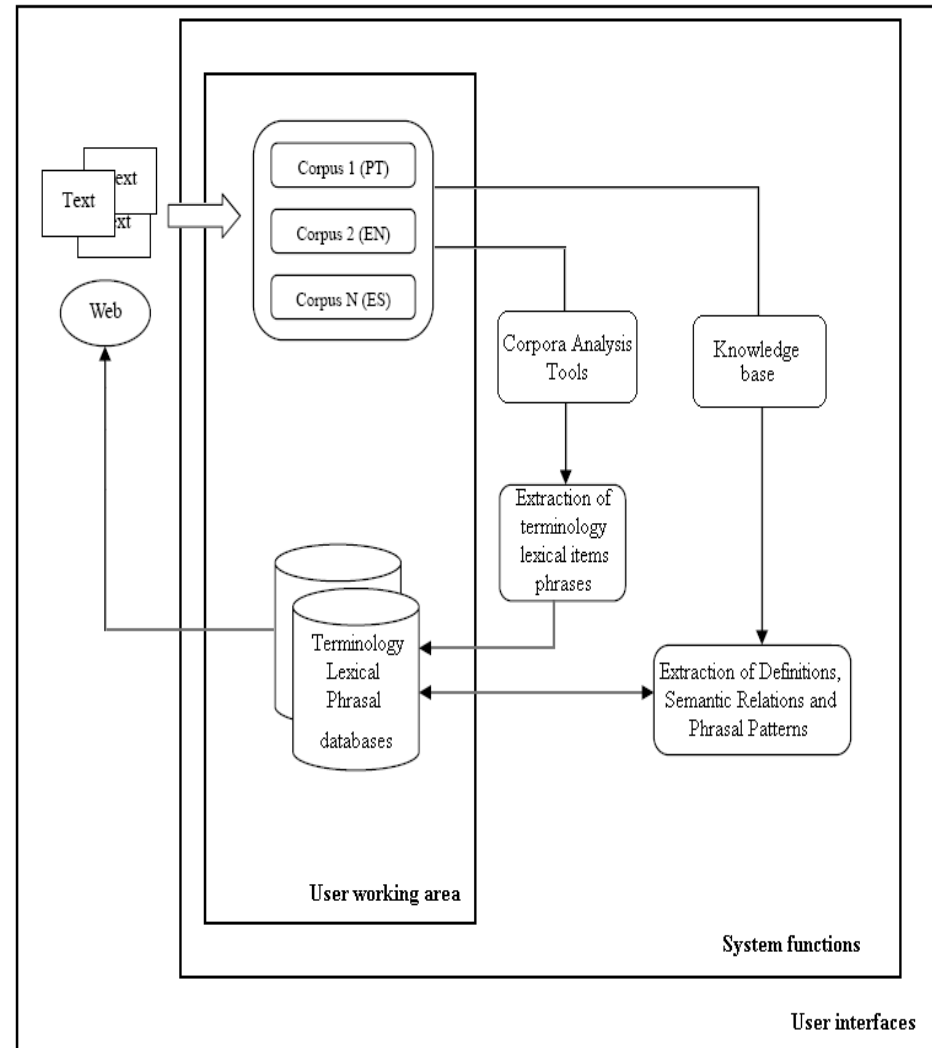# Corpógrafo and NooJ

Using Linguistic Resources to Obtain Aligned Concordances from Corpora

Sérgio Matos, Anabela Barreiro, Belinda Maia
April 2008

Corpógrafo

- Text manipulation

- Corpora analysis
  - Concordances
  - n-grams

- Terminology
  - Extraction of terms
  - Extraction of relations, definitions

Corpógrafo

- Concordances

    - Using regular expressions (e.g. "(the|a) \w+ is \w+")

- N-grams

    - Identification of multi-word expressions / lexical bundles

Corpógrafo

- N-grams

Corpus: Natura-PUBLICO (1726679 átomos)
Lista com 981104 N-Gramas de comprimento 3

[|<]   [<<]   [>>]   [>|]

#	n-grama	# oc.	freq. %
1	milhões de contos	628	0.036
2	, o que	514	0.029
3	, que se	348	0.020
4	no entanto ,	331	0.019
5	, com a	298	0.017
6	, com o	293	0.016
7	, depois de	292	0.016
8	o presidente da	249	0.014
9	por cento ,	246	0.014
10	milhões de dólares	239	0.013
11	secretário de estado	239	0.013
12	, e a	217	0.012
13	em lisboa ,	216	0.012
14	a partir de	215	0.012
15	por cento .	215	0.012
16	, e o	213	0.012
17	de acordo com	213	0.012
18	de contos .	212	0.012
19	, em que	211	0.012
20	, no entanto	192	0.011
21	do ano passado	186	0.010
22	, em lisboa	179	0.010
23	o facto de	179	0.010
24	ano passado ,	177	0.010
25	por cento do	176	0.010
26	, apesar de	174	0.010

