# Integrating Verbal Idioms into an NLP System

Jorge Baptista[1,3], Nuno Mamede[2,3], and Ilia Markov[1,3]

[1] Universidade do Algarve/FCHS and CECL,
Campus de Gambelas, 8005-139 Faro, Portugal
jbaptis@ualg.pt

[2] Instituto Superior Técnico, Universidade de Lisboa,
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
Nuno.Mamede@ist.utl.pt

[3] INESC-ID Lisboa/L2F – Spoken Language Lab,
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
{jbaptis,Nuno.Mamede,Ilia.Markov}@l2f.inesc-id.pt

**Abstract.** This paper describes the integration of verbal idioms into an Natural Language Processing (NLP) system, adopting a construction approach, which is based on the prior parsing stage, so that these Multi-Word Expressions (MWE) can be taken into account in subsequent tasks, such as semantic role labeling or whole-part relation extraction. The paper focuses on body-part nouns, which are often part of many verbal idioms, and uses a manually annotated corpus to evaluate its parsing strategy. Results showed a precision of 0.92, 0.83 recall, 0.87 f-measure and an accuracy 0.99.

**Keywords:** Verbal Idioms, Multi-Word Expressions, Body-part Nouns, Lexicon-Grammar, Natural Language Processing, Parsing, European Portuguese.

## 1 Introduction

Verbal idioms are idiomatic (semantically non-compositional) expressions consisting of a verb and at least one constraint argument slot, for which the overall meaning cannot be calculated from the meaning that the individual elements of the expression would present when used independently, in other contexts [10,11]: *O Pedro perdeu a cabeça* (lit: Pedro lost the=his head) 'Pedro became furious' (or 'Pedro lost his mind'). In this paper, we address the main issues raised in the process of integrating the lexicon-grammar of European Portuguese verbal idioms [2,3] into a fully-fledged natural language processing system, STRING[1] [13]. Our purpose is to highlight the detailed level of description required to identify this type of linguistic meaning units in texts, while maintaining these resources updated.

[1] https://string.l2f.inesc-id.pt/ [last access: 10/05/2014].

J. Baptista et al. (Eds.): PROPOR 2014, LNAI 8775, pp. 250–255, 2014.

## 2 Related Work

Work on multiword expressions (MWE) has drawn the attention of computational linguists for quite a long time [17]. Compound nouns, adverbs and other multiword lexical units pose specific problems to their automatic lexical acquisition and identification, but can be parsed using a *words-with-spaces* approach. For a recent comparison on different techniques for automatic multiword identification, see [15]. A *construction approach* [6] seems more appropriate to represent verbal idioms, provided lexical resources are available.

Extensive lists of verbal idioms, particularly the most frequent ones, have been systematically collected for Portuguese, both the European [2] and the Brazilian [18] varieties, along with their main distributional, syntactic and transformational properties, under the Lexicon-Grammar methodological and theoretical framework [10,11]. To our knowledge, so far these resources have not been integrated yet in any Portuguese NLP system, so it is difficult to ascertain the issues that may rise from the interaction of the different modules.

## 3 Integration of Verbal Idioms in STRING

The STRING system uses the XIP parser (Xerox Incremental Parser) [1] to segment sentences into chunks and extract dependency relations among chunks' heads. Considering that most idioms have a "normal" syntactic structure, which follows the ordinary word combinatory rules of the general grammar, STRING's strategy consists in parsing them *first* as ordinary sentences and *only then* to identify specific word combinations, whose meaning should not be calculated in a compositional way. This corresponds to the *construction approach* originally proposed by [6], and this is based on the results from the previous parsing stages, including the main syntactic dependencies such as `SUBJ`[ect], `MOD`[ifier] or direct object (`CDIR`), as well as auxiliary dependencies like `PREPD`, relating prepositions and the `PP` heads, or `DETD`, linking the determiners to the `NP` heads. It also involves using either surface forms or lemmas, or even restrictions in the morphological attributes of any given lemma. A large set of rules were semiautomatically built from the available lexicon-grammar of verbal idioms [2]. The idiomatic word combinations are identified by a new dependency, `FIXED`, which takes as arguments the verb and the frozen elements of the idiomatic expression (the number of arguments depends on the type of idiom involved). Figure 1 below illustrates the chunking tree and the relevant dependencies extracted for the sentence *O Pedro perdeu a cabeça* 'Pedro lost his mind'.

