Multilingual Text processing & Multilingual Lexical Databases
Structure, encoding, issues...

Gilles Sérasset
GETALP-LIG, Université Joseph Fourier, Grenoble, France

14th October 2015
1 Introduction
We are here for a 4.5 hours course

It means that you will have to participate!

Ask question whenever you want! Don’t hesitate to interrupt me!
Or we will all be Magn(bored)
Foreword

We are here for a 4.5 hours course

It means that you will have to participate!

Ask question whenever you want! Don’t hesitate to interrupt me!
Or we will all be bored to death
Who Am I?

Working on multilingual lexical database management since 1990.

PhD thesis on pivot approach to MLDB and generic MLDB system in 1994

Active in Electronic Lexicography, Multilingual Information Retrieval, Multilingual Computing
Part I

Multilingual Lexical Resources, a First Approach
Dictionaries and Other Lexical Data
- Monolingual Lexical Data
- Bilingual Lexical Data
- Multilingual Lexical Data

Computational Problems posed by ALL Lexical Data
- Denoting languages of the world
- Writing Systems
- Naming a language
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
- RDF and Linked Data
Dictionaries and Other Lexical Data
- Monolingual Lexical Data
- Bilingual Lexical Data
- Multilingual Lexical Data

Computational Problems posed by ALL Lexical Data
- Denoting languages of the world
- Writing Systems
- Naming a language
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
- RDF and Linked Data
Monolingual Lexical Data
A dictionary is a book of alphabetically listed words in a specific language, with definitions, etymologies, pronunciations, and other information; or a book of alphabetically listed words in one language with their equivalents in another, also known as a lexicon.\(^a\)

\(^a\)Webster’s New World College Dictionary, Fourth Edition, 2002
Dictionary

A First Definition

A dictionary is a book of alphabetically listed words in a specific language, with definitions, etymologies, pronunciations, and other information; or a book of alphabetically listed words in one language with their equivalents in another, also known as a lexicon.¹

¹Webster’s New World College Dictionary, Fourth Edition, 2002

1 a collection of words,
Dictionary

A First Definition

A dictionary is a book of alphabetically listed words in a specific language, with definitions, etymologies, pronunciations, and other information; or a book of alphabetically listed words in one language with their equivalents in another, also known as a lexicon.¹

¹Webster’s New World College Dictionary, Fourth Edition, 2002

1. a collection of words,
2. each word comes with a set of information that describes it,
Dictionary

A First Definition

A dictionary is a book of **alphabetically listed** words in a specific language, with definitions, etymologies, pronunciations, and other information; or a book of **alphabetically listed** words in one language with their equivalents in another, also known as a lexicon.\(^a\)

\(^a\)Webster’s New World College Dictionary, Fourth Edition, 2002

1. a collection of **words**,  
2. each word comes with a **set of information** that describes it,  
3. there is an implicit **access method** to each word.
Examples of monolingual dictionary

`com•pose | kəmˈpōz`  
verb [ trans. ]  
1 write or create (a work of art, esp. music or poetry): *he composed the First Violin Sonata four years earlier.*  
   • write or phrase (a letter or piece of writing) with care and thought: *the first sentence is so hard to compose.*  
   • form (a whole) by ordering or arranging the parts, esp. in an artistic way: *compose and draw a still life.*  
   • order or arrange (parts) to form a whole, esp. in an artistic way: *make an attempt to compose your images.*  
2 (usu. be composed) (of elements) constitute or make up (a whole): *the system is composed of a group of machines.*  
   • be (a specified number or amount) of a whole: *Christians compose 40 percent of the state's population.*  
3 calm or settle (oneself or one's features or thoughts): *she tried to compose herself.*  
   • archaic settle (a dispute): *the king, with some difficulty, composed this difference.*  
4 prepare (a text) for printing by manually, mechanically, or electronically setting up the letters and other characters in the order to be printed.  
   • set up (letters and characters) in this way.
Examples of monolingual dictionary