Corpógrafo

- Concordances
  - Using regular expressions (e.g. "(the|a) \w+ is \w+")

- N-grams
  - Identification of multi-word expressions / lexical bundles

- Term extraction
  - Based on n-grams (using filters / stop-words)

Corpógrafo

Corpógrafo

- Concordances

  – Using regular expressions (e.g. "(the|a) \w+ is \w+")

- N-grams

  – Identification of multi-word expressions / lexical bundles

- Term extraction

  – Based on n-grams (using filters / stop-words)

- Extraction of definitions / semantic relations

  – Pattern-based

Corpógrafo

- Extraction of Definitions

Example patterns:

```
(^|^the |^a |^an | the | a | an )__TERMO__ is (a|an|the) \w+ \w+ of the
(^|^the |^a |^an | the | a | an )__TERMO__ is (a|an|the) \w+ of
(^|^the |^a |^an | the | a | an )__TERMO__ is \w+ with
(^|^the |^a |^an | the | a | an )__TERMO__ is made of
(^|^the |^a |^an | the | a | an )__TERMO__ is made up of
(^|^the |^a |^an | the | a | an )__TERMO__ is the main
(^|^the |^a |^an | the | a | an )__TERMO__ is (that|the) part of
(^|^the |^a |^an | the | a | an )__TERMO__ is (part|composed) of
```

Corpógrafo

Corpógrafo

- Extraction of Semantic Relations

  Example patterns for holonymy / meronymy:

  (^|^a |^an |^the | a | an | the)__TERMO2__ is located in the __TERMO1__
  (^|^a |^an |^the | a | an | the)__TERMO1__ is (composed|made up) of .* : the .* ,
  (the|a|an) __TERMO2__
  (^|^a |^an |^the | a | an | the)__TERMO1__ is (composed|made up) of .* :
  (the|a|an) __TERMO2__
  (^|^a |^an |^the | a | an | the)__TERMO1__ which is composed of \w+ __TERMO2__
  (^|^a |^an |^the | a | an | the)__TERMO2__ is the \w+ of the __TERMO1__
  (^|^a |^an |^the | a | an | the)__TERMO2__ , the \w+ part of the __TERMO1__
  (^|^a |^an |^the | a | an | the)__TERMO1__ contains the \w+ , __TERMO2__
  (^|^a |^an |^the | a | an | the)__TERMO1__ consists of (the|a|an) __TERMO2__

