



**Extending an Information Extraction Tool Set  
to Central and Eastern European Languages**

**'Information Extraction for Slavonic and other  
Central and Eastern European Languages'**  
Workshop held at RANLP Conference  
8 September 2003



**Camelia Ignat, Bruno Pouliquen, António Ribeiro & Ralf Steinberger**

**Joint Research Centre, Ispra, Italy**  
<http://www.jrc.it/langtech>





**Agenda**

- Introduction: aim and focus of our work
- (Character sets: representation and recognition)
- Recognition of date expressions
- Place name recognition and visualisation

 **Introduction**  EUROPEAN COMMISSION  
DIRECTORATE-GENERAL  
Joint Research Centre


- Aim of our work: multi-component system for
  - Gathering of Documents
  - **Text Analysis**
  - **Visualisation of Contents**
- Focus:
  - Multilingual applications
  - Cross-lingual applications
- Small team, many languages
  - ➔ use linguistics-poor approaches / statistics

 **Character Sets**  EUROPEAN COMMISSION  
DIRECTORATE-GENERAL  
Joint Research Centre


- Most languages can be represented with various character sets (ASCII, ISO-8859-1, UTF-8, UCS-2, ...)

ISO-8859-1	most Western European languages	} UTF-8
ISO-8859-2	Central and Eastern European Languages	
ISO-8859-5	Cyrillic	
ISO-8859-7	Greek	
...		


- ➔ **Recognise character set** used in document and **convert to UTF-8**




## Character Set Recognition




- UTF-8 representation of letter    ü            195 + 188
  - ISO-8859-1 (Latin-1):            Ä            ¼
  - ISO-8859-2 (Latin-2):            Ā            Ž
  - ISO-8859-5 (Cyrillic):            Ч            М
  - ISO-8859-7 (Greek):            Γ            Ο
  - ...
  
- We use Ted Dunning's (1994) Markov Model for character set and language recognition (based on bigram and trigram distributions)
  
- ➔ Convert text to UTF-8 encoding




## Difficulties with Character Encoding




- Various software and programming languages have different inherent encoding:
  - Oracle 8i = UTF-8 compatible
  - JAVA uses UCS-2
  - Perl only deals with UTF-8 since version 5.8 (many difficulties)
  - ...
  
- Conversion is needed when they 'talk' to each other.
  
- Is the world ready for UTF-8 ???



## Recognition of Dates




- **Task:**
  - Recognise dates in various formats in running text
  - Represent in normalised form: DDMMYYYY
  - Store normalised form + onset + length in database
  
- **Advantage:**
  - Highlight in text for easier retrieval
  - Possibility to search for documents mentioning a date within a date range



## Example for Date and Place Recognition


Romanian text




**Armistițiu**

In fiecare an, la 11 noiembrie, in Franta si in Marea Britanie se celebreaza Armistițiu prin care s-a pus capat primului razboi mondial. In Franta este chiar zi oficiala de sarbatoare, jour férié, nu se lucreaza; in Marea Britanie, comemorarea are loc in cea mai apropiata duminica de data de 11 (se si cheama Remembrance Sunday, Duminica Aducerii-aminte, s-a tinut alaltaieri).


Nu se sarbatoreste nimic in sa in Germania, ceea ce pina la un punct este explicabil. Armistițiu, care a fost semnat la 11 noiembrie 1918 intr-un vagon-restaurant prefacut cu acel prilej in sala de reuniune, a parafat capitularea Germaniei. Acel vagon, aflat pe o linie secundara a garii satului Rethondes, in padurea linga orasul Compiègne, avea sa fie simbolic folosit si in 1940, de aceasta data pentru semnarea armistițiuului care consfintea infringerea Frantei in fata Germaniei lui Hitler.




