Creating a Technical Corpus

Summer School of Linguistics
PER-FIDE 2013

Pedro Carvalho
pg22760@alunos.uminho.pt

Agenda

- Introduction
- Thesis Abstracts
  - Sources
  - Vertical Alignment
- ECTS
  - Sources
  - Linking the Pages
- Examples

Introduction

Every day countless technical documents are produced inside our learning institutions walls.

Even a quick inspection of their web pages reveal lots of useful technical information and terminology.

If these resources exist and are open to the public in general why won't we use them?

Thesis Abstracts - Source

The Portuguese Open Access Scientific Repository (RCAAP) contain a great collection of thesis abstracts translated to one or more languages.

When obtaining those abstracts some problems arose, but in this presentation the main focus will be the vertical alignment problem.

Thesis Abstracts - Vertical Alignment

In 30% of the cases the abstracts ended up vertically alligned with no possibilty to detect the breaking point between different languages.

(...)

Acresce a todas estas actividades que a Opuntia ficus indica é muito rica nutricionalmente e daí o seu uso na alimentação. Medicinal plants have been used since Antiquity, by primitive civilizations, to fight illnesses. Nowadays, phytotherapy shows to be more and more sustained due to quality, safety and efficiency requirements, and therefore medicinal plants are being

(...)

Thesis Abstracts - Vertical Alignment

Strategy:

1. Break the chunk of text into phrases.

É um cacto arbustivo constituído por raiz, caule (cladódios) onde se encontram, consoante a época, flores e frutos (doces e suculentos). É uma planta cujos caules jovens ("nopalitos") e frutos são usados na dieta alimentar dos Mexicanos.

É um cacto arbustivo constituído por raiz, caule (cladódios) onde se encontram, consoante a época, flores e frutos (doces e suculentos).

É uma planta cujos caules jovens ("nopalitos") e frutos são usados na dieta alimentar dos Mexicanos.

Thesis Abstracts - Vertical Alignment

Strategy:

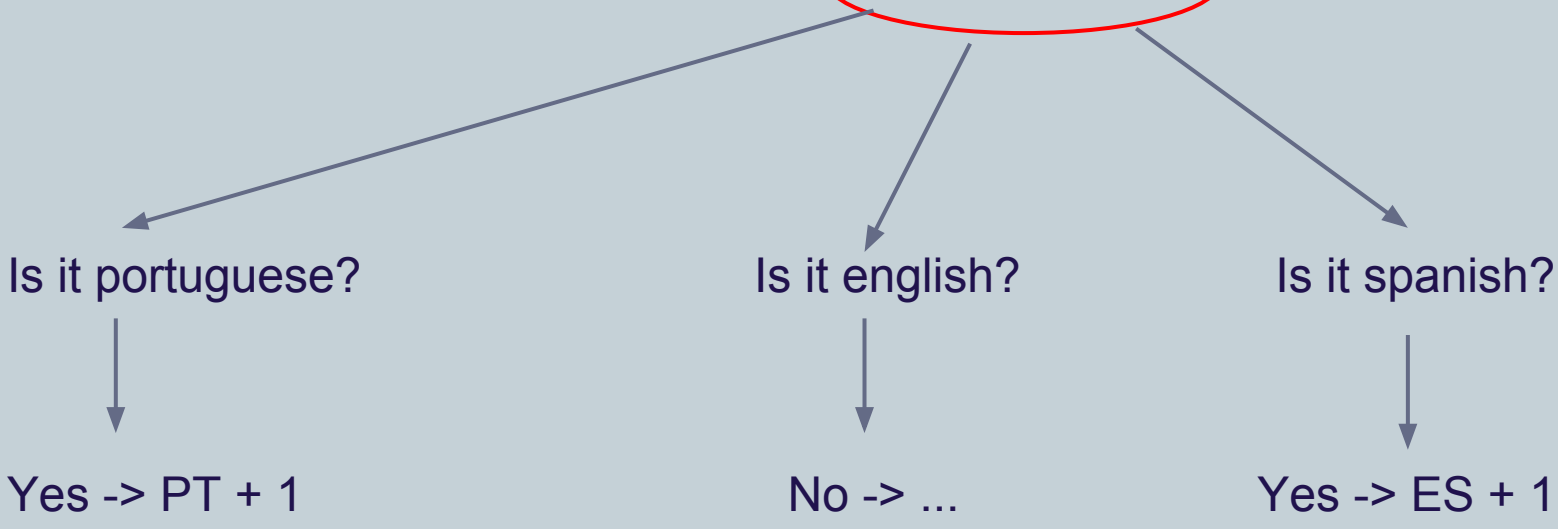
2. For each phrase:
   a. Break it in words.