Corpógrafo

Corpógrafo

- Aligned concordances
  - Parallel Corpora
  - Comparable Corpora

- Phraseology / Lexicography
  - Identification of phrases / lexi...
  - Discourse analysis
  - Genre analysis
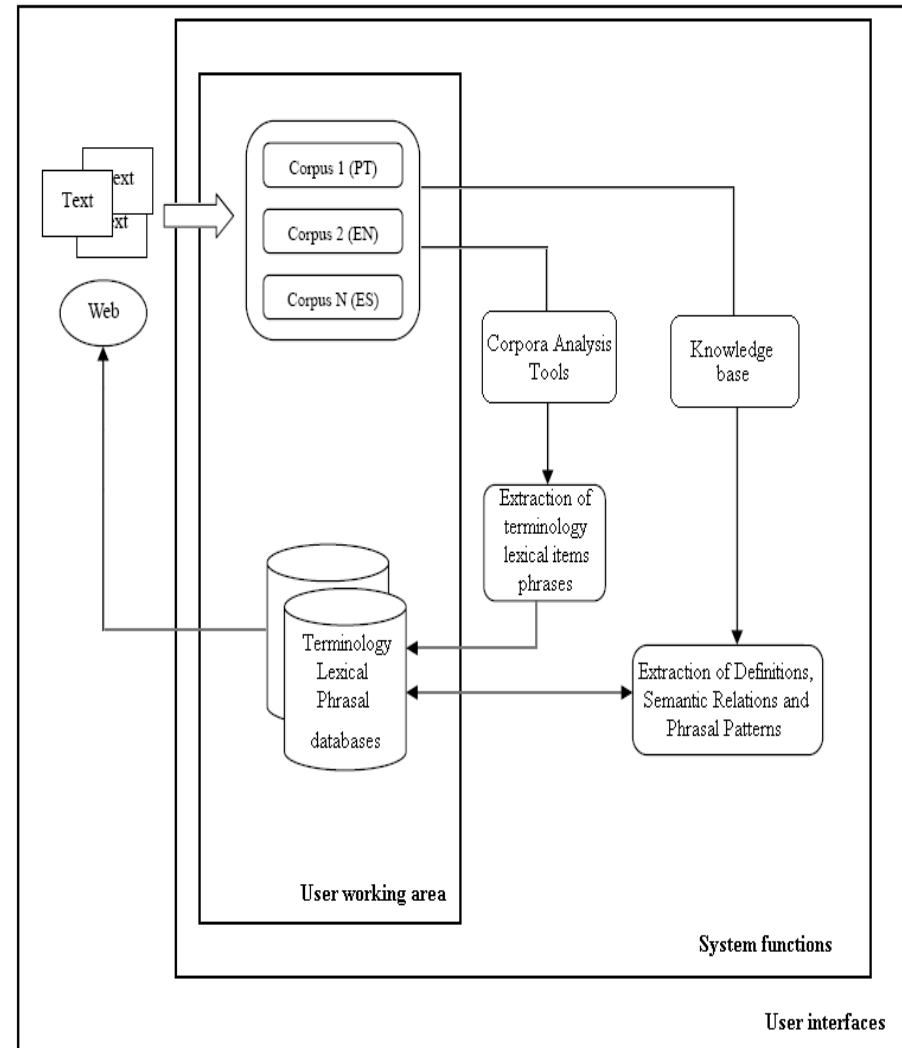
Corpógrafo
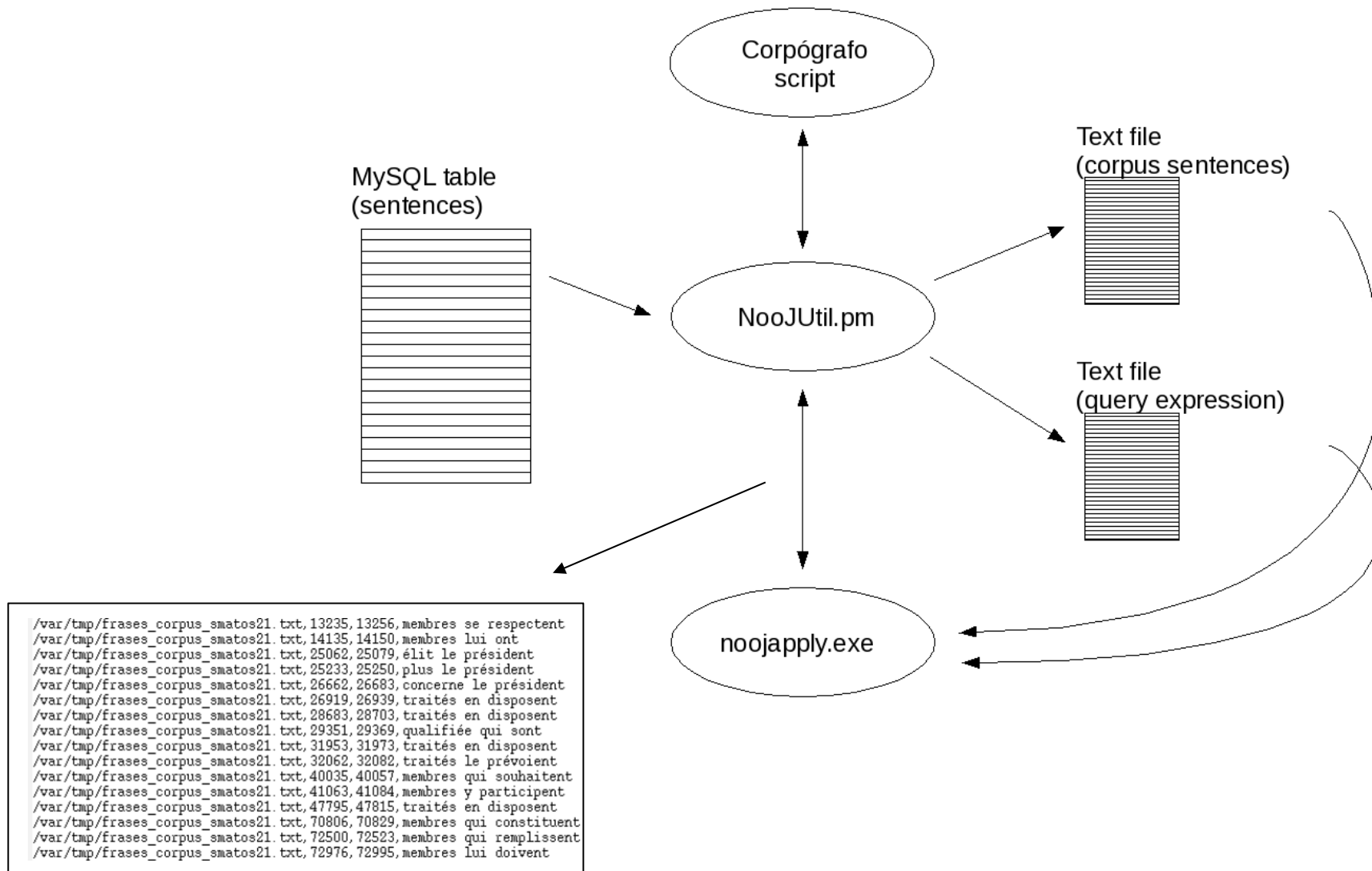
- Parallel Corpora

