Letras and Números: web-based corpus tools

Hiroto UEDA (University of Tokyo)

Antonio MORENO-SANDOVAL (UAM)

This story from the start…

- In CICL13 (Alicante) Aquilino Sánchez, Pascual Cantos and Antonio Moreno had a conversartion on developing a corpus tool similar to The Sketch Engine but FREE.
- In the same Conference, Ueda and Moreno met and decided to run a joint project for reusing and sharing resources (corpora and tools)

Do we need another corpus tool?

Possible reasons:

- I have my own texts but I need a powerful search and statistical outputs.
- I would like to use for free some existing corpora.
- I prefer a web tool than a stand-alone application.

What are Letras and Números?

- Web versions of previous stand-alone programs:
  - LETRAS: a concordancing tool for pre-loaded and new corpora
  - NUMEROS: a statistical tool for analysing pre-calculate data

  Both are fully integrated: numerical data extracted by LETRAS can be transferred to NUMEROS.

LETRAS-web: Input

- Over 10 pre-loaded multilingual corpora:
  - **spoken corpora**: CORLEC, C-ORAL-JP, MAVIR
  - **diachronical corpora**: CODCAR, CODEA, LEMI
  - **dialectal corpora**: ANDES
  - **learning corpora**: VARITEX
  - **parallel corpora**: BILING-es-jp; BILING-jp-es

- **You can 'cut and paste' your own text**

Concordance in context

- Search in context (large or narrow):
  - in paragraph
  - in sentence
- Filter on meta-data (only for some pre-loaded corpora):
- Item frequency table, ordered by some filter and with absolute, relative or normalised frequency:
  - e.g. two patterns ordered by filter of location and relative frequency

Search by regular expressions

- list of words: #[a-zA-Z]+#
- list of n-grams:
  - 2-grams: #[a-z]+\s[a-z]+#
  - 6-grams: #[a-z]+\s[a-z]+\s[a-z]+\s[a-z]+\s[a-z]+\s[a-z]+#
- list of Verbs in infinitive:
  - #[a-z]+ar#
  - #[a-z]+er#
  - #[a-z]+ir#
- verb patterns: V +w +w
  - #poder\s[a-z]+r# (PODER + INFINITIVE)
  - #ir\sa\s[a-z]+r# (IR + ‘a’ + INFINITIVE)

