Abstract

Dealing with idioms in Natural Language Processing systems is difficult, among other reasons, because their architecture must be conceived in such a way that it should not preclude the processing of both free word combinations and these, more constraint, expressions. On the other hand, many idioms do have syntactic structure, and can undergo several types of formal variation, thus making them hard to identify in a strictly string pattern-matching approach. Furthermore, many of these expressions are ambiguous between a literal (non-idiomatic) and figurative, non-compositional (idiomatic) use, depending of many linguistic and extra-linguistic factors. This paper presents the way (European) Portuguese verbal idioms have been integrated in fully STRING, a hybrid, statistical and rule-based, natural language processing system, and identify several of the problems that had (and some that still have) to be addressed, in order to adequately identify and process idioms in texts.

1. This paper focuses on verbal idioms, e.g. perder a cabeça, lit: ‘lose the head’ (lose one’s head), that is, 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 (M. Gross 1982, 1996).

Extensive lists of verbal idioms, particularly the most frequent ones, have been systematically collected for Portuguese, both the European (Baptista et al. 2004, 2005) and the Brazilian (Vale 2001) varieties, along with their main distributional, syntactic and transformational properties, under the Lexicon-Grammar methodological and theoretical framework (M. Gross 1996). Previous studies have shown that the identification of idioms cannot rely neither on strict pattern-matching techniques (Fernandes e Baptista 2007, 2008), nor the use of association measures suffices to identify many idioms (Baptista et al. 2010), hence much manual development of language resources by linguists is required.

In this paper, we address the main issues raised in the process of integrating the lexicon-grammar of European Portuguese verbal idioms into a fully-fledged natural language processing system, STRING (Mamede et al. 2012). In order to do so, we briefly present the system in the next section.

2. STRING (string.l2f.inesc-id.pt) is a hybrid statistical and rule-based natural language processing chain for Portuguese, with a modular structure, that performs all the basic NLP tasks in four main steps: (i) preprocessing and lexical analysis, (ii) rule-based and (iii) statistical part-of-speech (POS) disambiguation and (iv) parsing. The parsing step is performed by the Xerox Incremental Parser (Ait-Moktar et al. 2002), using a rule-based Portuguese grammar jointly developed by the INESC-ID Lisboa and Xerox. XIP first delimits the elementary phrases (or chunks, like NP, PP, etc.), and then it extracts the dependencies between the chunk’s heads; e.g. SUBJECT, MODifier, CDIR (direct complement), etc.

3. Considering that idioms have a syntactic structure, STRING’s strategy consists in parsing them first as ordinary sentences and only then to identify the word combinations whose meaning is not to be calculated in a compositional way, based on the results of the previous parsing. The idioms are identified by the dependency FIXED, which take as its arguments the verb and the frozen elements of the idiomatic expression (the number of arguments depends
on the type of idiom involved). The figure below illustrates the chunking tree and the relevant dependencies extracted for the sentence *O Pedro perdeu a cabeça* (Peter lost his head), where the dependency `FIXED` has been highlighted:

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

The identification of the idioms uses the previously calculated dependencies and is carried out by a rule like this:

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

which captures any form of the lemma of the verb (including compound tenses) and the surface form of the direct object (obligatorily after the verb). Around 2,400 rules were semi-automatically built for 10 formal classes of verbal idioms. A list of examples (one for each idiom) was used to test the rules. The remaining of the paper presents the different issues that had to be dealt with in the integration of these rules in a functional module. Finally, we present an estimation of the precision of the system on a large Portuguese corpus of news texts.

**References**