## Current Coverage of Date Formats Recognised



Type	Description	Dates recognised, e.g.	Dates not recognised, e.g.
Complete absolute dates	Numerical;  Containing month name (full or abbreviated)	<i>3-04-03 or 21.2.1983 or 1997/04/01 1999, the 2nd of May the sixth of March in the year nineteen eighty four</i>	<i>1.2.15 7 May 2003 in the period expression 7-8 May 2003</i>
Incomplete absolute dates	Containing month name (full or abbreviated)	<i>third February Jan. 2003</i>	<i>incomplete numerical dates: 1990 ; the 1970s; two thousand and two simple month name: in May</i>
Relative dates	Relative to a reference day; Relative month names; Month name + relative year	<i>yesterday, today, tomorrow next June, last September, February last year</i>	<i>last month; next Summer; Labour Day; on Tuesday; in the third quarter; February three years ago</i>



## Modular Program





- **Language-independent PERL code**  
+ one **language-specific parameter file** per language
  
- **Containing lists of:**
  - Names of days
  - Names of months
  - Year numbers
  - Relative date expressions (e.g. *tomorrow*)
  - Pre- and post-modifiers (e.g. *last, next, this June*)
  - Words allowed as part of date expression  
3<sup>rd</sup> of June  
trois juillet de l'année 2003

plus, where appropriate:

- abbreviations;
- with and without diacritics
- inflections (e.g. case)
- ...

- **Current language coverage:**  
English, French, German, Spanish, Italian, Portuguese, Romanian





## Recognition Procedure (1)

**1) Detect complete dates** in numerical form, using a PERL regular expression

e.g. 31.5.2003, 13/02/03

**2) Identify full or abbreviated month names**, then search their context

e.g. trois juillet de l'année 2003  
last july




## Recognition Procedure (2)

**3) Identify relative time expressions** and resolve using **document reference date**


e.g. Reference date = 8 September 2003  
last December → 00 12 2002

**4) Disambiguation:** US format MDY vs. European format DMY

- Check number ranges:  
12/31/03 → 31 12 2003  
01/02/03 → ???
- Check further dates in text to detect **text standard**
- Default is set to DMY



## Evaluation (English)




- 510 KB of text; 87000 words;
- marked up with MUC codes
- 70% are incomplete dates that we do not cover
- **613** of the types we cover.


	<u>Precision</u>	<u>Recall</u>
Relative dates:	86%	67%
Complete dates:	100%	100%
Incomplete dates:	98%	98%

**Error analysis**

- this **may** sound → 00 05 0000
- **late January** / **mid-August**
- **7 – 8 June** → 08 06 0000



## Evaluation (Romanian)




- 582 news articles (Newspaper *Evenimentul zilei*)
- 1.6 MB of text
- **1031** date expressions


	<u>Precision</u>	<u>Recall</u>
	97.7%	98%

**Error analysis**

- **mai** întâi (mai: again/more) → 01 05 0000 (= En: firstly)
- cei **doi mai** incercasera → 02 05 0000 (= En: both already tried)
- **intre 12 si 26 iunie** → 26 06 0000 (= En: between 12 and 26 June)



**Place Name Recognition and Visualisation**



**EUROPEAN COMMISSION**  
DIRECTORATE-GENERAL  
Joint Research Centre

**Task:**

**Input:**  
1 text or a collection of documents.

**Output:**


- onset and length of geographical expression
- list of place names per text
- **countries they belong to**
- % with which each country is mentioned
- **size class information**  
(1=capital; 2 = major city; ... 6 = village)
- **geographical co-ordinates** for each place name
- **visualisation** of all place names in a map

**Armistițiu**


In fiecare an, la **11 noiembrie**, în **Franta** și în **Marea Britanie** se celebrează Armistițiul prin care s-a pus capăt primului război mondial. În **Franta** este chiar zi oficială de sărbătoare, zi feriată, nu se lucrează; în **Marea Britanie**, comemorarea are loc în cea mai apropiată duminică de data de 11 (se și cheama Remembrance Sunday, Duminica Aducerii-aminte, s-a ținut alaltaieri). Nu se sarbatoreste nimic însă în **Germania**, ceea ce pînă la un punct este explicabil. Armistițiul, care a fost semnat la **11 noiembrie 1918** într-un vagon-restaurant prefăcut cu acel prilej în sala de reuniune, a parafat capitularea **Germaniei**. Acel vagon, aflat pe o linie secundară a gării satului **Rechnowes**, în pădurea lîngă orășul **Compiègne**, avea să fie simbolic folosit și în 1940, de această dată pentru semnarea armistițiului care consfințea înfrîngerea **Francei** în fața **Germaniei** lui Hitler.

