomatic equivalent(s) in the target language. For each idiom, the user is presented with an explanation of the meaning of the idiom and an example (both in English). The user is then asked to provide information about its syntactic variation (e.g. *Can the idiom be topicalised?*, *Does it allow internal modification?*, etc), and about its mapping to the source language (if it exists). As discussed in Section 2., for a particular language pair, there may be considerable variation in the realisation of equivalent idioms. In order to capture this variation, we adopt the following procedure:

1. If the idiom in the target language is lexically, syntactically and semantically equivalent to the idiom in the source language (e.g. *in the red* and *no vermelho*), the user is asked to provide a word-to-word mapping of the idiom;

2. Otherwise if they are syntactically and semantically equivalent, but not lexically (e.g. *in the black* and *no azul*), the user is asked to provide the mapping between the corresponding words, and for those that are lexically distinct, a translation to the source language;

3. Otherwise if they are only semantically equivalent, the user is asked to input each word of the idiom and its translation to the source language.

For each of these cases, the position of the word in the idiom is also recorded, to account for variations in word order. One example is *new blood* in English, where the adjective precedes the noun, and its equivalent in Portuguese *sangue novo* (*blood new*), where the adjective follows the noun.

If more than one equivalent exists, then the same process applies to each of the equivalents. After that, or if there are no equivalents, the next idiom is displayed and the user goes through the same process.

4. Test Data

In order to test the design, the database currently contains a sample of 100 high-frequency[2] English idioms extracted randomly from the Collins Cobuild Dictionary of Idioms (Villavicencio and Copestake, 2002). This is used as the starting point (source language seed) to collect translation-equivalent idioms in other languages. Initially, it is this mapping between English and other languages that is being tested, but the goal is to extend the database to support mappings between idioms in any two languages. This database can be accessed locally and also through a web interface, allowing users in different locations to browse the database and provide information about idioms in their native language.

5. Web Interface

The first step in the annotation process is to stipulate the target language, and optionally select the English idiom index number from which to start the annotation. At the present time, language selection is string-based and not normalized in any way, to avoid restricting the scope of annotation to any closed set of languages. The interface additionally has a cookie-based facility to identify the annotator for data maintenance purposes and also consistency in multi-session annotations.

Having chosen the language, the annotator works through each of the 100 English idioms in turn, supplying equivalent idiom(s) in the target language. For each target language idiom, the annotator is asked to give a monolingual judgement on its internal modifiability, and an evaluation of its lexical and syntactic equivalence to the source language idiom. In Figure 1, we provide a screen shot of the annotations for *no azul* as a translation for *in the black*.

The interface next presents the annotator with an alignment window to indicate lexical correspondences between the two idioms. In the case of target language idioms which are lexically equivalent to the source language idiom, this consists of matching up each target language word with its corresponding source language word(s), and the interface simply presents the annotator with a list of source language word indices with which to perform the alignment (see Figure 2 for the case of *new blood* and *sangue novo*); on submitting the alignment, the system then checks that the alignment is maximal—i.e. that all words map onto one or more words in the opposing language—and issues a warn-

[2] Flagged by the editors of the Collins Cobuild Dictionary of Idioms as occurring "at least once in every 2 million words of (their) corpus".