**compose** |ˈkʌmpəz|
---
verb [ trans. ]
1 write or create (a work of art, esp. music or poetry) : *he composed the First Violin Sonata four years earlier.*
   - write or phrase (a letter or piece of writing) with care and thought : *the first sentence is so hard to compose.*
   - form (a whole) by ordering or arranging the parts, esp. in an artistic way : *compose and draw a still life.*
   - order or arrange (parts) to form a whole, esp. in an artistic way : *make an attempt to compose your images.*
2 (usu. **be composed**) (of elements) constitute or make up (a whole) : *the system is composed of a group of machines.*
   - be (a specified number or amount) of a whole : *Christians compose 40 percent of the state's population.*
3 calm or settle (oneself or one's features or thoughts) : *she tried to compose herself.*
   - archaic settle (a dispute) : *the king, with some difficulty, composed this difference.*
4 prepare (a text) for printing by manually, mechanically, or electronically setting up the letters and other characters in the order to be printed.
   - set up (letters and characters) in this way.

**compose** v. composed ◇ composing ◇ composes |ˈkʌmpəz| 1. To write music; *"Beethoven composed nine symphonies"*; 2. To write prose; *"Many students compose their paper at the computer keyboard"*; 3. To put together out of existing material; 4. To form the substance of; *"Greed and ambition composed his personality."* 5. To calm (someone, esp. oneself); make quiet; *"She had to compose herself before she could reply to this terrible insult."*
Thesaurus

First Definition

A Thesaurus is a listing of words with similar, related, or opposite meanings (this new meaning of thesaurus dates back to Roget’s Thesaurus)
Thesaurus

First Definition

A Thesaurus is a listing of words with similar, related, or opposite meanings (this new meaning of thesaurus dates back to Roget’s Thesaurus)

1 a collection of words,
A Thesaurus is a listing of words with similar, related, or opposite meanings (this new meaning of thesaurus dates back to Roget’s Thesaurus)

1. a collection of words,
2. there is no definition or description of any word (this is left to the dictionary),
A Thesaurus is a listing of words with similar, related, or opposite meanings (this new meaning of thesaurus dates back to Roget’s Thesaurus)

1. a collection of **words**, 
2. there is no definition or description of any word (this is left to the dictionary), 
3. a thesaurus may be seen as a network of word **senses**.
An example of monolingual thesaurus

**compose**

verb

1. *a poem composed by Shelley* WRITE, formulate, devise, make up, think up, produce, invent, concoct; pen, author, draft; score, orchestrate, choreograph.

2. *compose a still life* ORGANIZE, arrange, set out.

3. *the subcommittee is composed of ten senators* MAKE UP, constitute, form.

**PHRASES**

**compose oneself** you have to compose yourself before you take the stand CALM DOWN, control oneself, regain one's composure, pull oneself together, collect oneself, steady oneself, keep one's head, relax; informal get a grip, keep one's cool, cool one's jets, decompress.
A first set of problems

What is a word?

- Composed and Compose: one or two words?
- Compose and Composition: one or two words?
A first set of problems

What is a word?
- *Composed* and *Compose*: one or two words?
- *Compose* and *Composition*: one or two words?

Different dictionaries do not agree on the granularity of lexical entries
- *Stock_v* and *Stock_n*: one or two entries?
- *Judgement_{judging}* and *Judgement_{estimate}*: one or two word senses?
A first set of problems

What is a word?
- *Composed* and *Compose*: one or two words?
- *Compose* and *Composition*: one or two words?

Different dictionaries do not agree on the granularity of lexical entries
- *Stock*$_v$ and *Stock*$_n$: one or two entries?
- *Judgement*$_{judging}$ and *Judgement*$_{estimate}$: one or two word senses?