The identification of the idiom uses the previously calculated dependencies, namely the direct object (`CDIR`) and the main verb, and is carried out by the following rule:

```
IF (VDOMAIN(?,#2[lemma:perder]) & CDIR[post](#2,#3[surface:cabeça])) FIXED(#2,#3)
```

This rule captures any form of the lemma of the verb *perder* 'lose' (inclunding any compound tenses) [4] and the surface form of the direct object (obligatorily

```
               TOP
     +----------+---------+
     |          |         |        MAIN(perdeu)
    NP         VF        NP        DETD(Pedro,O)
 +-------+      +     +-------+    DETD(cabeça,a)
 |       |      |     |       |    SUBJ_PRE(perdeu,Pedro)
ART   NOUN    VERB   ART    NOUN   CDIR_POST(perdeu,cabeça)
 +     +-       +    +-       +    FIXED(perdeu,cabeça)
 |      |       |    |        |    NE_PEOPLE_INDIVIDUAL(Pedro)
 O   Pedro   perdeu  a     cabeça  0>TOP{NP{O Pedro} VF{perdeu} NP{a cabeça}}
```

**Fig. 1.** Extraction of `FIXED` dependency for the sentence *O Pedro perdeu a cabeça* ‘Pedro lost his mind’

after the verb) *cabeça* ‘head’. Around 2,400 rules were semiautomatically build for 10 formal classes of verbal idioms [2]. A list of simple, manually-built examples (one for each idiom) provided with those classes was used to test the rules during the development stage.

For lack of space, we can not detail much the challenges that had to be met during the implementation of the rules, nor the solutions we provided; therefore, only the briefest overview is provided here.

**Compounds.** Since the idioms are being processed at a very late stage of parsing, some compound words have already been identified and the resulting parse is inadequate. For example, in: *O dinheiro subiu à cabeça do Pedro* ‘The money went to Pedro’s head’ the compound preposition *à cabeça de* ‘at the head of’ has been tokenized, at an earlier stage, so the rule that would identify the verbal idiom must take this preposition into account. Notice that, for the semantic representation of the sentence, since the idiom is to be considered semantically non-compositional, an overall meaning should be attributed to the sentence, and the original meaning of the compound preposition *à cabeça de* ‘at the head of’ is to be discarded, in much the same way as the (potential) meaning of the verb: `FIXED(subiu,à cabeça de)`.

**Intrinsically Reflexive Constructions.** In many idioms, the verb shows an intrinsically reflexive construction: *O Pedro atirou/mandou/amandou-se ao ar* (lit: Pedro threw himself to the air) ‘Pedro went mad/furious’. A dependency, named `CLITIC`, has already been extracted between the verb and the clitic pronoun. This dependency is also extracted even if the pronoun, under certain syntactic conditions, is moved to the front of the verb (proclisis), as is *O Pedro até se atirou ao ar* (lit: Pedro even himself threw to the air). Therefore, even sentences where this fronting take place are captured by the same rule. Notice that in this intrinsically reflexive construction, the reflexive pronoun should correspond to an object `NP`, but for the the purpose of the semantic representation of the idiom, one can ignore it altogether: `FIXED(atirar,a,ar)`.

**Obligatory Negation.** About 5% of all the Portuguese verbal idioms collected so far [7] involve an obligatory negation: *O Pedro não brinca em serviço* (lit: Pedro does not play in his job) ‘Pedro is always very serious/competent in his job’.

Even if in most cases negation is carried out by negation adverb *não* 'no/not', other negation adverbs (*nunca*, *jamais* 'never'; *nem* 'nor') can also be used. All these cases are captured by a special feature `NEG` on the `MOD`[ifier] dependency that links these adverbs to the verb. Notice that, as the negation is considered as an intrinsic component of idiom, thus it is captured as a feature of the fixed expression: `FIXED_NEG(brinca,em,serviço)`.

