MWEs in the QTLeap Corpus of Online Helpdesk Interactions

Andreia Querido, Rita de Carvalho, João Rodrigues, António Branco
(andreia.querido, rita.carvalho, joao.rodrigues, antonio.branco)@di.fc.ul.pt
University of Lisbon

Dataset

♦ Portuguese corpus made of 139 411 tokens in 9 959 sentences;

♦ Helpdesk corpus: real interactions between clients and experts in a support chat line;

♦ Information and Communications Technology domain;

♦ Corpus is translated from portuguese to seven languages.

Annotation methodology and tool

Tool
WebAnno (Yimam *et al.*, 2013)

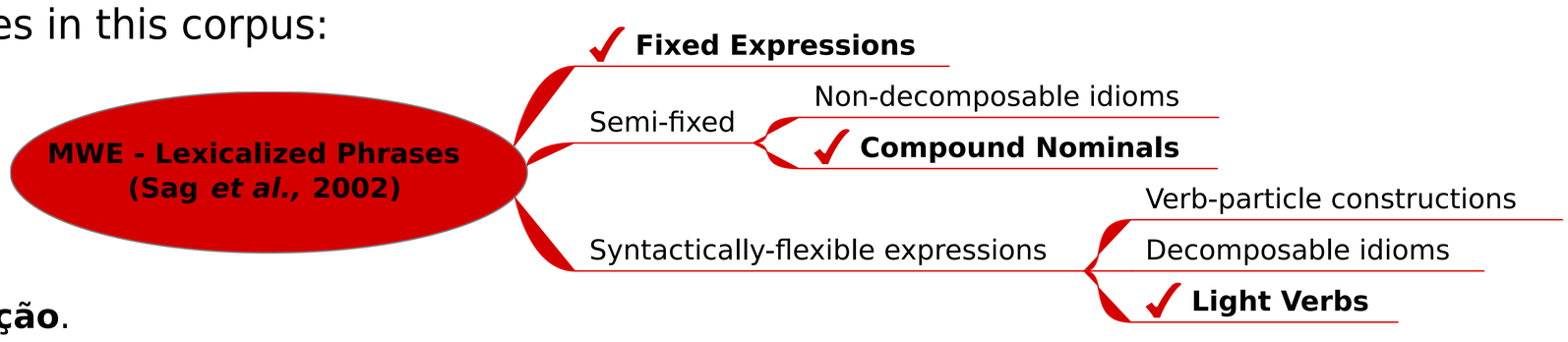
Empirical approach
♦ Double-blind annotation followed by adjudication;
♦ First stage: annotation of MWEs, Named Entities, Institutionalized Phrases;
♦ Second stage: annotation of MWEs with more fine grained categorization;

Theoretical guidance
Sag et al., 2002

Named Entities and Institutionalized Phrases were not included in this classification of MWEs.

Findings

♦ Total of MWEs: 1132.
♦ The corpus is in Portuguese, but most of the MWEs are in English, because in several cases the users do not translate these from the source language.
♦ Only 3 subclasses have occurrences in this corpus:



Examples:

♦ Fixed Expressions:
PT: Também reproduz vídeo de **alta definição**.
EN: It also plays **high-definition** video.

PT: No youtube como posso pesquisar videos em **HD**?
EN: In YouTube, how can I search for videos in **HD**?

♦ Compound Nominals:
PT: Arraste e solte os ficheiros selecionados do iTunes para o **Ambiente de Trabalho**.
EN: Drag and drop selected files from iTunes to **Desktop**.

PT: Tenho conflito de **IPs** e não consigo chegar a um servidor.
EN: I have an **IP** [address] conflict and I can´t connect to the server.

♦ Light Verbs:
PT: Onde posso **fazer download** do Direct X?
EN: here can I **download** Direct X?

Distribution of MWEs		
Fixed Expressions	**Compound Nominals**	**Light Verbs**
287 occurrences	799 occurrences	46 occurrences
25%	71%	4%
87 types	116 types	13 types

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.

This work was partially supported by the European Comission through project QTLeap under the contract #610516 and by FCT- Fundação para a Ciência e Tecnologia,the project DP4LT, under the contract PTDC/EEI-SII/1940/2012

Working Group: WG4