How do different lexical data interoperate?
- Can I merge information from the dictionary and from the thesaurus?
- Can I merge information from two different dictionaries?
A first set of problems

What is a word?
- *Composed* and *Compose*: one or two words?
- *Compose* and *Composition*: one or two words?

Different dictionaries do not agree on the granularity of lexical entries
- *Stock*$_v$ and *Stock*$_n$: one or two entries?
- *Judgement*$_{judging}$ and *Judgement*$_{estimate}$: one or two word senses?

How do different lexical data interoperate?
- Can I merge information from the dictionary and from the thesaurus?
- Can I merge information from two different dictionaries?

Can I extract data from a dictionary?
- How is the dictionary structured?
- How is this structure represented?
- Are definitions consistent?
Some basic vocabulary

There is no such things as a word...

**Textual Analysis View (usual in GETALP translation systems)**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrences</td>
<td>a &quot;word&quot; as it appears in a text (e.g. am, are, lists, listed...).</td>
</tr>
<tr>
<td>Lemmas</td>
<td>a set of occurrences (e.g. to be = {am, are, \ldots}, to list = {lists, list, listed\ldots})</td>
</tr>
<tr>
<td>Lexical Unit</td>
<td>a set of morpho-semantically related lemmas (e.g. {to build, building, builder, \ldots})</td>
</tr>
</tbody>
</table>

**Lexicological View (definitions may vary among ≠ dictionaries)**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inflected form</td>
<td>a &quot;word&quot; as it appears in a text (e.g. am, are, lists, listed...).</td>
</tr>
<tr>
<td>Vocable</td>
<td>a set of lexical entries (e.g. bleu = {bleu_n, bleu_adj})</td>
</tr>
<tr>
<td>Lexical Entry</td>
<td>a set of word senses with a common etymology and an unique Part Of Speech (e.g. {to build, building, builder, \ldots})</td>
</tr>
<tr>
<td>Word Sense/Acception</td>
<td>a set of morpho-semantically related lemmas (e.g. {to build, building, builder, \ldots})</td>
</tr>
</tbody>
</table>
Some basic vocabulary
Polysemy vs Homonymy

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spelling, pronunciation and meaning</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Polysemy</strong></td>
<td>a <em>Lexical Entry</em> is polysemic if it has different word senses.</td>
</tr>
</tbody>
</table>
|           | e.g. river
|           | a. a stream of water; b. flow of liquid; c. (poker) last card in a deal; |
| **Homonymy** | Relation between words that are spelled the same but has a different meaning. |
Some basic vocabulary

Polysemy vs Homonymy

**Homophone**
Same pronunciation, different meaning

**Homograph**
Same spelling, different meaning

- **Heterograph**
  - Different spelling and meaning
  - e.g. too / two

- **Homonym**
  - Different meaning
  - e.g. tire (car wheel) / tire (fatigue)

- **Heteronym**
  - Different pronunciation and meaning
  - e.g. desert (arid region) / desert (leave)

- **Synonym**
  - Different spelling and pronunciation
  - e.g. settee / sofa

- **Identical words**
  - Words with different spelling, pronunciation and meaning

Gilles Sérasset (GETALP-LIG)
2 Dictionaries and Other Lexical Data
- Monolingual Lexical Data
- Bilingual Lexical Data
- Multilingual Lexical Data

3 Computational Problems posed by ALL Lexical Data
- Denoting languages of the world
- Writing Systems
- Naming a language
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
- RDF and Linked Data
Bilingual Lexical Data
Bilingual Dictionary: an example

**compose**: English definition | in Italian | in Spanish
conjugar | in context | images

Écouter 🎧 US - UK

### Principal Translations/Principales traductions

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>compose (write)</td>
<td>écrire</td>
<td>escribir</td>
</tr>
<tr>
<td>compose (music, literature)</td>
<td>composer</td>
<td>(música, literatura)</td>
</tr>
<tr>
<td>compose (put together)</td>
<td>composer</td>
<td></td>
</tr>
<tr>
<td>compose (technical)</td>
<td>constituer</td>
<td>(compositor)</td>
</tr>
</tbody>
</table>

### Additional Translations:

- compose vtr rédiger

**compose** v. <k&m'pOz> 1. composer v. 2. rédiger v. 3. écrire v. p.p.: écrit (in writing)
A second set of problems

How do I translate in my context?

