RECURSOS LÉXICOS Y GRAMÁTICAS PARA RECUPERACIÓN DE INFORMACIÓN

Tecnologías Pregunta-Respuesta

Roberto Maestre Martínez

@rmaestrem @paradigmalabs

INDICE

Pasos básicos del Text Mining

Raw text

NER

Gramáticas

Information extraction

NAME ENTITY RECOGNITION

Reconocer y extraer menciones de entidades relevantes en el texto

"El presidente Barack Obama es el presidente de los Estados Unidos de América"

NAME ENTITY RECOGNITION

Enfoque basado en diccionarios

El objetivo es hacer matching de entidades relevantes utilizando diferentes recursos: ontologías, thesauros, ...

Este enfoque es muy sencillo, pero es sensible a errores de ortografía, master/MSc , etc ...

Dependiendo de la cobertura deseada puede ser un problema crear los diccionarios.

NAME ENTITY RECOGNITION

Enfoque basado en diccionarios

¿Como crear un diccionario para detectar los nombres de compañías de telefonía que operan en España?

¿Como crear un diccionario para detectar actores de television que tenga una covertura lo mas grande posible?

NAME ENTITY RECOGNITION

Enfoque basado en diccionarios

https://api.freebase.com/api/service/mqlread?query={"query":%20{"type":%20"/tv/tv_actor",%20"id":%20[]}}

```
import urllib
import json

# Query parameters
query = [{"type": "/tv/tv_actor", "id": []}]
query_envelope = {'query': query}
# Service url
service_url = 'http://api.freebase.com/api/service/mqlread'
url = service_url + '?query=' + json.dumps(query_envelope)
# Perform request
response = json.loads(urllib.urlopen(url).read())
# Read results
for actor in response['result']:
    print actor['id']
```

NAME ENTITY RECOGNITION

Enfoque basado en diccionarios

Para resolver el problema de los errores ortográficos podemos utilizar algún enfoque basado en distancias de palabras, e.g.: *Probabilistic Term Variant Generator for Biomedical Terms**

http://www.mendeley.com/research/a-comparison-of-string-metrics-for-matching-names-and-records/

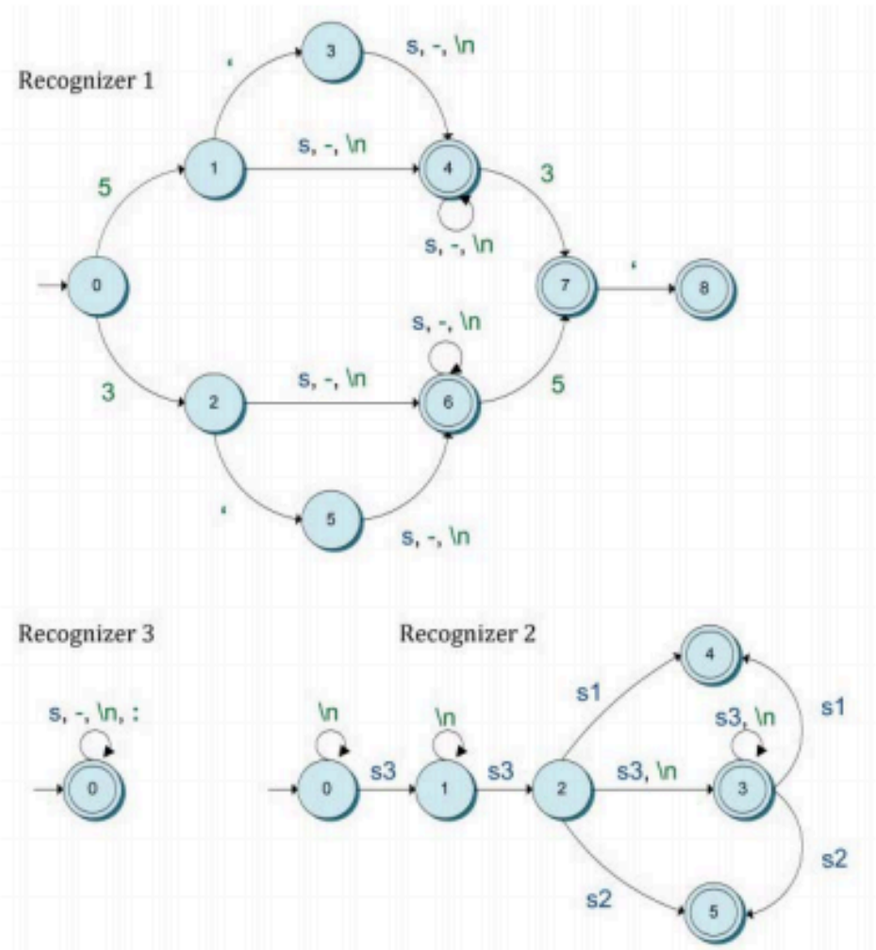
http://labs.paradigmatecnologico.com/2011/07/06/information-propagation-in-twitters-network/

$$DL(barack, barak) = 1$$
$$DL(barack, varaka) = 3$$
$$DL \leq 1$$

* Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2003. Probabilistic term variant generator for biomedical terms. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (SIGIR '03). ACM, New York, NY, USA, 167-173. DOI=10.1145/860435.860467 http://doi.acm.org/10.1145/860435.860467