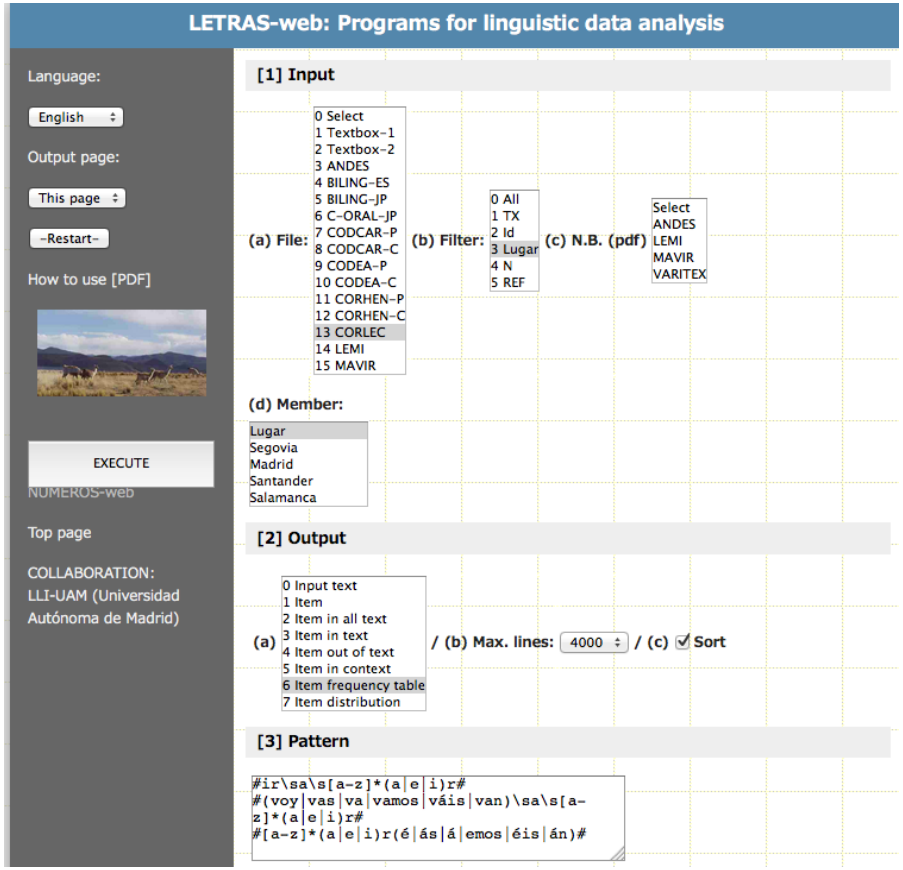
some shortcuts for RE

- & = any word with extended characters
- @ = any letter

4& = 4-grams of words

4@= 4-grams of letters

LETRAS-web


LETRAS-web: Programs for linguistic data analysis
Language:
English
Output page:
This page
–Restart–
How to use [PDF]
EXECUTE
NUMEROS-web
Top page
COLLABORATION:
LLI-UAM (Universidad Autónoma de Madrid)
[1] Input
(a) File:
0 Select
1 Textbox-1
2 Textbox-2
3 ANDES
4 BILING-ES
5 BILING-JP
6 C-ORAL-JP
7 CODCAR-P
8 CODCAR-C
9 CODEA-P
10 CODEA-C
11 CORHEN-P
12 CORHEN-C
13 CORLEC
14 LEMI
15 MAVIR
(b) Filter:
0 All
1 TX
2 Id
3 Lugar
4 N
5 REF
(c) N.B. (pdf)
Select
ANDES
LEMI
MAVIR
VARITEX
(d) Member:
Lugar
Segovia
Madrid
Santander
Salamanca
[2] Output
(a)
0 Input text
1 Item
2 Item in all text
3 Item in text
4 Item out of text
5 Item in context
6 Item frequency table
7 Item distribution
/ (b) Max. lines: 4000 / (c) Sort
[3] Pattern
#ir\sa\s[a-z]*(a|e|i)r#
#(voy|vas|va|vamos|váis|van)\sa\s[a-z]*(a|e|i)r#
#[a-z]*(a|e|i)r(é|ás|á|emos|éis|án)#

Examples: pattern search in context ordered by location

CORLEC (Spanish spoken corpus, 1.1 M. words)

Pattern: IR + a + INFINITIVE (going to)

#Op	#Anterior context	Item	Posterior context	Id	Lugar
1	por sorpresa, y ahora todos tienen que	ir a buscar	a este hombre que manda en la carrera.	ADEP009A	Madrid
2	H1: Sobre todo tú porque yo tengo que	ir a buscar	a los concursantes.	BENT009A	Barcelona
3	y hemos tenido que	ir a buscar	un preservativo de.	ALUD030A	Madrid
4	que habría que	ir a capturar	a cada guerrillero por la selva, en este caso por el desierto?	ADEB003A	Madrid
5	En la época del año en que los animales tienen los hijitos está prohibido	ir a cazar	.	cedu020b	Segovia

Examples: frequency table of verbs in infinitive ordered by pattern (-AR, -ER, -IR)

Absolute Frequency

Pattern	All
#[a-z]*ar#	11937
#[a-z]*er#	11462
#[a-z]*ir#	5060

Relative Frequency

Pattern	All
#[a-z]*ar#	41.9%
#[a-z]*er#	40.3%
#[a-z]*ir#	17.8%

X per thousand words

Pattern	All
#[a-z]*ar#	11.014
#[a-z]*er#	10.576
#[a-z]*ir#	4.669

Future: periphrastic vs. synthetic uses, relative distribution by city in CORLEC

- Periphrastic forms:
  - IR + a + INFINITIVE
  - (voy | vas | va | vamos | váis |van) + a + INFINITIVE
- Future Inflected forms: stem + r (é|ás|á|emos|éis|án)

Pattern	Segovia	Madrid	Santander	Salamanca	Pirineo	Empresa	Varios	Barcelona
#(voy|vas|va|vamos|váis|van)\sa\s[a-z]*(a|e|i)r#	65.6%	57.8%	8.3%	32.7%	50.0%	41.9%	81.2%	57.8%
#[a-z]*(a|e|i)r(é|ás|á|emos|éis|án)#	30.2%	41.7%	91.7%	67.3%	50.0%	58.1%	18.8%	38.6%
#ir\sa\s[a-z]*(a|e|i)r#	4.2%	0.6%						3.6%