- What is my word sense?
- What is the resulting word sense?
A second set of problems

How do I translate in my context?

- What is my word sense?
- What is the resulting word sense?

Can I use this info to translate the other way round?

- How much of the word sense is covered by the translation?
- How marginal is the target word sense, compared to word senses of the target word?
2 Dictionaries and Other Lexical Data
   - Monolingual Lexical Data
   - Bilingual Lexical Data
   - Multilingual Lexical Data

3 Computational Problems posed by ALL Lexical Data
   - Denoting languages of the world
   - Writing Systems
   - Naming a language
   - String Encoding Problems
     - Presentation of the problem
     - What do we code?
     - UNICODE
   - Searching and Sorting
   - RDF and Linked Data
Multilingual Lexical Data
An example of multilingual terminological database

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
<th>Spanish</th>
<th>Russian</th>
<th>Chinese</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>child poverty</td>
<td>pauvreté touchant les enfants</td>
<td>pobreza infantil</td>
<td>детская нищета</td>
<td>儿童贫穷</td>
<td>فقر الأطفال</td>
</tr>
<tr>
<td>alien</td>
<td>étranger</td>
<td>extranjero</td>
<td>иностранный</td>
<td>外侨</td>
<td>أجنبة；غريب；دخيل</td>
</tr>
<tr>
<td>Judicial Support Division</td>
<td>Division de l'appui judiciaire</td>
<td>División de Apoyo Judicial</td>
<td>Отдел вспомогательного обслуживания судопроизводства</td>
<td>司法支助司</td>
<td>شعبة الدعم القضائي</td>
</tr>
<tr>
<td>Finance Assistant</td>
<td>assistant (finances)</td>
<td>Auxiliar de finanzas</td>
<td>помощник по финансовым вопросам</td>
<td>财务助理</td>
<td>مساعد مالي</td>
</tr>
<tr>
<td>service pack</td>
<td>-</td>
<td>-</td>
<td>пакет обновлений</td>
<td>服务包</td>
<td>حزمة خدمة</td>
</tr>
<tr>
<td>population pyramid</td>
<td>pyramide des âges</td>
<td>pirámide por edades, pirámide de la población por edades</td>
<td>возрастная структура населения</td>
<td>人口金字塔</td>
<td>هرم أعمار السكان</td>
</tr>
<tr>
<td>sighe marriage</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>نكاح (زواج) المتعبة</td>
</tr>
<tr>
<td>Office of the Under-Secretary-General for Field Support</td>
<td>Bureau du Secrétaire général adjoint à l'appui aux missions</td>
<td>Oficina del Secretario General Adjunto de Apoyo a las Actividades sobre el Terreno</td>
<td>Канцелярия заместителя Генерального секретаря по полевой поддержке</td>
<td>主管外勤支助事务副秘书长办公室</td>
<td>مكتب وكيل الأمن العام للدعم الميداني</td>
</tr>
<tr>
<td>raster</td>
<td>-</td>
<td>-</td>
<td>1. растр; 2. двумерный массив точек; 3. раstroвое изображение</td>
<td>光栅</td>
<td>نمط خطوط المسح</td>
</tr>
<tr>
<td>nuclear bomb;</td>
<td>bombe nucléaire;</td>
<td>bomba nuclear; arma nuclear</td>
<td>ядерное оружие</td>
<td>核弹；核武器</td>
<td>قنبلة نوعية：سلاح نووي</td>
</tr>
<tr>
<td>nuclear weapon</td>
<td>arme nucléaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
An example of a multi-bilingual dictionary