NAME ENTITY RECOGNITION

Ruled based

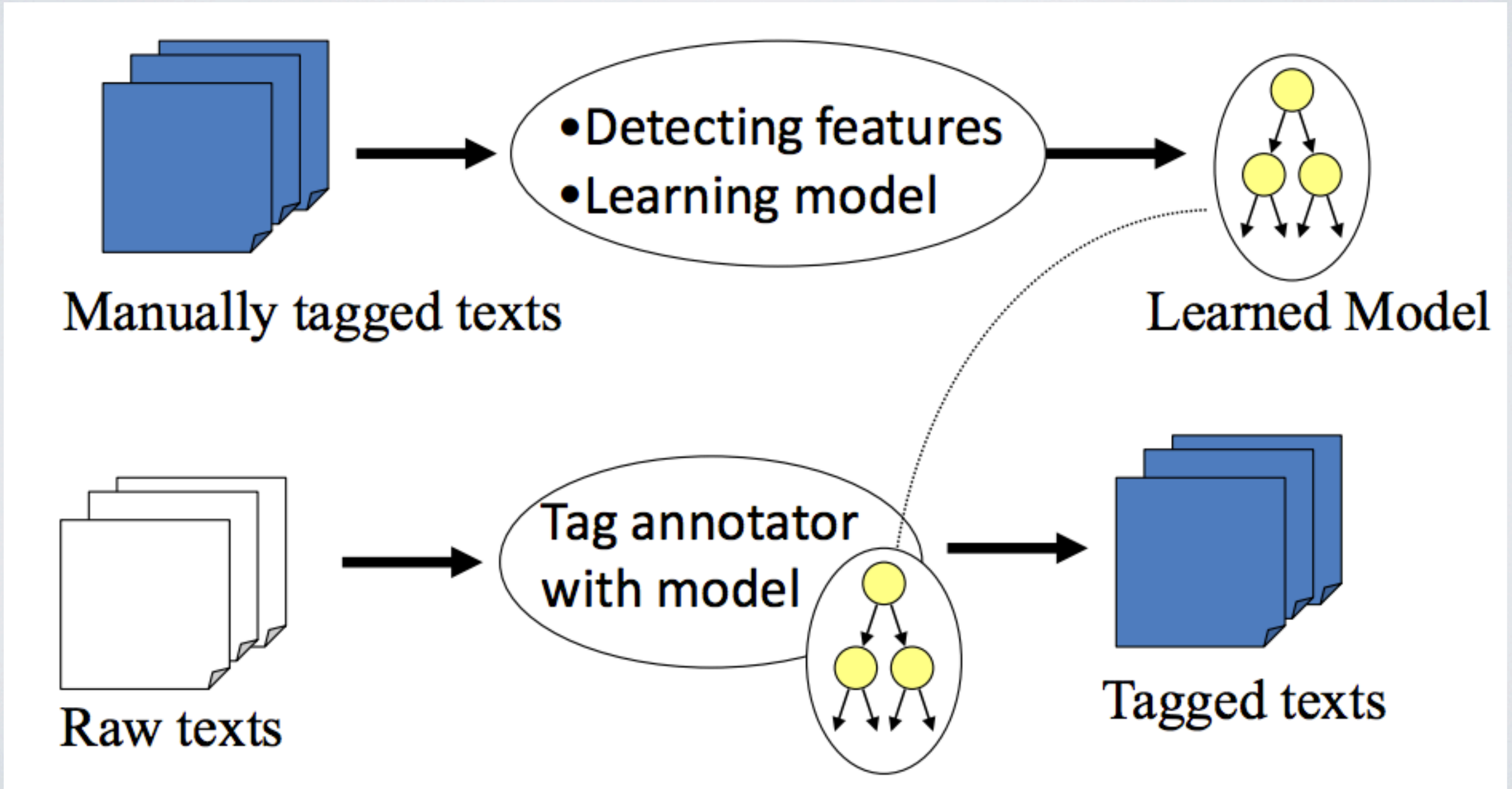


Detector	PMID	Text String	List of Tokens
1	19781080	...primers AA247 (5'-TGCCATTGCCAAAGAGAC-3') and pLQ510-rp1...	{"TGCCATTGCCAAAGAGAC"}
1	19664269	...mecA gene, mecAR (5'-TTACTCATGCCATACATAAATGGATA-\nGACG-3') and mecAF...	{"TTACTCATGCCATACATAAATGGATA", "GACG"}
1	19379498	...specific primer pair traD-F (5'-caatgcttgatctatttggtag-3') and traD-R...	{"caatgcttgatctatttggtag"}
1	19758438	...MY 09, 5-CGT CCM\n ARR GGA WAC TGA TC-3; where M = A/C, W = A/T...	{"CGT", "CCM", "ARR", "GGA", "WAC", "TGA", "TC"}
2	19799780	B-globin outside R @ CTC AAG TTC TCA GGA TCC A @ 1st round PCR primer for Human Beta globin DNA	{"CTC", "AAG", "TTC", "TCA", "GGA", "TCC", "A"}
2	18847469	btherm @ GAT GTG CCG GGC TCC TGC ATG @ This study	{"GAT", "GTG", "CCG", "GGC", "TCC", "TGC", "ATG"}
2	18154687	Stx1 @ GTA CGT CTT TAC TGA TGA TTG ATA GTG GCA CAG GG @ 35 @ 73.5	{"GTA", "CGT", "CTT", "TAC", "TGA", "TGA", "TTG", "ATA", "GTG", "GCA", "CAG", "GG"}
2	19558693	...are listed below.\n EP1- F ATG GTG GGC CAG CTT GTC\n EP1- R...	{"ATG", "GTG", "GGC", "CAG", "CTT", "GTC"}
3	19754958	...with primer N309 (ACATGCGGATCCCTCGAGCCTTTGAA-\nGATGACTAACTCCCCA) and N297...	{" ACATGCGGATCCCTCGAGCCTTTGAA", "GATGACTAACTCCCCA"}
3	19737401	...and 3' AAGCT TGGTA CCTCA CTGCA\nGCAGA GCGCT GAGGC CCAGC AGCAC. The resulting PCR...	{"AAGCT", "TGGTA", "CCTCA", "CTGCA", "GCAGA", "GCGCT", "GAGGC", "CCAGC", "AGCAC"}
3	19149882	1 @ XAC0340 @ 432 @ gATACCCCATATgAATgCgAT	{"gATACCCCATATgAATgCgAT"}
3	19775435	20 @ F:GAGATGGATTAACCAGATGTCTTAAAAACTATCGTAAC	{":","GAGATGGATTAACCAGATGTCTTAAAAACTATCGTAAC"}