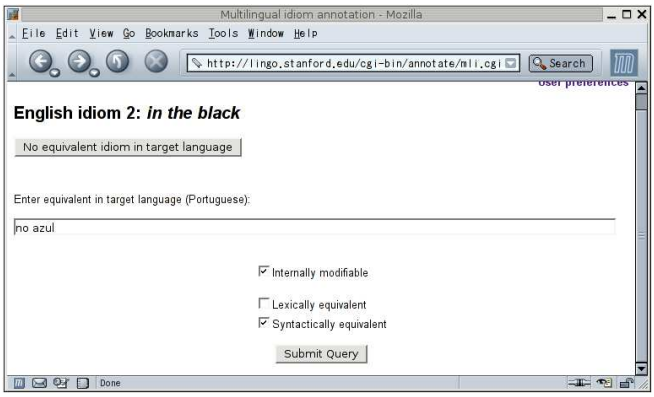

Figure 1: Providing a translation and basic idiom properties

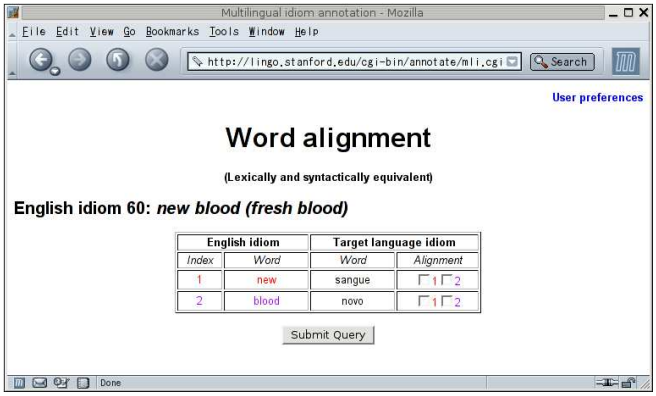

Figure 2: Word alignment (1)

ing in the case that any non-connected words are found.[3] If the idioms are not lexically equivalent, on the other hand, an additional column is supplied for translation glosses of non-aligned words in the target language idiom (see Figure 3 for the case of *in the black* and *no azul*). Note that it is possible for there to be partial lexical correspondence (as seen for *no* with *in the*), and for this reason, we provide the word alignment facility as for lexically-equivalent idioms. We do, however, check for the existence of both aligned and glossed target language words (indicating non-lexical equivalence). In the case that these conditions are not met, a warning message is issued. At present, we do not attempt to make any further classification of the nature of mismatch for idioms that are not syntactically equivalent, nor do we attempt to classify the construction type of syntactically-equivalent idioms.

After annotating each idiom pair, the annotator is given the option of adding an additional translation for the source language idiom, or alternatively proceeding to the next idiom. Additionally, the annotator can flag a source language idiom as having no target language equivalent (see Figure 1).

[3]Note that it is arguably possible for a lexically-equivalent idiom to not strictly align between languages. E.g., the Japanese translation of *in the black* is *kuroji*, which is lexically equivalent to *black*. It is possible to argue that *in* and *the* are function words and that the two idioms are thus lexically equivalent in terms of their content words.

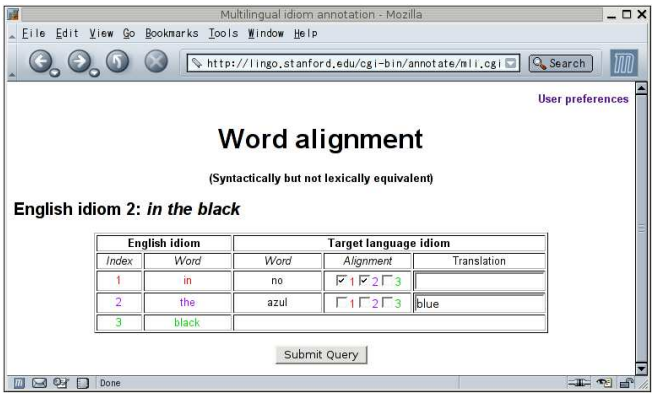

Figure 3: Word alignment (2)

The web interface is publicly accessible at `lingo.stanford.edu/cgi-bin/annotate/mli.cgi` in the form of a CGI script.

6. Lexical Database

The work reported in the paper relates to a larger project to develop a lexical database (Copestake et al., 2004). This lexical database is primarily for use within a grammar development environment. It provides a resource for the association of stems with grammatical, that is syntactic and semantic, information. In addition to grammatical information entries are associated with bookkeeping information (such as language and dialect) and other information. For example by linking to a semantic database containing detailed fully-expanded lexical semantics we can provide an efficient index for generation, or a data source for purposes. The existence of such a base lexical component within a grammar development environment provides a number of advantages over alternative approaches, including ease of maintainance, efficiency, and the benefits gained by utilising bookkeeping information and data from secondary sources.

By taking advantage of database functionality we can link idioms in the database of idioms discussed in the this paper with idiomatic entries in the lexical database.

As well as basic simplex lexical entries such as *bombard* the lexical database supports multiword expressions. These we may divide into two classes: those which allow for internal variation, and those which do not.

