MWEs in the QTLeap Corpus of Online Helpdesk Interactions

Andreia Querido, Rita de Carvalho, João Rodrigues, António Branco
University of Lisbon

1 Introduction

Considerable effort has been devoted to the analysis and extraction of multiword expressions (MWEs) mostly in newspaper texts, but also in literary and didactic texts or documents of parliament sessions, etc. Less attention has been given to the analyses of MWEs in some other kind of texts, as the one addressed in the present study.

The corpus under consideration is being developed in the context of the QTLeap project, whose goal is to research on and deliver a methodology for machine translation that explores deep language engineering approaches as a way of improving translations of higher quality. In order to pursue its objectives, this project is organized around the deployment of machine translation pilots that are trained and evaluated in a real-use scenario corpus that results from the gathering of written interactions via an online chat channel that offers 24h/24 professional expert help on ICT for domestic users and laypersons. This dataset had the potential to present interesting features to study MWEs, as will be shown ahead.

2 Dataset

The QTLeap corpus can be categorized as belonging to the ICT domain and is composed of linguistic interactions, which are provided by a company, Higher Functions, that ensures technical support to its clients. This corpus is composed by written questions made by users and by written answers provided by professionals working in the helpdesk. The corpus is thus composed of real interactions between clients and experts in a support chat line, that originally were in Portuguese, and that were translated into the other seven languages of the project. The Portuguese *corpus* is made of 139 441 tokens in 9 959 sentences.

The annotation of this corpus is crucial to train and evaluate tools and components in the scope of the project. Besides that, it is also useful because it represents a domain and a genre that is seldom studied, even though ICT is a growing and dynamic area. To the best of our knowledge, there is no annotated corpus with these characteristics.

3 Annotation methodology and tool

In order to ensure the reliability of the resulting linguistically interpreted dataset, we followed the methodology of double-blind annotation followed by adjudication: independently of each other, two experts in linguistics annotate the same data and for those cases where their decisions differ, a third annotator makes the final decision.

For this task we used WebAnno, which is a general purpose web-based platform for linguistic annotation (Yimam *et al.*, 2013). It is an annotation tool for a range of linguistic annotations, including various layers of morphological, syntactical, and semantic annotations.

The annotation of MWEs was undertaken on a par with the annotation of Named Entities and Institutionalized Phrases. This task was performed in two stages. In the first stage, the corpus was annotated along three classes: Named Entities (NE), Multiword Expressions (MWE) and Institutionalized Phrases (IP). The goal was to give priority to an empirical approach, by inspecting the corpus without any pre-determined categorization that might bias the annotation. It was important to first check what kinds of MWEs and NEs could be found. After this first stage of annotation, we understood that, in fact, we could proceed with a more detailed annotation for MWEs, and in the second stage, we re-annotated the corpus having adopted a more fine grained categorization and tag set.

4 Theoretical guidance

To guide our work along a principled approach, we made an overview of the key literature concerning MWEs and decided to follow Sag *et al.*'s (2002) proposal. These authors propose a classification of MWEs in two groups: institutionalized phrases and lexicalized phrases.

The institutionalized phrases are statistically idiosyncratic but can be semantically and syntactically compositional. On the other hand, lexicalized phrases represent MWEs with some kind of idiomaticity. They are divided in three subclasses: fixed expressions, semi-fixed expressions and syntactically-flexible expressions.

Fixed expressions are strings without internal modification or morphosyntactic variation. Semi-fixed are non-decomposable idioms and compound nominals. The first one can undergo lexical variation, but not internal modification or passivation; the second type can inflect in gender and number. The syntactically-flexible expressions, in turn, are verb-particle constructions, decomposable idioms and light verbs. They can undergo morphosyntactic variation, internal modification, passivation and changes in the order of its constituents.

We adopted this classification but we marked as MWEs only what Sag *et al.* (2002) considered lexicalized phrases given IPs have no subclasses and had already been annotated in the first stage of annotation.

5 Findings

We found 1 095 MWEs in the first round of annotation. In the second round, we found three subclasses: fixed expressions, compound nominals (a subtype of semi-fixed expressions) and light verbs (a subtype of syntactically-flexible expressions):

MWEs types		
Fixed expressions	Compound nominals	Light verbs
25%	71%	4%

There are only a few types of MWEs in this *corpus*, and in spite of adopting the Sag *et al.* (2002), a taxonomy with 6 types does not appear fully instantiated. The *corpus* is in Portuguese, but most of the MWEs are in English, because in several cases the users do not translate these terms from the source language. Some examples of English technical terms used in Portuguese are *PIN* (*Personal Identification Number*) or *HDMI* (*High-Definition Multimedia Interface*), among several others.

6 Final Remarks

We have presented a corpus that contributes to a more accurate analysis of MWEs in an uncommon domain and genre. Given that the corpus has been translated into multiple languages, this type of work can be the first step in the creation of a domain specific multilingual aligned lexicon. This resource could be useful, not only for machine translation, but also as data for domain adaptation of a machine learning MWE recognition tool, etc.

Acknowledgments

This work was partially supported by EU FP7 project 610516 “QTLeap” and by FCT project PTDC/EEI-SII/1940/2012 "DP4LT".

References

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of* CICLing'02., pp.1–15, Springer-Verlag.

Yimam, S.M., Gurevych, I., Eckart de Castilho, R., and Biemann C. (2013): WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of ACL-2013, demo session*, Sofia, Bulgaria.