García-Remesal M, Cuevas A, López-Alonso V, López-Campos G, de la Calle G, de la Iglesia D, Pérez-Rey D, Crespo J, Martín-Sánchez F, Maojo V. A method for automatically extracting infectious disease-related primers and probes from the literature. BMC Bioinformatics. 2010 Aug 3;11:410

NAME ENTITY RECOGNITION

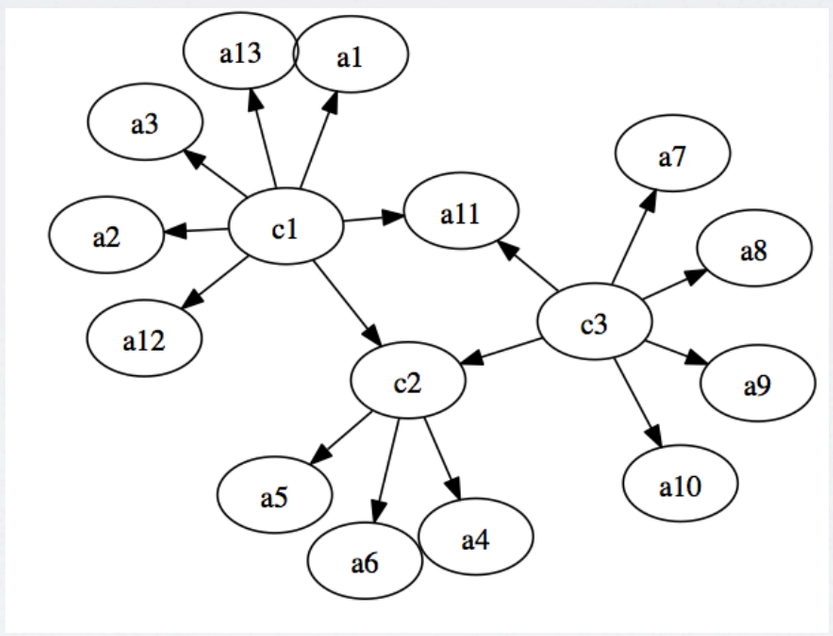
Machine learning based



NAME ENTITY RECOGNITION

Paradigma labs. Wikipedia NER

Usar las relaciones entre categorias, subcategorias y artículos



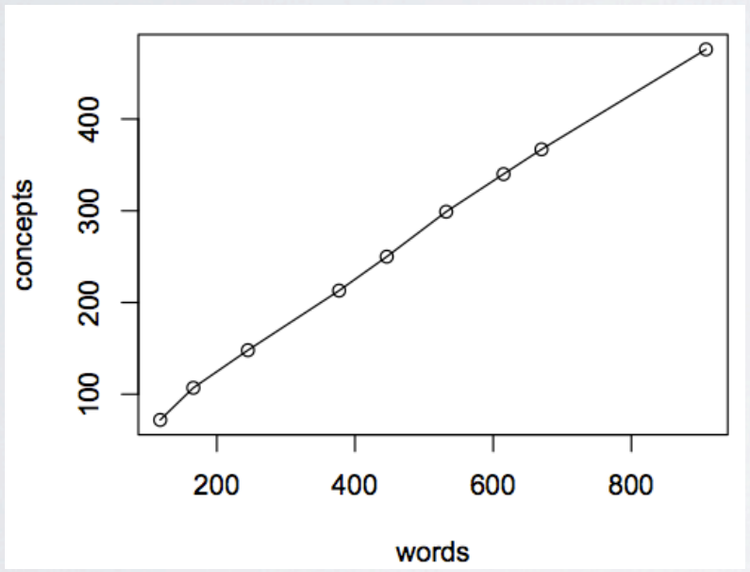
https://twitter.com/#!/rmaestrem/status/162656378555596801/photo/1

Networkx (20%) + Kyoto cabinet (80%)

NAME ENTITY RECOGNITION

Paradigma labs. Wikipedia NER

Podemos detectar 25.000.000 de categorias



Podemos utilizar los enlaces entre categorias

```
{
  - clusters: {
      - mercado financiero: {
            mariano rajoy: 7,
            jose luis rodriguez zapatero: 7,
            politicos: 6,
            crisis economica: 6
        },
      - actimel: {
            yogur: 3,
            danone: 2,
            lanjaron: 4
        }
    },
    sse_error: 3
}
```

NAME ENTITY RECOGNITION

Paradigma labs. Wikipedia NER

- La desambiguation la hacemos en base al "surface text"

$$E(w) = -\sum_{i \in R_s} P(r_i|w) \log P(r_i|w)$$

Zapatero va a arreglar mis zapatos

Zapatero

El término ***zapatero*** puede referirse a:

- cualquier cosa perteneciente o relativa a los zapatos;
- la profesión de zapatero, consistente en fabricar, arreglar o vender zapatos;
- el mueble zapatero, un mueble para guardar zapatos.