3. For each word:
   a. Check if it belongs to any dictionaries.

4. In the end choose the dictionary with more words associated.

Thesis Abstracts - Vertical Alignment

É uma planta cujos caules jovens ("nopalitos") e frutos são usados na dieta alimentar dos Mexicanos.



What should we do in case of a tie?

Thesis Abstracts - Vertical Alignment

Strategy:

3. In case of a tie
   a. Check the neighborhood.
   b. Choose the dictionary with more words associated in the global context.

Thesis Abstracts - Vertical Alignment

(...)

Actualmente, a fitoterapia está cada vez mais sustentada por requisitos de eficácia, segurança e qualidade, sendo as plantas medicinais bastante procuradas como recurso terapêutico (Cunha, Silva e Roque, 2003). → PT

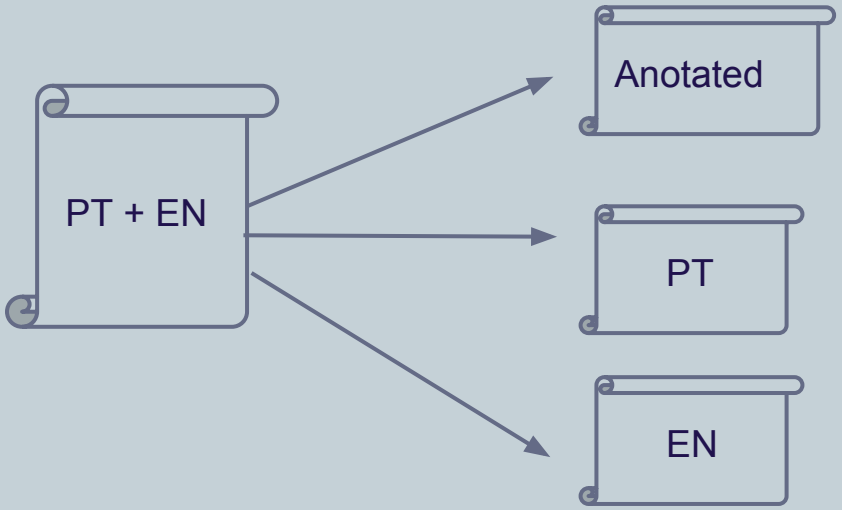
A Opuntia ficus indica é uma planta originária do México que cresce de forma selvagem em regiões áridas e semi-áridas de todo o mundo (Ennouri et al., 2006b). → PT

É um cacto arbustivo constituído por raiz, caule (cladódios) onde se encontram, consoante a época, flores e frutos (doces e suculentos). → PT

(...)

Thesis Abstracts - Vertical Alignment

Each original document originates a final anotated document and a new document for each detected language.



We randomly chose 50 files that were produced by this tool.

In 1043 phrases that were analyzed we detected 13 errors.

Giving us a success percentage of 98.75%.