**composer** /kon-po-ze-/  

v.tr.  
préparer ; make (up) ; menyediakan ; mengadun  
Valérie a composé une salade avec tous les légumes du jardin ; Valérie took all the vegetables of the garden to make up a salad ; Valérie telah mengadun salad dgn sayur-sayuran yg diambil dari kebun  
former ; form ; menubuhkan  
sélectionner ; select ; memilih  
disposer ; arrange ; menyusun  
créer, écrire ; compose ; mencipta ; mengarang ; menyusun ; menggubah  
faire un numéro ; dial ; memutar ; memusing ; mendail  
v.intr. ; scol. ; take an exam ; mengambil peperiksaan ; menduduki peperiksaan  
les élèves des classes terminales composeront lundi ; the pupils of form 6 will take their exam on monday ; pelajar tingkat 6 akan menduduki peperiksaan pada hari Isnin  
transiger ; << ; <<  
s'accomoder avec ; compromise ; bertolak ansur ; bersetujuse compose /sekon-po-ze-/  
v.pr.  
se composer de ; be made up of ; terdiri drpd  
le spectacle se compose de deux parties avec un entracte ; the show is made up of two parts with an interlude ; pertunjukan itu terdiri drpd dua bhg dan ada waktu rehat di antaranya
Did you ever use a really multilingual printed dictionary?

I mean: a dictionary for which you can use ANY language as a source and ANY language as a target...
Did you ever use a really multilingual printed dictionary?

I mean: a dictionary for which you can use ANY language as a source and ANY language as a target...

No?
Did you ever use a really multilingual printed dictionary?

I mean: a dictionary for which you can use ANY language as a source and ANY language as a target...

No?

That’s normal... such dictionary does not exist...

... because nobody knows how to layout/present such information on paper...

... except in very specific situations (i.e. terminology)
Did you ever use a really multilingual printed dictionary?

I mean: a dictionary for which you can use ANY language as a source and ANY language as a target...

No?

That’s normal... such dictionary does not exist...
... because nobody knows how to layout/present such information on paper...
... except in very specific situations (i.e. terminology)

But you may find such a database online.
Before we go further, let’s see what are the technical difficulties that have to be handled...
Dictionaries and Other Lexical Data
- Monolingual Lexical Data
- Bilingual Lexical Data
- Multilingual Lexical Data

Computational Problems posed by ALL Lexical Data
- Denoting languages of the world
- Writing Systems
- Naming a language
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
- RDF and Linked Data
2 Dictionaries and Other Lexical Data
- Monolingual Lexical Data
- Bilingual Lexical Data
- Multilingual Lexical Data

3 Computational Problems posed by ALL Lexical Data
- Denoting languages of the world
- Writing Systems
- Naming a language
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
- RDF and Linked Data
Languages of the world

The biggest database on the matter is the “Ethnologue\(^1\)” database. It currently contains 6912 living languages in the world that are geographically distributed as follows:

- Africa: 30%
- America: 16%
- Europe: 3%
- Asia: 32%
- Pacific: 19%

\(^{1}\)http://www.ethnologue.com/
LANGUAGES OF THE WORLD

EACH DOT REPRESENTS THE PRIMARY LOCATION OF A LIVING LANGUAGE LISTED IN THE ETHNOLOGUE.

image taken from http://www.ethnologue.com/
This distribution is the result of difficult work and the precision of the numbers is not easy to evaluate. The problems are:

- Heterogeneity of the sources
- The difficult distinction between languages and dialects.\(^2\)

Depending on the sources, estimations vary between about 3000 and 7000 living languages.

\[^2\text{Languages are generally the object of statistics, but the definition of a language is based on subjective and variable criteria.}\]

---

Max Weinreich

"א שפרַך איז א דיאַלאָקט מיט אן אַרְמי און פלוֹט"

(a shprakh iz a dialekt mit an armey un flot)

a language is a dialect with an army and a navy.
Written languages, spoken languages

It is even more difficult to know the number of written languages. Depending on sources, 67% to 90% of the languages are spoken languages. This gives around 1500 written languages among the 6912. The Bible Society\(^3\) indicates that the Bible has been translated into 2426 different languages (at least one of the book has been translated). Certain languages, considered as oral languages can be written using a transcription using an existing writing system.

