



Lexicography in the grid environment

Villegas Marta, Bel Nuria, Bel Santiago,
Alemany Francesca & Martínez Héctor
(Universitat Pompeu Fabra, Spain)

Index

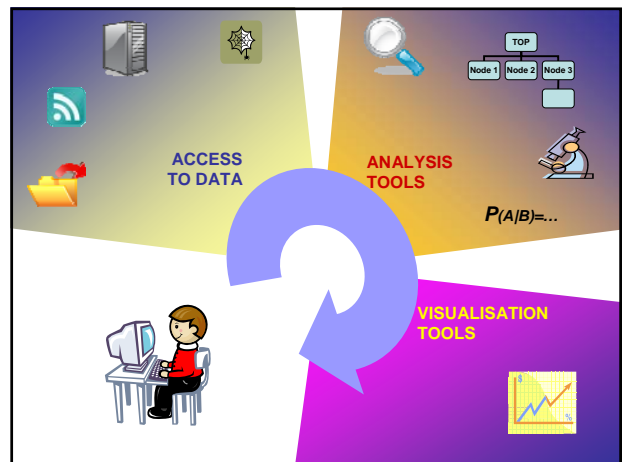


- Introduction
 - CLARIN
 - Use case: *Lexicography in a grid environment*
- Using corpus data
 - Manual classification
 - Problems
 - Goals
- Clustering by lexical similarity
 - Application structure
- Results and discussion

CLARIN



The CLARIN (*Common Language Resources and Technology Infrastructure*) is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable.



Use case scenario



Support for lexicographers creating a dictionary for Spanish as a second language.

Distributional hypothesis



“A word is characterized by the company it keeps”
(Harris, 1951; Firth, 1957)

Manual classification – raw corpus data



- [1] sea un poco más complicada ya que hay abundancia de construcciones asintéticas Es evidente que el autor
- [2] celulosa polisacárido No hay abundancia de vacuolas hay abundancia de vacuolas La pared nuclear es doble aunque
- [3] El Principat tiene como preocupación prioritaria * la abundancia de trigo y vino para el natural sustento
- [4] El Principat tiene como preocupación prioritaria * la abundancia de trigo y vino para el natural sustento
- [5] las formas mejor adaptadas a cada ambiente la abundancia de formas coloniales con lo que se asegura
- [6] concedido a l sector privado también corrobora la abundancia de liquidez En cuanto a los indicadores recogidos
- [7] alcanzar el mismo punto que Italia donde la abundancia de estrellas extranjeras no hiere las prestaciones de
- [8] el medio que los rodea por ejemplo la abundancia de plantas en un río o lago provoca
- [9] diversas La diversidad refleja las diferencias en la abundancia de las diferentes especies independientemente de cual sea
- [10] El análisis de las curvas de dominancia o abundancia relativa de cada especie en las muestras analizadas
- [11] explicaciones que ofrecían se centraron fundamentalmente en la abundancia relativa de factores los países tenderían a exportar
- [12] persisten y se refugian muchas especies pero la abundancia de las más es pequeña y contribuyen en
- [13] de cantidad de son inadecuadas para predecir la abundancia de los recolectores en el gradiente Por tanto
- [14] mercantilistas dicen o quieren decir es que la abundancia de dinero de oro y de plata produce
- [15] el delito de la misma manera que la abundancia de ladrones no convierte el robar en cosa
- [16] en segmentos de extensión aleatoria y que la abundancia de cada especie es proporcional a l espacio
- [17] de logaritmos de base dos para representar la abundancia de las especies puede ser cómoda En plantas
- [18] relación entre la tasa de multiplicación y la abundancia de las especies pero siempre que se habla

Manual classification – lexical patterns



- group 1**
- [3] El Principat tiene como preocupación prioritaria * la **abundancia** de trigo y vino para el natural sustento
- [4] El Principat tiene como preocupación prioritaria * la **abundancia** de trigo y vino para el natural sustento
- group 2**
- [10] El análisis de las curvas de dominancia o **abundancia relativa** de cada especie en las muestras analizadas
- [11] explicaciones que ofrecían se centraron fundamentalmente en la **abundancia relativa** de factores los países tenderían a exportar
- group 3:**
- [9] diversas La diversidad refleja las diferencias en la **abundancia** de las diferentes **especies** independientemente de cual sea
- [17] de logaritmos de base dos para representar la **abundancia** de las **especies** puede ser cómoda En plantas
- [18] relación entre la tasa de multiplicación y la **abundancia** de las **especies** pero siempre que se habla