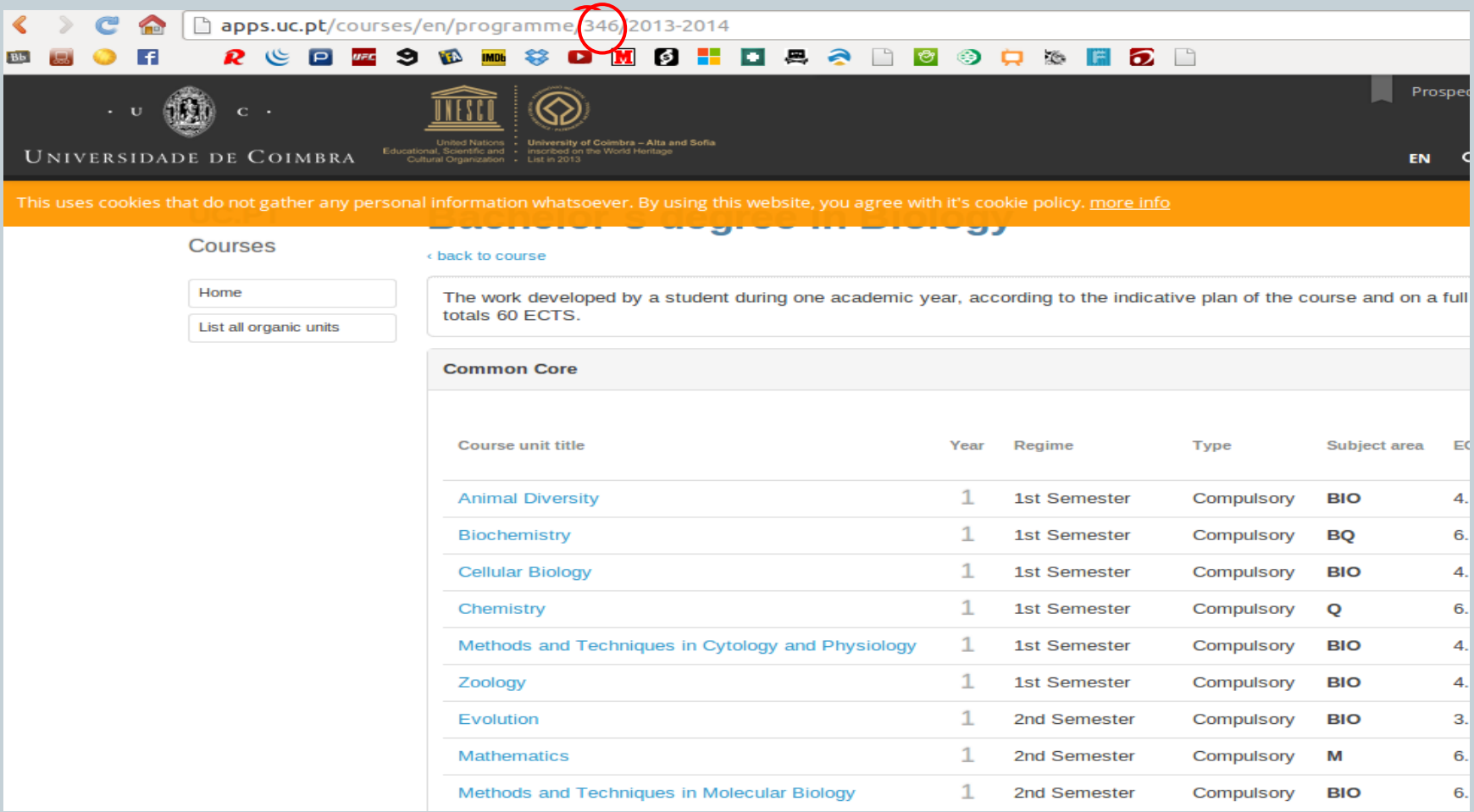
ECTS - The Source

In pratically any learning institution's web page we can easily find study plans and information about all the courses they offer.

These pages are usually translated at least to english.

The goal was to obtain these pages and somehow link all the original versions to it's counterpart translated versions.