## 4 Evaluation

We have framed the evaluation of the idioms identification module in the task of part-whole relation extraction. We took advantage of an already existing corpus with annotated frozen expressions, in this case, with idioms involving *Nbp*. This corpus consists of a random stratified sample of 1,000 sentences, selected from a large set of around 17,000 sentences extracted from the $1^{st}$ fragment (6,25 million words) of the CETEMPúblico corpus [16] using a small dictionary of around 300 *Nbp*.

The number of sentences with each *Nbp* in the sample is proportional to the number of its occurrences in the CETEMPúblico fragment. This sample had been previously annotated by 4 different native speakers, under the scope of another work on the extraction of whole-part relations involving body-part nouns and human entities [14]. The output sentences were divided into 4 subsets of 225 sentences each. Each subset was then given to a different annotator, and a common set of 100 sentences was added to each subset in order to assess inter-annotator agreement. From the 100 sentences that were annotated by all the participants in this process, we calculated the Average Pairwise Percent Agreement (0.85), the Fleiss' Kappa [8] (0.625; observed agreement 0.85/expected agreement 0.601), and the Cohen's Kappa coefficient of inter-annotator agreement [5] (0.629) using ReCal3: Reliability Calculator [2], for 3 or more annotators. According to Landis and Koch [12] these figures correspond to the lower bound of the "substantial" agreement; however, according to Fleiss [9], these results correspond to an inter-annotator agreement halfway between "fair" and "good". In view of these results, we can assume as a reasonable expectation that the remaining, independent and non-overlapping annotation of the corpus by the four annotators is sufficiently consistent, and will use it for the evaluation of the system output.

There are about 400 frozen expressions involving *Nbp* in the lexicon-grammar of European Portuguese verbal idioms, while the corpus features 40 types of these (43 instances). Table 1 shows the results from this experiment. While the observation on this sample can not be extended to the general lexicon, it may suggest the level of adequacy of the methods and resources here used.

We now describe the main errors still to be addressed by the system. Some of the false-positive cases result from the structural ambiguity between the idiom

[2] `http://dfreelon.org/utils/recalfront/recal3/` [last access: 08.02.2014].

[3] TP: true-positives; TN: true-negatives; FP: false-positives; FN: false-negatives.

**Table 1.** Results

| Number of sentences | TP | TN | FP | FN [3] | Precision | Recall | F-measure | Accuracy |
|---|---|---|---|---|---|---|---|---|
| 1,000 | 33 | 957 | 3 | 7 | 0.92 | 0.83 | 0.87 | 0.99 |

and another construction. For example, in the next sentence, a support verb construction of the predicative noun *bofetada* ‘slap’ involves the same verb *dar* ‘give’ and the *Nbp cara* ‘face’: *Para mim dizer bem de Castro é o mesmo que dar uma bofetada na cara do meu pai* ‘For me, to say nice things about Castro is the same as giving a slap in my father’s face’. In this case, the dependency `FIXED(dar,na cara de)` has been extracted, which happens to be a frozen expression with a similar meaning.

Finally, the next sentence came from a title that is immediately followed by a quotation from the actress it mentions: *Catarina Furtado estreou Uma noite de Sonho na SIC “Fecho os olhos e oiço aplausos”* ‘Catarina Furtado debuted (the show) *Uma noite de Sonho* in SIC (a tv station) “I close my eyes and I hear aplauses”’. The sentence is ambiguous with the idiom *fechar os olhos* ‘close the eyes’, which is an euphemism for ‘to die’. Naturally, this kind of situation can not be solved without information from the context.

The false-negatives are often just idioms that were still missing in the lexicon-grammar. For example, at the end of the next sentence, the idiom *dar a cara por* (lit: to give the face for (sth.)) ‘to represent (sb./sth.)’ was not captured simply because it had not been listed. But, curiously, the sentence opens with another missed idiom: *Com o projeto ainda na gaveta* ‘With the project still in the drawer’, which is probably a reduction from the sentence form *meter na gaveta* (lit: put (sth.) in the drawer) ‘to shelve (sth.); i.e., to prevent something from developing’. These complex prepositional phrases resulting from the reduction of longer sentence structures are still not described in the lexicon-grammar.