Personas con el apellido Zapatero:

- José Luis Rodríguez Zapatero (1960 -), presidente del Gobierno de España (2004 -);
- Virgilio Zapatero Gómez (1946 -), ministro español de Relaciones con las Cortes (1986 - 1993);
- Luis Arroyo Zapatero (1951 -), rector español de la Universidad de Castilla-La Mancha (1988 - 2003);
- Ismael Piñera Zapatero (1977 -), futbolista español;
- Carlos Arroyo Zapatero (1964 -), arquitecto español;
- Gonzalo Ruiz Zapatero (1954 -), catedrático español en Prehistoria.

Personas conocidas con el sobrenombre de Zapatero:

- Teódoto el Zapatero (finales del s. II), escritor cristiano de Bizancio;
- Simón el Zapatero (s. IV - s. III a.C.), discípulo directo de Sócrates.

```
<result name="response" numFound="20" start="0" maxScore="0.8154754">
  <doc>
    <float name="score"> 0.8154754 </float>
    <str name="from"> zapatero </str>
    <str name="texto"> profesion consistente fabricar arreglar vender zapatos </str>
    <str name="to"> profesion de zapatero </str>
  </doc>
  <doc>
    <float name="score"> 0.63237333 </float>
    <str name="from"> zapatero </str>
    <str name="texto"> mueble guardar zapatos </str>
    <str name="to"> mueble zapatero </str>
  </doc>
  <doc>
    <float name="score"> 0.31618667 </float>
    <str name="from"> zapatero </str>
    <str name="texto"> zapato cualquier cosa perteneciente relativa s </str>
    <str name="to"> zapato </str>
  </doc>
  <doc>
    <float name="score"> 0.0 </float>
    <str name="from"> zapatero </str>
    <str name="texto"> jose luis rodriguez 1960 presidente gobierno españa 2004 </str>
    <str name="to"> jose luis rodriguez zapatero </str>
  </doc>
```

Mendes, P. y Jakob, M. y García-Silva, A. y Bizer, C. (2011) *DBpedia Spotlight: Shedding Light on the Web of Documents.* In: 7th International Conference on Semantic Systems, 7–9 September 2011, Graz, Austria.

POS TAGGING

NLTK

http://www.nltk.org/

```
import nltk

###
#   Simple example of tokenize and POS tag with NLTK
###
text = "All that is gold does not glitter. not all those that wander
are lost."

sentences = nltk.sent_tokenize(text)

for sent in sentences:
    tokens = nltk.word_tokenize(sent)
    pos_tags = nltk.pos_tag(tokens)
    print pos_tags
```

POS TAGGING

http://nlp.lsi.upc.edu/freeling/demo/demo.php

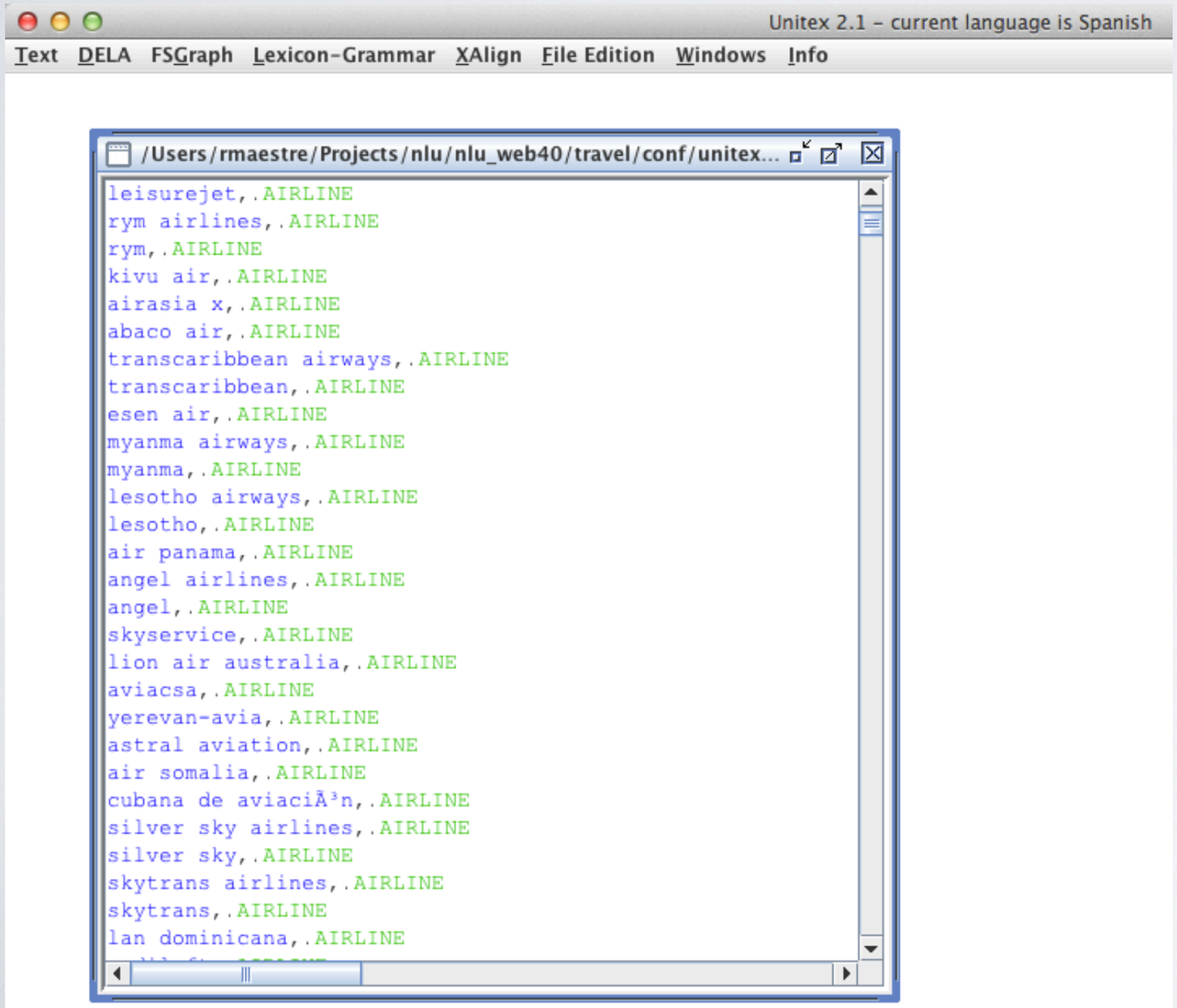
El gato come pescado y bebe agua.