5 geographical references found in 10 words, Ratio = 50%


BG: 3 (60%): sofiya, rila, bulgaria  
IT: 2 (40%): ispra, italy  
Total: 5



*Sofiya and Rila are in Bulgaria. Ispra is in Italy.*




**Example: Visualisation of Place Names Recognised in RANLP Web Site Text**



**EUROPEAN COMMISSION**  
DIRECTORATE-GENERAL  
Joint Research Centre



geoplace LangTech/DMA/JRC

WORKSHOP on Information Extraction for Slavonic and other Cer


Visualisation using JRC's Digital Map Atlas (<http://dma.jrc.it>). ISFEREA project.




## Need for Gazetteers



- Unlike for people and companies, **there are few reliable textual clues.**
- ➔ Lists of place names are required.
- For a list of freely available resources, see Gey (2000)
  
- We use: *Global Discovery* commercial geographic database
  - Containing approx. 500 000 place names world-wide (we use only 85 000)
  - In local language and in English
  - Plus size information (classes 1 to 6)
  - Plus co-ordinates
  
- We added
  - country ISO codes
  - Currency names
  - Adjectives referring to countries and their people




## Challenges




Simple recognition via database lookup, but:


- Multiple place names have the same name:
  - 14 places called 'Paris'
  - 13 places called 'Roma'
  
- Place names have homographs with other words or with other types of names
  - 'And' (Iran)
  - 'Split' (Croatia)
  - 'Annan' (UK)
  
- Local variations of place names: Venezia - Venice / Venedig / Venise / ..
  
- Character encoding: Βενετία




## Recognition Procedure (1)



- 1) **Database lookup** of all (uppercase) words in the text.
  
- 2) 1<sup>st</sup> word of **multi-word place name**? → look for the rest.
  - Stara +
  - Stara Zagora (Bulgaria)
  - Stara Lubovna (Slovakia)
  - Stara Reka (Bulgaria)
  - Stara Tura (Slovakia)
  - Stara Wrona (Poland)
  - ...
  
- 3) **Geo stop word list** for homographs with common words ('And', 'Split', 'Chirac', 'Estaing', 'Harden', ...)



## Recognition Procedure (2)




- 4) **Disambiguation** of place name homographs: 'Paris', 'Roma' (It/Ro), ...
  - Using size information  
Roma (IT) = 1  
Roma (RO) = 4
  - Except if the size class is similar and there are other references to the same country:  
"Birmingham in the United States" → US
  - **Other possibilities:**
    - source country of document
    - Analysis at sentence level
    - Use administrative units at various levels
    - Use geographic distance
    - ...


NAME	COUNTRY	CLASS	LONGITUDE	LATITUDE
Roma	Italy	1	12,4906	41,8955
Roma	Sweden	4	18,4488	57,5319
Roma	Zambia	4	28,3845	-15,375
Roma	Romania	4		
Roma	Lesotho	5		
Roma	Brazil	5		
Roma	Australia	5		
Roma	Colombia	6		
Roma	Canada	6		
Roma	Zambia	6		
Roma	Peru	6		
Roma	Chile	6		

NAME	COUNTRY	CLASS
Birmingham	United Kingdom	2
Birmingham	United States	2
Birmingham	United States	4
Birmingham	United States	4
Birmingham	United States	6
Birmingham	United States	6
Birmingham	United States	6




## Evaluation




- 80 English texts (730 KB), containing 853 references
  - Precision: 96.8%
  - Recall: 96.5%

**Error Analysis**

- **Recall:**
  - Missing regions in database (**Montenegro**)
  - Missing names of peoples (**Timorese, Bosnian**)
  - [Missing names of mountains, rivers, islands, stretches of sea (**Adriatic**)]
- **Precision:**
  - Homographs with people's names, e.g. (**Cox**, Nigerian General **Abacha**)
  - **Malta**: city in Portugal