\(^3\)www.biblesociety.org
## Most spoken languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Native speakers</th>
<th>Total sp.</th>
<th>Other estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>335 (2003–2011)</td>
<td>&gt; 765</td>
<td>Approximately 375 million L1 speakers, 375 million L2 speakers, and 750 million EFL speakers. Totaling about 1500 million speakers.[9]</td>
</tr>
<tr>
<td>Arabic</td>
<td>206 (1999)</td>
<td>452</td>
<td>280 million native.[13]</td>
</tr>
<tr>
<td>Bengali</td>
<td>193 (2001)</td>
<td>250</td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>162 (2010)</td>
<td>272</td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>122 (1985)</td>
<td>123</td>
<td></td>
</tr>
</tbody>
</table>

4100 million are not proficient in Standard Arabic
<table>
<thead>
<tr>
<th>Language</th>
<th>Native speakers</th>
<th>Total sp.</th>
<th>Other estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>83.8 (1990)</td>
<td>111.8</td>
<td>99M native, 60 second language in EU + 5–20 worldwide.</td>
</tr>
<tr>
<td>Lahnda</td>
<td>82.7 (1998–2000)</td>
<td>—</td>
<td>90</td>
</tr>
<tr>
<td>Wu (Shanghainese)</td>
<td>77.2 (1984)</td>
<td>—</td>
<td>90</td>
</tr>
<tr>
<td>Telugu</td>
<td>74.0 (2001)</td>
<td>79</td>
<td>84.6 (2011 census)</td>
</tr>
<tr>
<td>Marathi</td>
<td>71.8 (2001)</td>
<td>—</td>
<td>72 (2001 census)</td>
</tr>
<tr>
<td>Tamil</td>
<td>68.8 (2001)</td>
<td>76.8</td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>68.5 (1987-2011)</td>
<td>118.5</td>
<td>128 M native/real, 72 M &quot;bilinguals&quot;. &gt; 200 M total.</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>67.8 (1999)</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Urdu</td>
<td>63.4 (1998)</td>
<td>167.4</td>
<td></td>
</tr>
<tr>
<td>Yue (Cantonese)</td>
<td>62.2 (1984–2006)</td>
<td>—</td>
<td>70</td>
</tr>
<tr>
<td>Persian</td>
<td>56.6 (2011)</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>50.7 (1987)</td>
<td>—</td>
<td>74 &amp; 83 (2005)</td>
</tr>
<tr>
<td>Italian</td>
<td>—</td>
<td>61</td>
<td></td>
</tr>
</tbody>
</table>

\(^5\) the boundary between this and Eastern Punjabi is spurious.
Dictionaries and Other Lexical Data

- Monolingual Lexical Data
- Bilingual Lexical Data
- Multilingual Lexical Data

Computational Problems posed by ALL Lexical Data

- Denoting languages of the world
- Writing Systems
- Naming a language
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
- RDF and Linked Data
Writing Systems

Writing appeared under different forms in different human civilizations: Egypt, Mesopotamia, China, India and, later, mid-america (the Mayas). The most ancient writing system seems to have appeared in Mesopotamia, around 3500-3400 B.C.

We use the term **writing system**.