El	gato	come	pescado	y	bebe	agua	.
el DA0MS0 1	*gato* NCMS000 1	*comer* VMIP3S0 0.75	*pescado* NCMS000 0.833333	*y* CC 0.999812	*beber* VMIP3S0 0.994868	*agua* NCCS000 0.973333	*.* Fp 1
		comer VMM02S0 0.25	*pescar* VMP00SM 0.166667	*y* NCFS000 0.000188324	*beber* VMM02S0 0.00513196	*aguar* VMIP3S0 0.0133333	
						aguar VMM02S0 0.0133333	

UNITEX

http://igm.univ-mlv.fr/~unitex/UnitexManual2.1.pdf

Diccionarios.

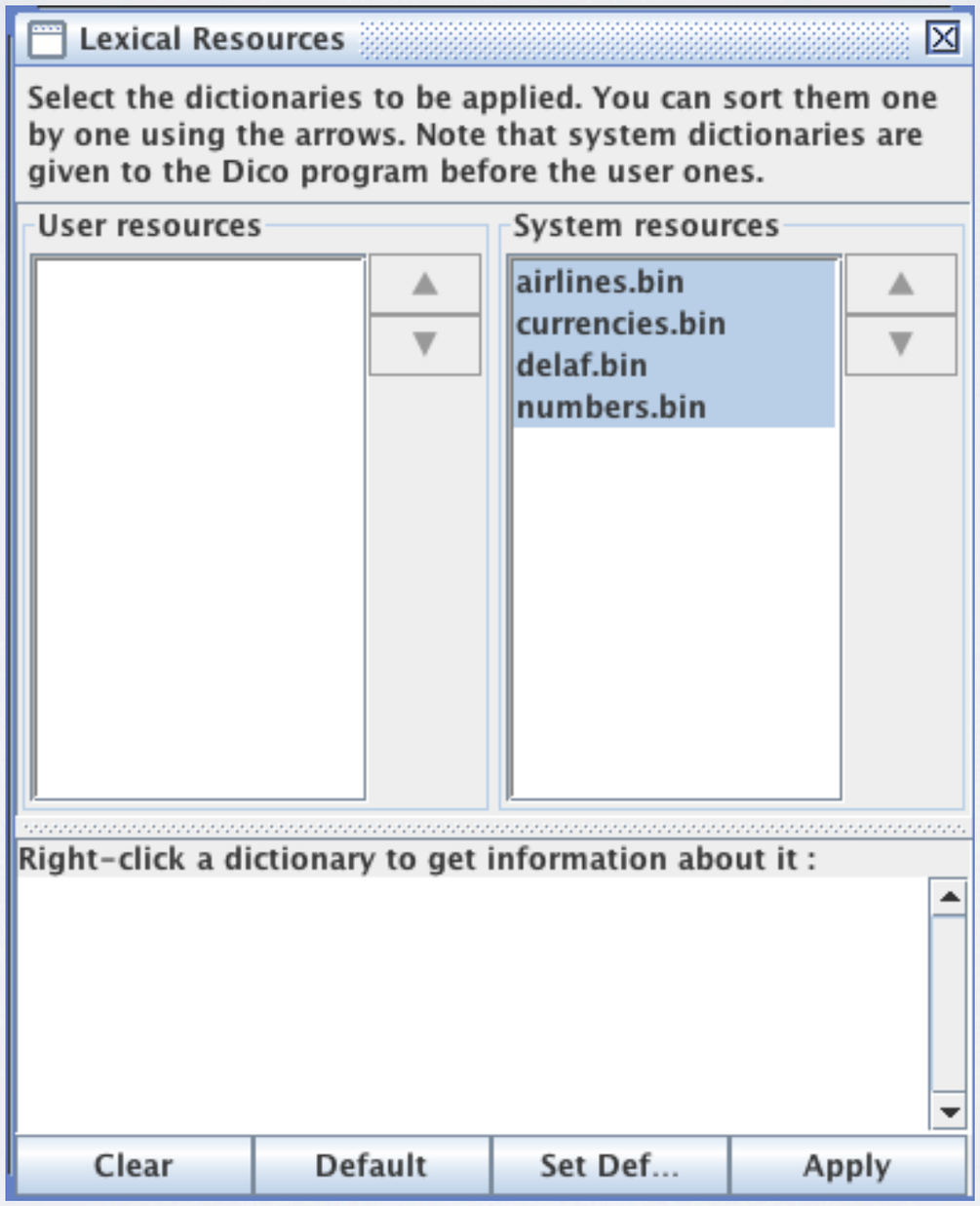


Crear 3 diccionarios

Numbers
Currency
Airlines

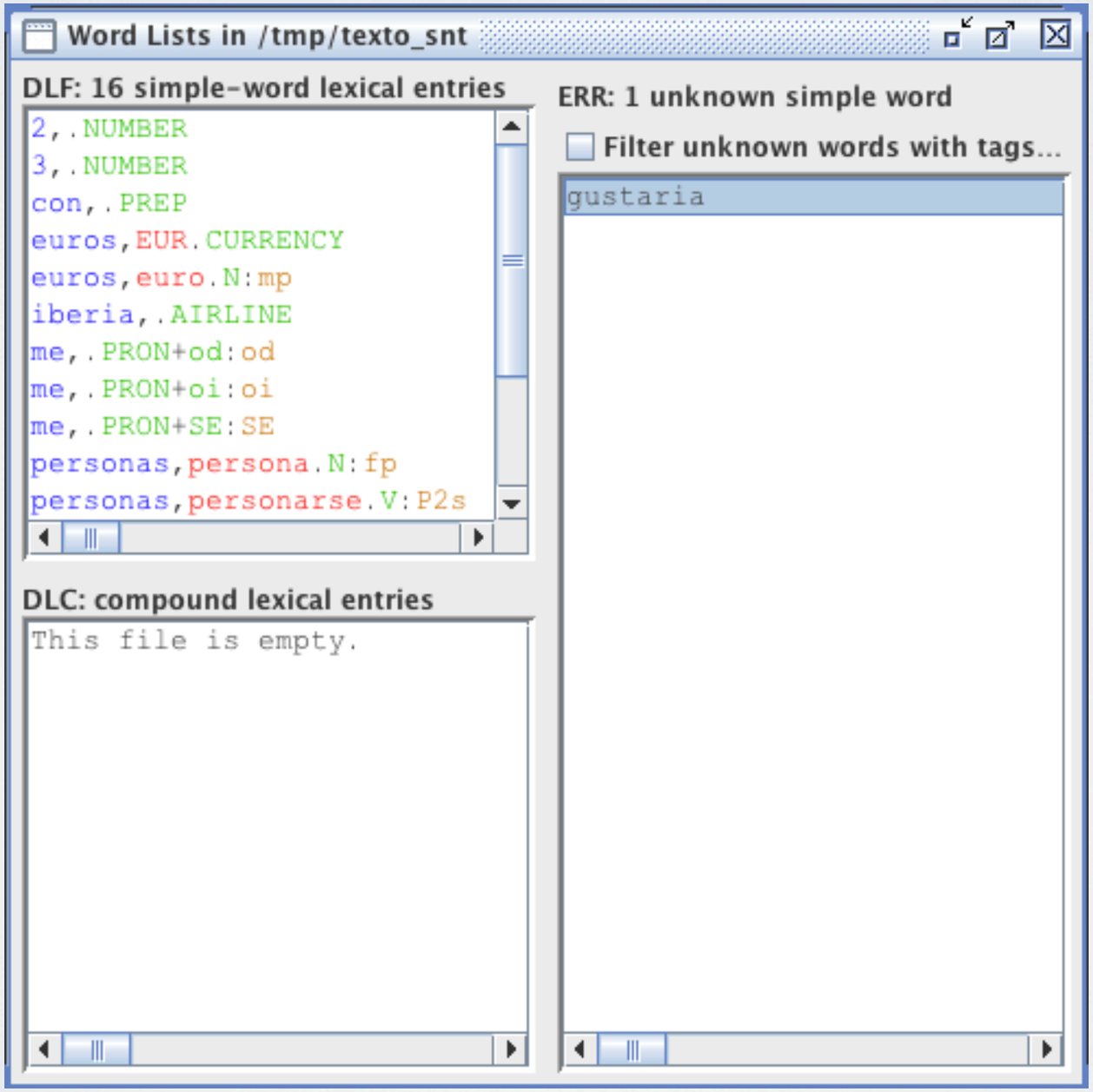
UNITEX

Aplicar recursos.