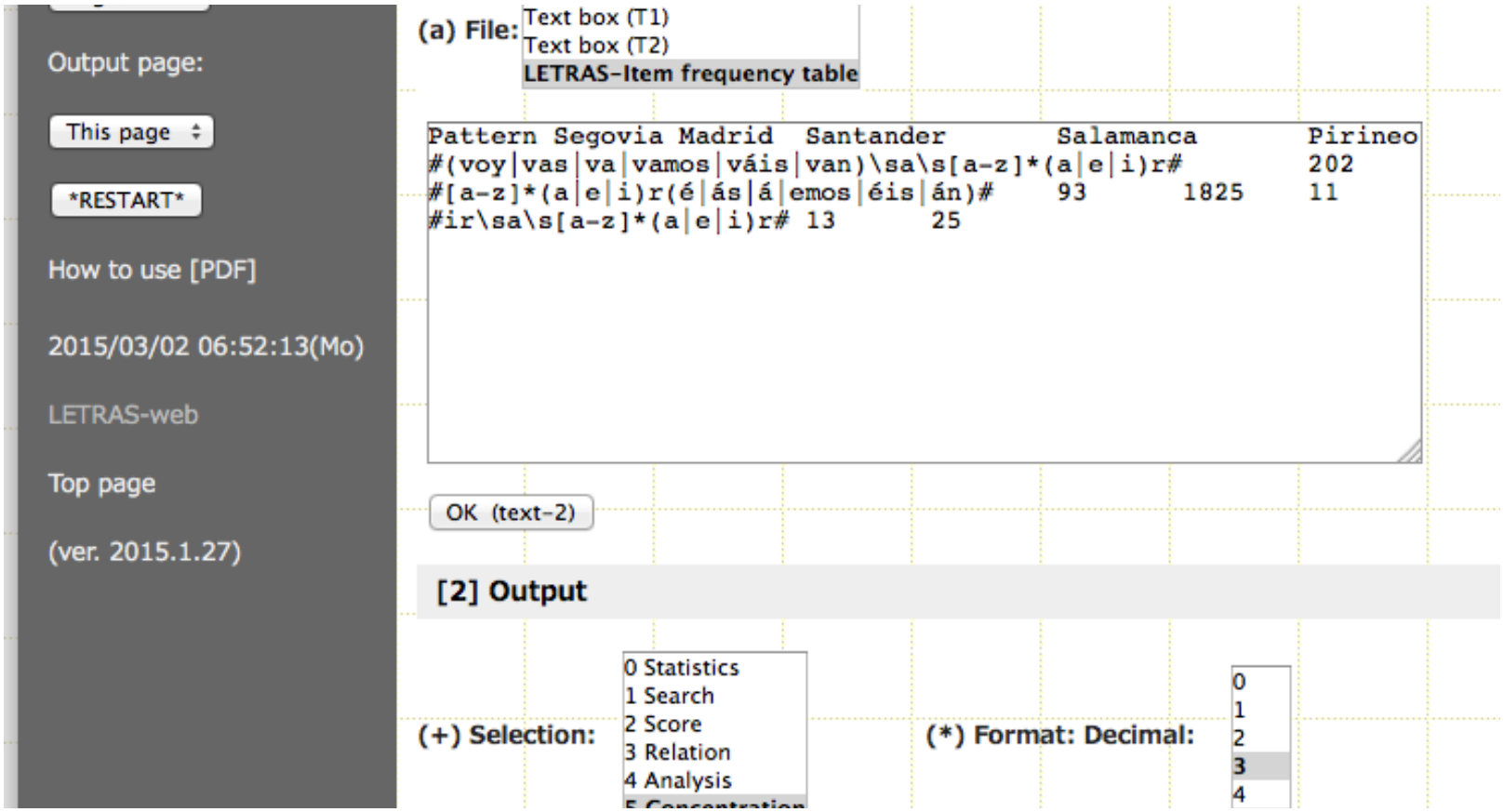
Future: absolute and normalised distributions by city in CORLEC

Pattern	Segovia	Madrid	Santander	Salamanca	Pirineo	Empresa	Varios	Barcelona
#(voy|vas|va|vamos|váis|van)\sa\s[a-z]*(a|e|i)r#	202	2530	1	17	11	13	13	48
#[a-z]*(a|e|i)r(é|ás|á|emos|éis|án)#	93	1825	11	35	11	18	3	32
#ir\sa\s[a-z]*(a|e|i)r#	13	25						3