Manual classification – raw corpus data



- group 1**
- [3] El Principat tiene como preocupación prioritaria * la **abundancia** de trigo y vino para el natural sustento
- [4] El Principat tiene como preocupación prioritaria * la **abundancia** de trigo y vino para el natural sustento
- group 2**
- [10] El análisis de las curvas de dominancia o **abundancia relativa** de cada especie en las muestras analizadas
- [11] explicaciones que ofrecían se centraron fundamentalmente en la **abundancia relativa** de factores los países tenderían a exportar
- group 3:**
- [9] diversas La diversidad refleja las diferencias en la **abundancia** de las diferentes **especies** independientemente de cual sea
- [17] de logaritmos de base dos para representar la abundancia de las especies puede ser cómoda En plantas
- [18] relación entre la tasa de multiplicación y la abundancia de las especies pero siempre que se habla

Project - Goals



- Clustering based on lexical similarity
- Language independence
 - No external resources
- Precision over recall
- Overall usefulness
- Focus on large data due to grid environments

Problems I



Form	Occurrences ¹	Senses ²
<i>Puente</i>	6,066	15
<i>Prueba</i>	15,987	14
<i>Camino</i>	29,161	8
<i>Sentido</i>	50,891	11

¹ Data from the *Corpus de Referencia del Español Actual* (CREA)

² Senses prescribed by the Diccionario de la Real Academia d

Problems II

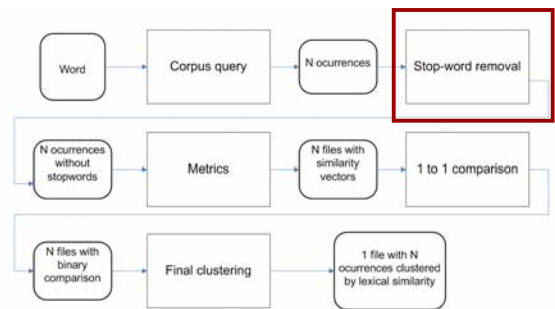


- Data sparseness
- Small context size
- How to measure lexical similarity?

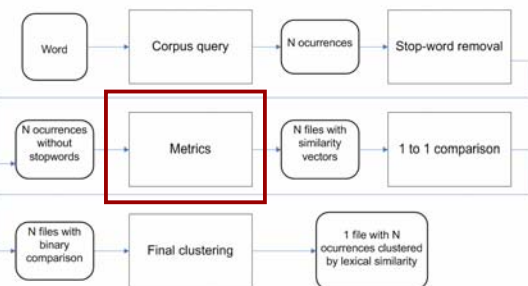
Possible technical approaches

- Latent Semantic Analysis & Bag of Words
 - No order information, need lots of data
- Text::Similarity (Pedersen et al.)
 - *WordNet*-based
- Fuzzy Matching (WER, etc)
- BLEU
 - Scoring system for automatic translation

System workflow I



System workflow II



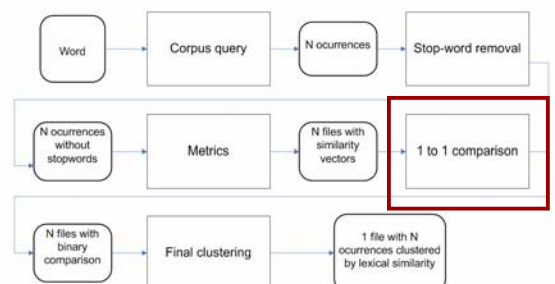
Metrics I

- Order-dependent
 - Hamming
 - Equal element
- Set operations
 - Jaccard
 - Coverage index

Metrics II - stemming

- *y · han · form · ado · una · asociaci · ón · de · comerciantes · de · hierr · o · una · socie · dad · de · responsa · bilidad · limit · ada*
- *no · es · completa · mente · intercambia · ble · por · ejempl · o · hierr · o · de · chatarr · a · compar · ado · con · lingote · de · hierr · o · entonces · habrá · lug · ar*

System workflow III



1 to 1 comparison

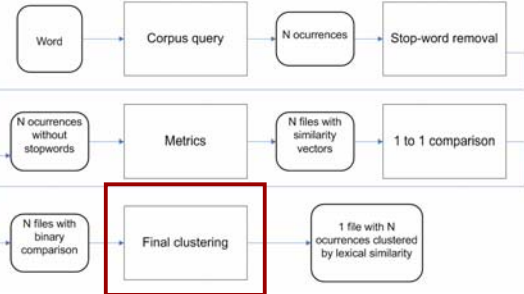


The number of 1-to-1 comparisons is $(N^2/2) - N$.