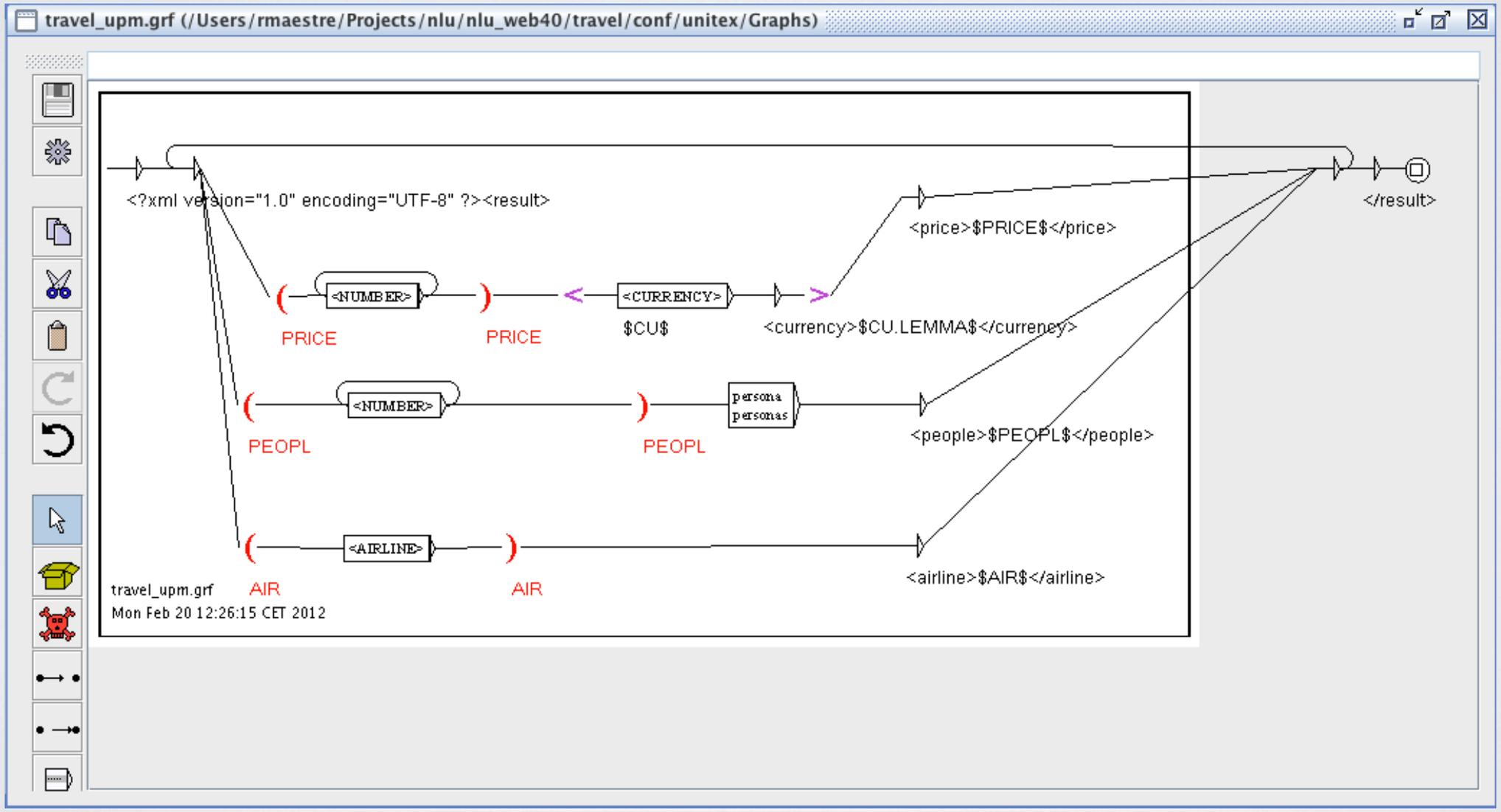
UNITEX

Resultado.