ECTS - Linking the Pages



ECTS - Linking the Pages

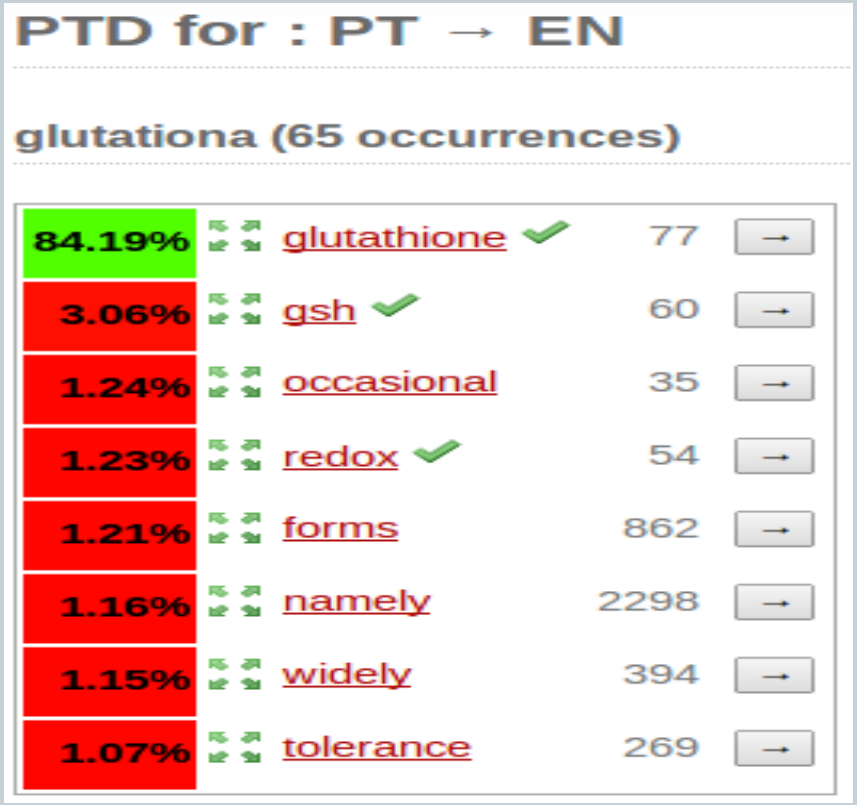
The pages that cannot be directly associated through it's url can be linked through mechanization. However each page has it's own characteristics.

To detect and possibly correct the alphabetically sorted lists we can use ucts (unambiguos-concept translation sets).

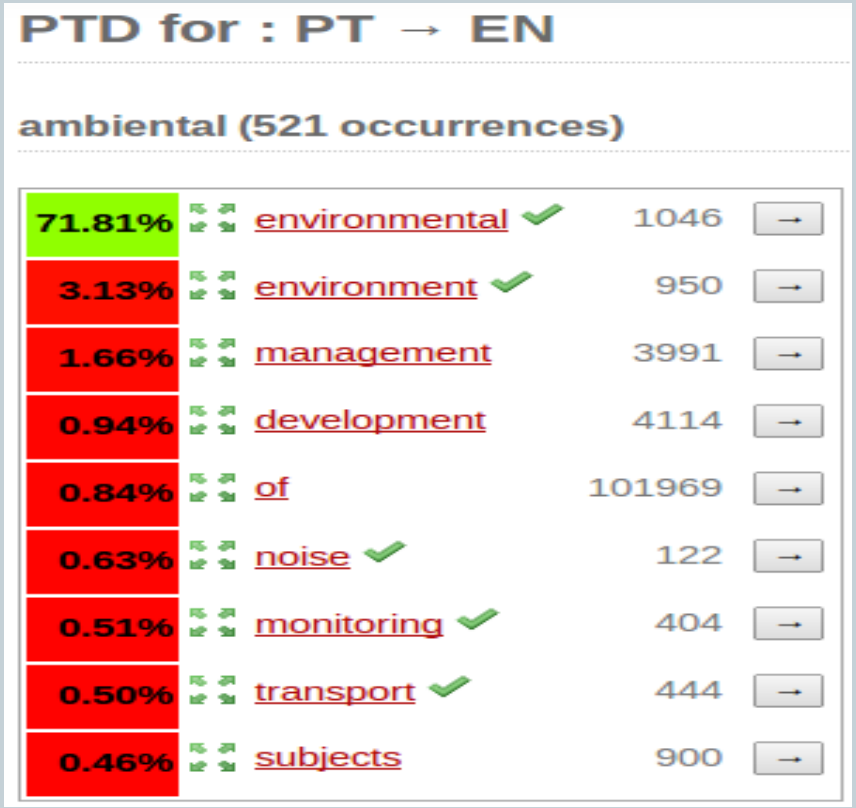
New ideas are welcome!!

Examples

THESIS - ABSTRACT



ECTS



Creating a Technical Corpus

Summer School of Linguistics
PER-FIDE 2013

Pedro Carvalho
pg22760@alunos.uminho.pt
Thank You :)