Integrating NooJ

- Motivation
  - Integrate POS information for obtaining concordances and extracting lexical items / phrases
  - Syntactic and semantic analysis

- Why NooJ?
  - Comprehensive corpus processing engine
  - Extensible through the use of lexical and syntactic grammars
  - Linguistic resources available for various languages

Integrating NooJ

- Practical aspects
  - Executable in Linux under the Mono platform (http://www.mono-project.com/)
  - Command-line programme (noojapply.exe)
  - Communication through text files

- Implementation
  - Perl module (NooJUtil.pm)
    - **init**: write corpus sentences to file; select dictionary
    - **query**: obtain NooJ concordances with noojapply.exe
    - **concordance**: obtain results and convert to Corpógrafo's concordance format

Integrating NooJ

- NooJUtil module

Diagram showing NooJUtil.pm at the center, connected to: Corpógrafo script, MySQL table (sentences), Text file (corpus sentences), Text file (query expression), noojapply.exe, and a text output listing:

```
/var/tmp/frases_corpus_smatos21.txt,13235,13256,membres se respectent
/var/tmp/frases_corpus_smatos21.txt,14135,14150,membres lui ont
/var/tmp/frases_corpus_smatos21.txt,25062,25079,élit le président
/var/tmp/frases_corpus_smatos21.txt,25233,25250,plus le président
/var/tmp/frases_corpus_smatos21.txt,26662,26683,concerne le président
/var/tmp/frases_corpus_smatos21.txt,26919,26939,traités en disposent
/var/tmp/frases_corpus_smatos21.txt,28683,28703,traités en disposent
/var/tmp/frases_corpus_smatos21.txt,29351,29369,qualifiée qui sont
/var/tmp/frases_corpus_smatos21.txt,31953,31973,traités en disposent
/var/tmp/frases_corpus_smatos21.txt,32062,32082,traités le prévoient
/var/tmp/frases_corpus_smatos21.txt,40035,40057,membres qui souhaitent
/var/tmp/frases_corpus_smatos21.txt,41063,41084,membres y participent
/var/tmp/frases_corpus_smatos21.txt,47795,47815,traités en disposent
/var/tmp/frases_corpus_smatos21.txt,70806,70829,membres qui constituent
/var/tmp/frases_corpus_smatos21.txt,72500,72523,membres qui remplissent
/var/tmp/frases_corpus_smatos21.txt,72976,72995,membres lui doivent
```

Integrating NooJ

- Converting results

Select corresponding sentence

```
/var/tmp/frases_corpus_smatos21.txt,13235,13256,membres se respectent
/var/tmp/frases_corpus_smatos21.txt,14135,14150,membres lui ont
/var/tmp/frases_corpus_smatos21.txt,25062,25079,élit le président
/var/tmp/frases_corpus_smatos21.txt,25233,25250,plus le président
/var/tmp/frases_corpus_smatos21.txt,26662,26683,concerne le président
/var/tmp/frases_corpus_smatos21.txt,26919,26939,traités en disposent
/var/tmp/frases_corpus_smatos21.txt,28683,28703,traités en disposent
/var/tmp/frases_corpus_smatos21.txt,29351,29369,qualifiée qui sont
/var/tmp/frases_corpus_smatos21.txt,31953,31973,traités en disposent
/var/tmp/frases_corpus_smatos21.txt,32062,32082,traités le prévoient
/var/tmp/frases_corpus_smatos21.txt,40035,40057,membres qui souhaitent
/var/tmp/frases_corpus_smatos21.txt,41063,41084,membres y participent
/var/tmp/frases_corpus_smatos21.txt,47795,47815,traités en disposent
/var/tmp/frases_corpus_smatos21.txt,70806,70829,membres qui constituent
/var/tmp/frases_corpus_smatos21.txt,72500,72523,membres qui remplissent
/var/tmp/frases_corpus_smatos21.txt,72976,72995,membres lui doivent
```