X per 1000 words

Pattern	Segovia	Madrid	Santander	Salamanca	Pirineo	Empresa	Varios	Barcelona
#(voy|vas|va|vamos|váis|van)\sa\s[a-z]*(a|e|i)r#	2.500	2.739	0.186	0.806	1.269	2.565	1.608	1.918
#[a-z]*(a|e|i)r(é|ás|á|emos|éis|án)#	1.151	1.976	2.045	1.660	1.269	3.551	0.371	1.279
#ir\sa\s[a-z]*(a|e|i)r#	0.161	0.027						0.120

Integration of LETRAS and NUMEROS



Statistical analysis with NUMEROS: factor concentration

Pattern	Soria	Varios	Guadalajara	barcelona	Segovia	Barcelona	Madrid	Pirineo
#ir\sa\s[a-z]*(aleli)r#					13	3	25	
#(voy\vas\va\vamos\vái s\van)\sa\s[a-z]*(aleli)r#	14	13	7		202	48	2530	11
#[a-z]*(aleli)r(élás\álemos\éis\án)#	3	3	2		93	32	1825	11

Column	Soria	Varios	Guadalajara	barcelona	Segovia	Barcelona	Madrid	Pirineo
Weight	-0.085	-0.067	-0.009	0.000	0.145	0.389	0.463	0.687

Cct. coef.	Value
Seq.mean dist.	68.581
Ref.mean dist.	36.496
Seq.corr.coef	0.113
Ref.corr.coef	0.109
mean linkage index	161.970
Std.Union.c.	0.500
Cramer c.	0.106

NUMEROS-web

- the **statistical counterpart of LETRAS-web** designed for performing quantitative analyses on the results from LETRAS, but also new data uploaded by the user:
  - around **80 different operations**, from basic median calculation to complex analyses based on matrix manipulations.
  - special interest: concentration analysis by correspondence, distance or factor.

Samples: combination of preffixes and suffixes to form medical terms (1)

- Distance with respect to zero point

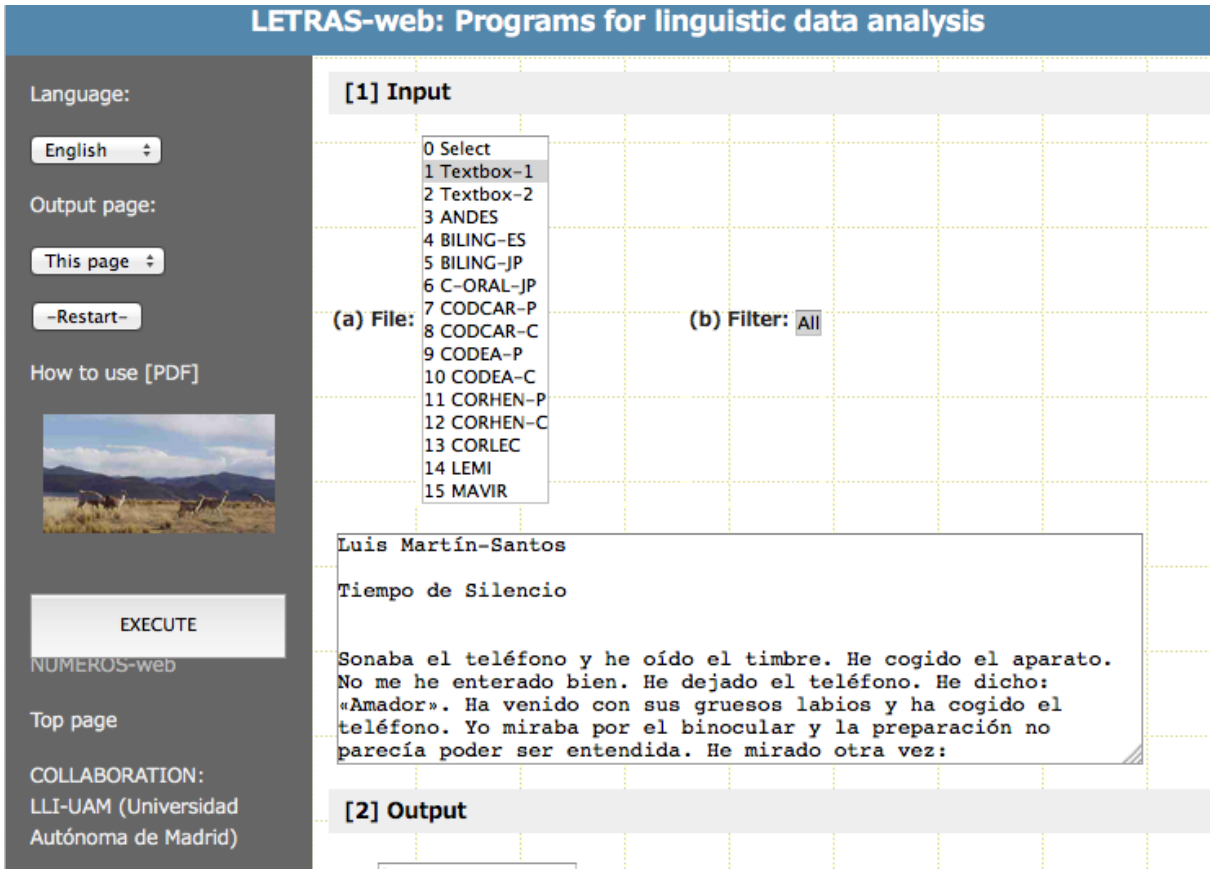
Dst.cct.	-algia	-génesis	-itis	-malacia	-osis	-oma	-tomía	-cito	-oide	-cele	-blasto	-patía	-megalia	-tóxico	Value
cefal(o)-	1														1,00
cerebr(o)-			2		1										3,79
dermat(o)-			4		7	2									4,67
arteri(o)-		1	2		4		1					1			5,92
artr(o)-	3		3		2		1					2			6,16
oste(o)-		2	3	1	8	5	1	1	1		2	2			6,67
neum(o)-			1		2	1	1	1				1			7,09
tiroid(o)-			1				1					1			8,21
miel(o)-			1		2	2		1	1	1	1		1	1	8,85
hepat(o)-			2		1	3	1	1			1	1	2	1	9,04
hem(o)-					2		1		1		2	3			9,74
cardi(o)-					1							2	1	1	11,64
ocul(o)-												1			12,00
Value	4,36	5,42	5,86	6,00	6,36	7,39	7,66	8,15	8,91	9,00	9,08	9,15	10,31	10,41	

