The Prolex Project

A multilingual platform to Proper Name Processing
Motivations
Proper Names

• The common denominator of many definitions of proper names proposed by linguists may be

Proper name refers as opposed to
Common noun is defined.
Proper Names

• The prototypes of proper names are personal names and toponyms...

Lech Kaczyński, Jacques Chirac, Poland, France
Warsaw, Paris
Proper Names

- But we want also accept as proper names: names of organizations, events, etc. as in Named Entity task...

the independent trade union Solidarity
the Warsaw Pact
the Eiffel Tower
Unknown words

- Somebody considers capitalized unknown words to proper names...
- It is not exact...above all multiword units!
- See four examples extracted to newspaper *Le Monde*
## Unknown words

<table>
<thead>
<tr>
<th>Proper names</th>
<th>Unknown words</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Banque Centrale Européenne</em></td>
<td>0</td>
</tr>
<tr>
<td><em>Banco Real</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Banca commerciale italiana</em></td>
<td>2</td>
</tr>
<tr>
<td><em>Banca di Roma</em></td>
<td>3</td>
</tr>
</tbody>
</table>
Relations

- Texts often play with anaphoric relation between Proper names.

John Paul II
Karol Jozef Wojtyla
the Polish pope
• Somebody thinks that Proper Names are not translated but it is not true...

Warszawa
Warsaw
Varsovie

Чехов
Chekhov
Tchekhov, Tchékhov

Karol Józef Wojtyła
Karol Jozef Wojtyla

Czechow (nominative)
Czechowa (accusative)
Czechowowi (dative)
Some realized works
Inhabitant Names in French

• The derivation from City Name to Inhabitant Name in French is not obvious (because, we frequently use an auxiliary base)

Inhabitant of Metz
Messin
Metzois
Messois
Ambiguity of dot in French

- In French, dot is used at the end of sentences, but also as abbreviation...

M. (Mr) L. Kaczynski a rencontré (has met)
M. J. Chirac à l'O.N.U. (in UNO)

cé matin (this morning).

Abbreviation
First name
Acronym
End of sentence
French Proper Noun recognition

<Person>
  <Status> M. </Status>
  <First> L. </First>
  <Name> Kaczynski </Name>
</Person>
<table>
<thead>
<tr>
<th>Newspaper</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le Monde</td>
<td>93.2</td>
<td>94.4</td>
</tr>
</tbody>
</table>
Prolexbase

A multilingual relational database of Proper Name
Two main concepts

Two parts

Language independent

Language dependent

PIVOT
NUM_PIVOT

A_pour_concept

PROLEXEME
NUM_PROLEXEME
LIBELLE_PROLEXEME

Two parts
Two main concepts

Two parts

Conceptual Level

Linguistic Level
Two main concepts

Two parts

The conceptual Proper Name

A set of word semantically linked to the Proper Name
Two main concepts
Conceptual Level

Accessibility
(Location: family, authors, capital...)

Metaconceptual Level
(hyperonymies)

Meronymy

Diachronic Synonymy
(Saint Petersburg, Petrograd, Leningrad)

Synonymy
(Diasystem: diachronic, diastatic, diaphasic)
Conceptual Level

Accessibility
(Location: family, authors, capital...)

Metaconceptual Level
(hyperonymies)

Meronymy

Diastatic Synonymy
(John-Paul II, Karol Jozef Wojtyla)

Synonymy
(Diasystem: diachronic, diastatic, diaphasic)
Conceptual Level

Accessibility
(Location: family, authors, capital...)

Metaconceptual Level
(hyperonymies)

Meronymy

Synonymy
(Diasystem: diachronic, diastratic, diaphasic)

Diaphasic Synonymy
(Paris, The City of Light)
Conceptual Level

Accessibility
(Location: family, authors, capital...)

Metaconceptual Level
(hyperonymies)

Meronymy

Synonymy
(Diasystem: diachronic, diastatic, diaphasic)
Conceptual Level

Meronymy

Diastratic Synonymy

en {Poland Polish Polish}

fr {Pologne Polonais polonais Polak}

en {John-Paul II}

fr {Jean-Paul II}

en {Karol Jozef Wojtyla}

fr {Karol Jozef Wojtyla}
Metaconceptual Level

Typology (Celebrity, Country, Hydronym...)

Existence (historical, fictive, religious)
## Typology

<table>
<thead>
<tr>
<th>Proper Name</th>
<th>Anthroponym</th>
<th>Toponym</th>
<th>Ergonym</th>
<th>Pragmonym</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anthroponym</strong></td>
<td>Individual</td>
<td>Group</td>
<td>Territory</td>
<td></td>
</tr>
<tr>
<td>Celebrity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Name</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patronym</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudo-anthroponym</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dynastie</strong></td>
<td>Collective</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnonyme</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Association</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Firm</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institution</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Astronym</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>City</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Geonym</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Supra-national</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hydronym</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Way</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Object</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thought</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vessel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Work</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Disaster</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Event</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Feast</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hystory</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Meteorology</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Linguistic Level

Derivate
(Relative noun,
Relative adjective,
Prefix...)

Alias
(variation,
abbreviative,
acronym...)

A representative
Proper Noun
Linguistic Level

Prolexeme

United States of America

American (Noun)

American (Adjective)

United States

USA

46929

Alias

Derivatives
Linguistic Level

Rules to generate derivatives

A_pour_dérivation3 → DERIVATION
NUM_DERIVATION
LIBELLE_DERIVATION
A_pour_dérivation1

A_pour_dérivation2 → ALIASISATION
NUM_ALIASISATION
LIBELLE_ALIASISATION
A_pour_aliasisation

Rules to generate aliases

PROLEXEME
NUM_PROLEXEME
LIBELLE_PROLEXEME
Inflexion
Instance Level

United States

USA

American (Adjective)

American (Noun)

United States of America

American Americans

Instances
Further relations

- Classifier
- Eponymy
- Phonetic
- Data pertinence
- Sort
- Determination
- Language
Eponymy

Antonomasia
(biro, pampersy)

Terminology
(Alzheimer's disease)

Idiom
(not for all the tea in China)
Data pertinence

Details,
National,
European,
International

From a corpus study?
Some numbers...

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemmas</td>
<td>4,092 anthroponyms</td>
</tr>
<tr>
<td></td>
<td>49,751 toponyms</td>
</tr>
<tr>
<td></td>
<td>174 ergonyms</td>
</tr>
<tr>
<td></td>
<td>222 pragmonyms</td>
</tr>
<tr>
<td>Relations</td>
<td>2,249 accesibilities</td>
</tr>
<tr>
<td></td>
<td>47,670 meronymies</td>
</tr>
<tr>
<td></td>
<td>648 synonymies</td>
</tr>
<tr>
<td>Instances</td>
<td>123,859 instances</td>
</tr>
</tbody>
</table>
Merci !