Consider firstly those idioms which allow for internal variation; for example *spill the beans* and variations thereof. In the lexical database we associate each such idiom with a template. This template specifies the necessary syntactic and semantic components of the idiom. For example *spill the beans* and *rock the boat* are syntactically composed of a verb and associated object; in the first case we require that the verb be (an idiomatic form of) the verb *spill*; in the second case, we require (an idiomatic form of) the verb *rock*; and so on. We also require that the simplex lexicon be augmented to include entries for these idiomatic word forms (these idiomatic simplex forms are generated by overriding certain grammatical information in the non-

idiomatic basic simplex entry; e.g. the idiomatic *spill* differs only from the non-idiomatic *spill* in specifying an idiomatic semantics). For a discussion of a specific approach to encoding such idioms within a grammar see (Copestake et al., 2002).

Those idioms which do not allow for internal variation (*ad hoc* being an example) may trivially be treated in the same manner as basic simplex entries.

The two classes of idiom outlined above are stored within distinct tables in the lexical database, each idiom being indexed by a unique identifier. Using the identifiers of the idioms in the two data sources, entries in the database of idioms are linked to the grammatical and other information contained in the lexical database, and via the lexical database to further potentially useful sources of information.

7. Discussion

The multilingual idiom database provides important information for linguistic, computational linguistic and psycholinguistic use, and allows for the comparison of different phenomena in different languages. For instance, it may be the case that families of languages have very similar idiom equivalents and the same patterns of modification within them, and this can provide the basis for a better understanding of regularities in idioms across languages. Orthogonally, the semantic mappings may provide evidence supporting the claim that languages base idioms on common metaphors (Neumann, 1999). Moreover, the possibility of analyzing the different degrees of flexibility allowed by different languages for the same idiom is also valuable (e.g. in analysing *idiom avoidance* in bilinguals (Laufer, 2000)), and the presence (or absence) of certain idioms in different languages may also be of interest (e.g. for historical studies). Finally, such a database may contain data from different speakers of the same language, and provide grounds for investigation of the variation in individuals' intuitions into, e.g. modification effects and semantic alignment.

8. Conclusion

This paper has outlined the architecture of a multilingual database of idioms. The database is set up to capture basic monolingual and crosslinguistic properties of idioms in a uniform fashion, via an English "interlingua" which is additionally linked to an implemented grammar of English.

Our primary short-term objective is to populate the database with as many languages as possible and proceed with a crosslinguistic study of idioms. We also hope to expand the scope of the annotation process to analyze the syntactic correspondences between idioms in different languages.

9. Acknowledgements

This research was supported in part by the Research Collaboration between NTT Communication Science Research Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University, and also the National Science Foundation under Grant No. BCS-0094638.

This document was generated partly in the context of the DeepThought project, funded under the Thematic Programme User-friendly Information Society of the 5th Framework Programme of the European Community (Contract No. IST-2001-37836).

10. References

Collins. 2000. *Collins Cobuild Dictionary of Idioms*. Harper Collins Publishers.

Ann Copestake, Fabre Lambeau, Aline Villavicencio, Francis Bond, Ivan Sag Timothy Baldwin, and Dan Flickinger. 2002. Multiword Expressions: Linguistic Precision and Reusability. In *In Proceedings of the 4th International Conference On Language Resources and Evaluation (LREC-2002)*, Las Palmas, Canary Islands.

Ann Copestake, Fabre Lambeau, Benjamin Waldron, Francis Bond, Dan Flickinger, and Stephan Oepen. 2004. A Lexicon Module for a Grammar Development Environment. In *4th International Conference On Language Resources and Evaluation, LREC-2004*, Lisbon, Portugal.

Batia Laufer. 2000. Avoidance of idioms in a second language: The effect of L1-L2 degree of similarity. *Studia Linguistica*, 54(2):186–96.

Christoph Neumann. 1999. *Formal languages for fuzzy subjects-modality, metaphor and machine translation*. Ph.D. thesis, Tokyo Institute of Technology Dissertation.

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

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

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation: A feasibility study on shallow processing. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 17–24, Sapporo.

Aline Villavicencio and Ann Copestake. 2002. The nature of idioms. *LinGO Working Paper No. 2002-04*.