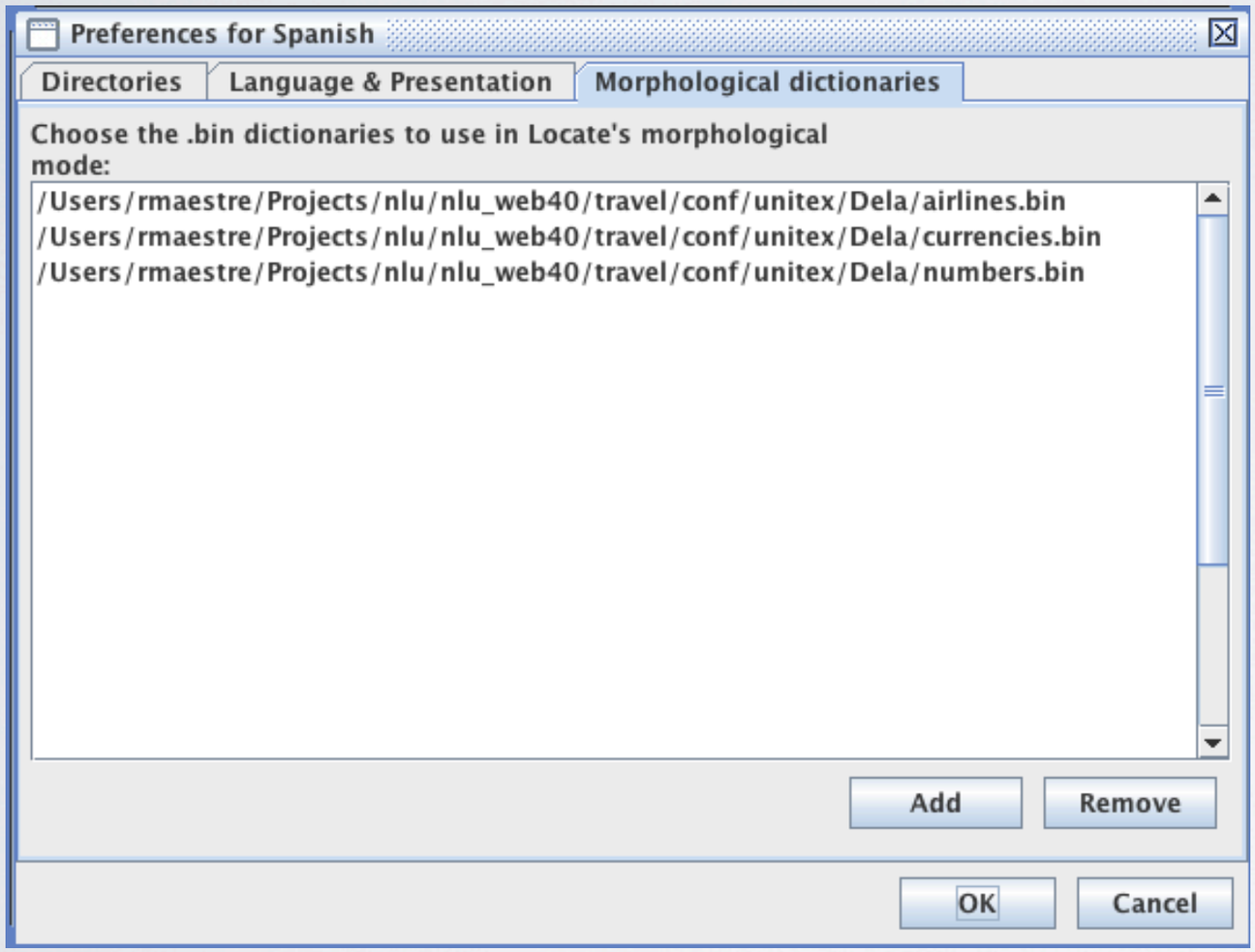
UNITEX

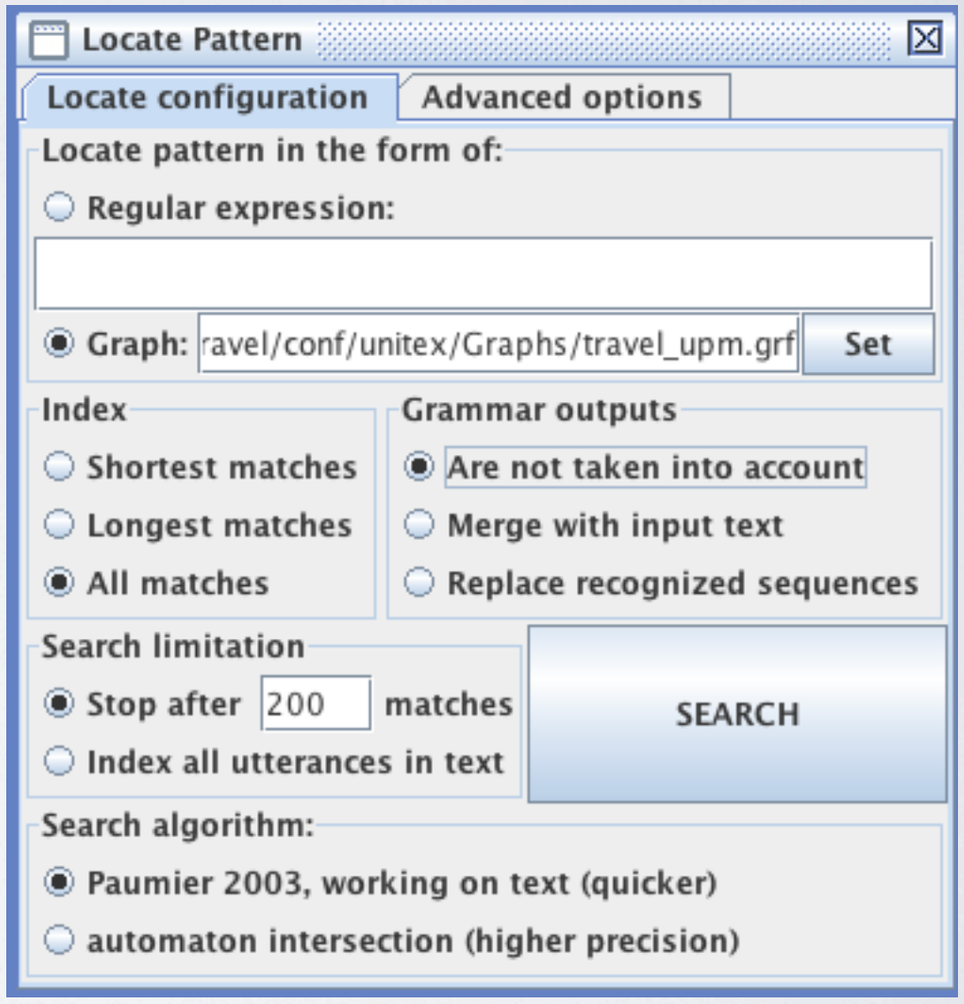
Diseño de la Gramática.



GRAMATICAS

Aplicar gramática






```
import pyunitex
u = pyunitex.Unitex()
u.Convert('-s', 'UTF8', 'file_name', '-o', 'file_name_converted')
```

https://github.com/moliware/pyunitex

RECURSOS LÉXICOS Y GRAMÁTICAS PARA RECUPERACIÓN DE INFORMACIÓN

Tecnologías Pregunta-Respuesta

Muchas gracias

Roberto Maestre Martínez

@rmaestrem @paradigmalabs