Two main classes:

- Phonologic scripts where **phonemes** are written (alphabets + syllabaries)
- Semantic scripts where **morphemes** are written (Chinese script, hieroglyphs or cuneiforms)
Writing Systems of the World

- Wikipedia
- Latin
- Arabic
- Chinese
- Cyrillic
- other writing systems

image taken from wikipedia
Some examples

Japanese writing system:

This system uses:

- “kanji”, Chinese ideograms: 漢字 (characters)
- “hiragana”, syllabary used for Japanese words, grammatical particles and flexion endings: 分かりません: wakarimasen (don’t understand)
- “katakana”, syllabary used for loan words: コード: côdo (code)
- “romaji”, i.e. latin alphabet, often used for acronyms: UNICODE
- arabic numbers, besides Chinese ones; special punctuation marks, etc.

Note: there is no separation between words...

Some computer science problems

How to pronounce kanji; segmenting a text in words; inputing (more characters than available key in a keyboard...), typography according to the writing direction (horizontal/vertical), etc.
Some examples

Arabic writing system:

- is an abjad, i.e. a writing in system which one may only write consonants (almost),
- written from right to left (except numbers, written from left to right...)
- no upper/lower case, letter are joined together,
- letter changes depending on the context (alone, beginning of a word, end of a word, between letters...)
- a space is found between words, but you can also find one (smaller) inside words containing letters that do not join

Some computer science problems

What are the (short) vowels (they encode morphology), fine typography (you do not justify text by extending white spaces but letters), discontinuous selection (e.g.: if there is a number), etc.
Some examples

Briefly describe particularities of your own native writing system...
Dictionaries and Other Lexical Data
- Monolingual Lexical Data
- Bilingual Lexical Data
- Multilingual Lexical Data

Computational Problems posed by ALL Lexical Data
- Denoting languages of the world
- Writing Systems
- Naming a language
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
- RDF and Linked Data
How do you name a language?

Giving a name to a language is difficult. Indeed, the name of a language is defined by the user language.

ex.: allemand, Deutsch, German, nemško, tedesco, ... are names for the German language (in French, German, English, Slovene, Italian, ...)

It is tedious to chose a particular language to make this designation (e.g. name every language in English...).

You can’t either chose to name each language with its "proper" name (e.g. English for English, français for French, Tiếng Việt for Vietnamese,...). This would be difficult to write for oral languages and difficult to pronounce for languages that use writing systems that are not known by the reader...
How to name a language?

You need to find a system that is:

- easy to set up,
- easily transportable (uses only ASCII characters),
- understandable by all (a standard).

It is the ISO-639 standard.
ISO-639 standard

ISO 639 contains 3$^6$ sets of language codes:

- ISO 639-1 (alpha-2) defines codes with 2 characters, and associates them with French and English names, along with the name in the described language itself;

- ISO 639-2 (alpha-3) uses 3 characters with 2 possible encodings: ISO 639-2/B (bibliographic) and ISO 639-2/T (terminological); the codes are associated with their French and English names.

- ISO 639-3 completes ISO 639-2 that has many missing languages; SIL$^7$ is the main author of this norm and is in charge of its maintenance (code addition, etc.); codes are associated with their English names.

\[6\] since 2007\[7\] www.sil.org
ISO-639 standard (suite)

These three code sets have different motivations:

- ISO 639-1 (alpha-2) is mainly intended for terminological, lexicographic and linguistic use; it contains main languages for which many linguistic resources exist;

- ISO 639-2 (alpha-3) is mainly intended for bibliography and terminology; it contains languages of part 1 plus many others having a large literature;

- ISO 639-3 intends to enumerate as many languages as possible, including living languages, dead languages, ancient languages and artificial languages\(^8\), regardless of the number of its speakers/readers and regardless of its written/spoken nature.