For a corpus query with 1,000 occurrences,
499,000 comparisons are necessary.

For $N=50,000$, we need 1,249,950,000.

System workflow IV



Sample output



group_60:
[98] de dedicar una columna entera a cantar la gallina y disculpar se Luego con ironía ha seguido

group_61:
[99] y hasta en Adolfo Suárez que cantó la gallina el viernes declarando que nunca hubo nadie mejor

group_62:
[115] salir la metáfora de l juego de la gallina ese a l que jugaba el personaje de

group_63:
[88] de la prioridad de l huevo o la gallina En el hemisferio Norte tanto por la dirección
[95] terrorista Todo se andará El huevo y la gallina Antonio Eiorza Algunas de las críticas que me

group_64:
[116] patio de l colegio para jugar a la gallina ciega mientras la invidente oposición se apresura a
[117] patio de l colegio para jugar a la gallina ciega mientras la invidente oposición se apresura a

group_65:
[119] Gobierno ha resuelto aplicar el juego de la gallina ciega a la educación secundaria y tal vez
[118] de gente Sin embargo este juego de la gallina ciega en el que se persigue la obtención
[121] Gobierno ha resuelto aplicar el juego de la gallina ciega a la educación secundaria y tal vez

Conclusions I – evaluation against gold standard



	Accurac y	Coverag e	Occurren s	Groups
Melanoma	1,000	0,411	38	38+0
Alzamiento	1,000	0,562	41	41+0
Chatarra	1,000	0,703	71	71+0
Gancho	0,899	0,690	113	65+20
Gallina	0,904	0,685	125	74+14
Tambor	0,787	0,682	127	60+15
Dama	0,765	0,384	208	90+23

Conclusions II - Underspecificity



A	retación	entre	la	ex	primera	dama	hipina	y
A	presidencial	en	La	ex	primera	dama	y	senadora
A	paz	Claude	Pompidou	ex	primera	dama	de	Francia
A	la	senadora	y	ex	primera	dama	arrasaría	Tiene
A	l	recinto	La	ex	primera	dama	de	Francia
A	comparecencia	de	la	ex	primera	dama	en	la
A	está	presente	la	ex	primera	dama	de	este
B	Johnson	entre	ellos	la	primera	dama	estadounidense	Laura
B	su	portavoz	Aunque	la	primera	dama	estadounidense	no
C	el	rival	de	la	primera	dama	Hillary	Clinton
C	le	enfrenta	a	la	primera	dama	Hillary	Clinton
D	oscuros	negocios	de	la	primera	dama	peruana	La
D	oscuros	negocios	de	la	primera	dama	peruana	BUSCADOR
E	en	Hillary	Clinton	una	primera	dama	tradicional	con
E	de	Hillary	Clinton	La	primera	dama	asegura	que
F	cuestionada	candidatura	de	la	primera	dama	En	total
F	la	candidatura	de	la	primera	dama	Es	el
G	l	Gobierno	incluida	la	primera	dama	tenían	millonarios
G	yo	quiero	ser	primera	dama	se	manifestaron	
G	viajes	y	actividades	como	primera	dama	La	prensa
G	a	l	Senado	La	primera	dama	no	se
G	adquiriera	el	rango	de	primera	dama	adelantando	en
G	escaño	a	pulso	la	primera	dama	ha	ganado
G	Elkane	Karp	de	Toledo	primera	dama	de	Peru

Conclusions III – use case results



- Overall usefulness
- Precision over recall
- Generalization i.e. underspecificity
- Linguistic Portability
 - Stoplist
 - Suffix list



- Workflow editing
 - Changing modules on the fly
- Distributed networking
 - Deal with larger amounts of data
- System integration
 - WS architecture
- User-layer interface



Thank you!

for more information:
<http://www.clarin.eu>