## 5 Conclusions

This paper set out to describe the issues raised by the integration of a large-sized database of verbal idioms of European Portuguese into a fully-fledged NLP system. The evaluation on an available corpus of sentences involving body-part nouns, a type of lexical item that is very prone to form idioms in many languages, showed promising results. Several aspects of the syntax of these idiomatic expressions have been described and taken into account for future work. Perhaps the most challenging task ahead is the complete automatization of the rule-generation process.

**Acknowledgments.** This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013; and Erasmus Mundus Action 2 2011-2574 Triple I - Integration, Interaction and Institutions. We would like to thank the comments of the anonymous reviewers, which helped to improve this paper.

## References

1. Ait-Mokhtar, S., Chanod, J., Roux, C.: Robustness beyond shallowness: incremental dependency parsing. Natural Language Engineering 8(2/3), 121–144 (2002)
2. Baptista, J., Correia, A., Fernandes, G.: Frozen Sentences of Portuguese: Formal Descriptions for NLP. In: Workshop on Multiword Expressions: Integrating Processing. Intl. Conf. of the European Chapter of the ACL, Barcelona, Spain, pp. 72–79 (2004)
3. Baptista, J., Correia, A., Fernandes, G.: Léxico Gramática das Frases Fixas do Portugués Europeo. Cadernos de Fraseoloxía Galega 7, 41–53 (2005)
4. Baptista, J., Mamede, N., Gomes, F.: Auxiliary verbs and verbal chains in European Portuguese. In: Pardo, T.A.S., Branco, A., Klautau, A., Vieira, R., de Lima, V.L.S. (eds.) PROPOR 2010. LNCS (LNAI), vol. 6001, pp. 110–119. Springer, Heidelberg (2010)
5. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46 (1960)
6. Copestake, A.: Representing idioms. Presentation at the HPSG Conference, Copenhagen (1994)
7. Fernandes, G., Baptista, J.: Frozen sentences with obligatory negation: linguistic challenges for natural language processing. In: Mellado-Blanco, C. (ed.) Colocaciones y Fraseología en Los Diccionarios, pp. 85–96. Peter Lang, Frankfurt (2008)
8. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psych. Bull. 76(5), 378–382 (1971)
9. Fleiss, J.L.: Statistical methods for rates and proportions, 2nd edn. John Wiley, New York (1981)
10. Gross, M.: Une classification des phrases "figées" du français. Revue Québécoise de Linguistique 12(2), 1–16 (1982)
11. Gross, M.: Lexicon-Grammar. In: Brown, K., Miller, J. (eds.) Concise Encyclopedia of Syntactic Theories, pp. 244–259. Pergamon, Cambridge (1996)
12. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159–174 (1977)
13. Mamede, N., Baptista, J., Diniz, C., Cabarrão, V.: STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In: Intl. Conf. on Computational Processing of Portuguese, Propor 2012, vol. Demo Session (2012), Paper available at `http://www.propor2012.org/demos/DemoSTRING.pdf`
14. Markov, I.: Automatic Identification of Whole-Part Relations in Portuguese. Master's thesis. U. Algarve, Faro (2014)
15. Ramisch, C., Araújo, V., Villavicencio, A.: A Broad Evaluation of Techniques for Automatic Acquisition of Multiword Expressions. In: Proceedings of the ACL 2012 Student Research Workshop, pp. 1–6. ACL (2012)
16. Rocha, P., Santos, D.: CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Nunes, M.G. (ed.) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000), pp. 131–140. ICMC/USP, São Paulo (2000)
17. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for NLP. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 1–15. Springer, Heidelberg (2002)
18. Vale, O.: Expressões Cristalizadas do Português do Brasil: uma proposta de tipologia. Ph.D. thesis. Universidade Estadual Paulista, Araraquara, SP (2001)