## Limitations (now partially overcome)



- Context not used for disambiguation: **President Chirac**

- Local variants of place names missing
- Only Roman alphabet

} **KNAB database:**  
Institute of the Estonian Language

KNAB contains historic, geographic and alphabetic variations of place names



**ipsc**  
Institute for the Promotion  
and Security of the Citizen

## Experiments with New EU Languages



**EUROPEAN COMMISSION**  
DIRECTORATE-GENERAL  
Joint Research Centre


**Bulgarian:** Членовете на Европейския парламент се избират чрез пряко гласуване по пропорционалната система на регионална основа, като например в **Италия**[Italy], **Великобритания**[United Kingdom] и **Белгия**[Kingdom of Belgium], на национална основа, като във **Франция**[France], **Испания**[Spain], **Австрия**[Republic of Austria], **Дания**[Denmark], **Люксембург**[Luxembourg] и други, или при смесена система (**Германия**[Germany]).

**Czech:** Poslanci Evropského parlamentu jsou voleni na základě všeobecného přímého volebního práva podle zásad poměrného zastoupení, a to buď na základě regionálního, jako například v **Itálii**[Italy], **Spojeném království**[UK] a **Belgii**[Kingdom of Belgium], nebo národním, jako ve **Francii**[France], **Španělsku**[Spain], **Rakousku**[Republic of Austria], **Dánsku**[Denmark], **Luxembursku**[Luxembourg] a dalších zemích, nebo na základě smíšeného systému (**Německo**[Germany]).

**Estonian:** Euroopa Parlamendi liikmed valitakse otseselt ja üldiselt valimiste teel, kasutades proportsionaalse esindatuse süsteemi, kas siis regionaalsel alusel, nagu näiteks **Itaalia**[Italy], **Ühendkuningriigi**[United Kingdom]gis ja **Belgias**[Kingdom of Belgium], või üleriigiliselt, nagu see on Prantsusmaal, **Hispaanias**[Spain], **Austrias**[Republic of Austria], **Taanis**[Denmark], **Luksemburgis**[Luxembourg] ja teistes riikides, või siis kombineeritud süsteemi põhjal (**Saksamaal**[Germany]).


**Slovene:** Poslance Evropskega parlamenta volijo prebivalci držav članic na osnovi splošne in neposredne volilne pravice po proporcionalnem sistemu zastopnosti, bodisi na regionalni osnovi, kot npr. v **Italiji**[Italy], Velika **Britanija**[United Kingdom] in **Belgiji**[Kingdom of Belgium], bodisi na nacionalni osnovi, kot npr. v **Franciji**[France], **Spaniji**[Spain], **Avstriji**[Republic of Austria], **Danski**[Denmark], **Luksemburgu** in drugje, ali pa po mešanem sistemu (**Nemčija**[Germany]).

**Polish:** Deputowani do Parlamentu Europejskiego są wybierani w powszechnych wyborach bezpośrednich zgodnie z systemem proporcjonalnej reprezentacji albo na podstawie regionalnej, jak na przykład we **Włoszech**[Italy], **Wielkiej Brytanii**[UK]i oraz **Belgii**[Kingdom of Belgium], albo krajowej, tak jak we **Francji**[France], **Hiszpanii**[Spain], **Austrii**[Republic of Austria], **Danii**[Denmark], **Luksemburgu** i innych krajach, lub według systemu mieszane go (**Niemcy**[Germany]).



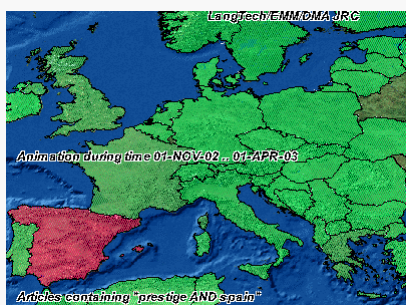
**ipsc**  
Institute for the Promotion  
and Security of the Citizen

## Prestige Tanker Accident: Change of Country Involvement over Time in News Reports



**EUROPEAN COMMISSION**  
DIRECTORATE-GENERAL  
Joint Research Centre




LangTech/EMM/DMA/JRC  
Animation during time 01-NOV-02.. 01-APR-03  
Articles containing "prestige AND spain"


**JRC Collaborative Effort:**

Language Technology  
ISFEREA-DMA  
WT-EMM


  




November




Mid-November



December



Mid-December



January