Concordancia.pm

- left and right context
- origin file
- occurrences

Using NooJ Expressions
in Corpógrafo

Using NooJ Expressions in Corpógrafo

Using NooJ Expressions in Corpógrafo

Using NooJ Expressions
in Corpógrafo

Using NooJ Expressions in Corpógrafo

Using NooJ Expressions in Corpógrafo

Using NooJ Expressions in Corpógrafo

Applications   Places   System   Sérgio Matos   Fri 18 Apr, 12:09

Corpógrafo - neurodemo - Mozilla Firefox

File   Edit   View   History   Bookmarks   Tools   Help

http://poloclup.linguateca.pt/cgi-bin/corpografo/user/sagi/sagi.pl   Google

Disable▾ Cookies▾ CSS▾ Forms▾ Images▾ Information▾ Miscellaneous▾ Outline▾ Resize▾ Tools▾ View Source▾ Options▾

Corpógrafo - Linguateca

- Gestor
- Pesquisa
  - Concordância Frase
  - Estudo de N-Gramas
  - Concordância Janela
  - Concordância KWIC
  - Concordâncias paralelas
- Centro de Conhecimento
- Centro de Comunicação

Concordâncias

* Voltar ao menu
* Analisar Frequências
* Contexto Janela 1-1
* Contexto Janela 2-2

Imprimir!

Informações

Expressão de pesquisa: **<WF> <V> <WF> (of + for + in) <WF>**
Corpus pesquisado: Neurons (29181 átomos : 37 ficheiros)

Nº de concordâncias obtidas: 323 distribuídas por 31 ficheiros
Frequência da totalidade das instâncias da concordância: 1.10 %

[I<]   [<<]   [>>]   [>I]

#	f	Frase onde ocorre a concordância	Info
1	0	Activation of the transcription factor NF-[kappa]B in Schwann **cells is required for peripheral** myelin formation	
2	1	The human body **is made up of trillions** of cells .	
3	1	Nerve cells - or neurons - are the cells in our body **that are responsible for bringing** information from the surrounding , analyze this input , and finally create a perception .	
4	1	**Dentrites branch out in treelike** fashion and serve as the main apparatus for receiving signals from other nerve cells .	
5	1	This site will help you learn about the **five major sections of a** Nerve cell :	
6	1	A neuron is **the functioning unit of the** nervous system ; specialized to receive , integrate , and transmit information .	
7	1	The gaps (approximately 1 micrometer wide) formed between myelin sheath cells long the axons **are called Nodes of Ranvier** .	
8	1	Plant cells are eukaryotic **and have many of the** structures found in animal cells .	
9	1	The brain **is made up of more** than just nerve cells (neurons) .	
10	1	Activation of the transcription factor NF-B in Schwann **cells is required for peripheral** myelin formation .	
11	2	Nervous **tissue is composed of two** main cell types : neurons and glial cells .	
12	2	The axon is the **main conducting unit of the** neuron , capable of conveying electrical signals along distances that range from as short as 0.1 mm to as long as 2 m .	
13	2	In addition , all eukaryotic cells **contain elaborate systems of internal** membranes which set up various membrane-enclosed compartments within the cell .	
14	2	The energy **released is trapped in the** form of ATP for use by all the energy-consuming activities of the cell.	
15	2	The membrane with its mosaic of **proteins is responsible for many** important functions : keeping certain ions and small molecules out of the cell and letting others in , accumulating nutrients , rejecting harmful substances , catalyzing enzymatic reactions , and conducting an impulse being sensitive to particular neurotransmitters and modulators .	
16	2	Cilia and Flagella Cells **contain elaborate arrays of protein** fibers that serve such functions as : establishing cell shape providing	

Done

Corpógrafo - neurodemo - Mozilla Firefox

What next?

- Annotations
- Grammars
  - Syntactic / semantic analysis
  - Disambiguation
- Lexical / phrasal items
  - Rules and grammars
- Discourse analysis
  - Discourse markers

Acknowledgment

- This work was done in the scope of the Linguateca, contract nº339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union.