Samples: combination of preffixes and suffixes to form medical terms (2)

- Cluster analysis

Dst.cct.	-algia	-blasto	-oide	-cito	-oma	-génesis	-malacia	-itis	-osis	-patía	-tomía	-cele	-megalia	-tóxico	Value
arteri(o)-						1		2	4	1	1				
dermat(o)-					2			4	7						,06
oste(o)-		2	1	1	5	2	1	3	8	2	1				,14
neum(o)-				1	1			1	2	1	1				,26
artr(o)-	3							3	2	2	1				,49
cerebr(o)-								2	1						,67
hepat(o)-		1		1	3			2	1	1	1		2	1	1,02
miel(o)-		1	1	1	2			1	2			1	1	1	1,23
cardi(o)-									1	2			1	1	1,65
ocul(o)-										1					1,78
hem(o)-		2	1						2	3	1				1,97
tiroid(o)-								1		1	1				2,27
cefal(o)-	1														2,84
Value		,55	,61	,86	,97	1,21	1,26	1,58	1,71	2,11	2,31	2,81	3,09	3,13	

Try your own text



Summary

- An **integrated toolbox** for concordancing and statistical analyses
- It is a **free** resource
- It is **web-based**
- It provides access to small but **curated corpora in different languages** (Spanish, English, Japanese).
- **Sophisticated statistics** as in propietary software (SPSS)

Conclusion

- A useful tool **aimed to linguists and philologists** who are looking for a free and on-line software for studying their corpora or the pre-loaded texts.
- **You DON'T have to register** or download/install software.

Future work

- Adding more corpora: ANYONE who wants to contribute with his/her texts are welcome.
- Handling tagged corpora: integration of some POS taggers in the tool.
- Handling graphics

→ CICL13 was a very successful experience for the authors!!

Contact information

- LETRAS-web:
  - http://lecture.ecc.u-tokyo.ac.jp/~cueda/letras/
- NUMEROS-web:
  - http://lecture.ecc.u-tokyo.ac.jp/~cueda/numeros/

- Hiroto UEDA: University of Tokyo
  - hiroto.ueda.tokio@gmail.com
- Antonio MORENO-SANDOVAL: Autonomous University of Madrid
  - antonio.msandoval@uam.es