\(^8\)like Esperanto, Volapük, Klingon, etc. i.e. artificial natural languages (not java, xml, etc.)
Details of ISO 639-3

Set of codes:

CREATE TABLE ISO_639-3 (  
   Id       char(3) NOT NULL,  
       -- The three-letter 639-3 identifier  
 Part2B   char(3) NULL,  
       -- Equivalent 639-2 identifier of the bibliographic applications code set, if there is one  
 Part2T   char(3) NULL,  
       -- Equivalent 639-2 identifier of the terminology applications code set, if there is one  
 Part1    char(2) NULL,  
       -- Equivalent 639-1 identifier, if there is one  
 Scope    char(1) NOT NULL,  
       -- I(ndividual), M(acrolanguage), S pecial  
 Type     char(1) NOT NULL,  
       -- A(ncient), C(onstructed),  
       -- E(xtinct), H(istorical), L(iving), S pecial  
 Ref_Name varchar(150) NOT NULL)  
       -- Reference language name
### Extract from the codes table

<table>
<thead>
<tr>
<th>Id</th>
<th>Part2B</th>
<th>Part2T</th>
<th>Part1</th>
<th>Scope</th>
<th>Type</th>
<th>Ref_Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>aao</td>
<td>I</td>
<td>L</td>
<td>Algerian Saharan Arabic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>abh</td>
<td>I</td>
<td>L</td>
<td>Tajiki Arabic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ara</td>
<td>ara</td>
<td>ara</td>
<td>M</td>
<td>L</td>
<td>Arabic</td>
<td></td>
</tr>
<tr>
<td>czh</td>
<td>I</td>
<td>L</td>
<td>Huizhou Chinese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deu</td>
<td>ger</td>
<td>deu</td>
<td>I</td>
<td>L</td>
<td>German</td>
<td></td>
</tr>
<tr>
<td>epo</td>
<td>epo</td>
<td>eo</td>
<td>I</td>
<td>C</td>
<td>Esperanto</td>
<td></td>
</tr>
<tr>
<td>fra</td>
<td>fre</td>
<td>fr</td>
<td>I</td>
<td>L</td>
<td>French</td>
<td></td>
</tr>
<tr>
<td>frm</td>
<td>frm</td>
<td>frm</td>
<td>I</td>
<td>H</td>
<td>Middle French (ca. 1400-1600)</td>
<td></td>
</tr>
<tr>
<td>fro</td>
<td>fro</td>
<td>fro</td>
<td>I</td>
<td>H</td>
<td>Old French (842-ca. 1400)</td>
<td></td>
</tr>
<tr>
<td>mis</td>
<td>mis</td>
<td>mis</td>
<td>S</td>
<td>S</td>
<td>Uncoded languages</td>
<td></td>
</tr>
<tr>
<td>mjl</td>
<td>E</td>
<td>Mahican</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mul</td>
<td>mul</td>
<td>mul</td>
<td>S</td>
<td>S</td>
<td>Multiple languages</td>
<td></td>
</tr>
<tr>
<td>nut</td>
<td>I</td>
<td>L</td>
<td>Nung (Viet Nam)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>und</td>
<td>und</td>
<td>und</td>
<td>S</td>
<td>S</td>
<td>Undetermined</td>
<td></td>
</tr>
<tr>
<td>vie</td>
<td>vie</td>
<td>vi</td>
<td>I</td>
<td>L</td>
<td>Vietnamese</td>
<td></td>
</tr>
<tr>
<td>zho</td>
<td>chi</td>
<td>zh</td>
<td>M</td>
<td>L</td>
<td>Chinese</td>
<td></td>
</tr>
<tr>
<td>zxx</td>
<td>zxx</td>
<td>zxx</td>
<td>S</td>
<td>S</td>
<td>No linguistic content</td>
<td></td>
</tr>
</tbody>
</table>
Details of ISO 639-3

“macrolangues” mapping:

CREATE TABLE ISO_639-3_Macrolanguages (  
  M_Id   char(3) NOT NULL,  -- The identifier for a macrolanguage  
  I_Id   char(3) NOT NULL)   -- The identifier for an individual language  
                         -- that is a member of the macrolanguage
<table>
<thead>
<tr>
<th>M_Id</th>
<th>I_Id</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ara</td>
<td>aao</td>
<td>Algerian Saharan Arabic is part of Arabic</td>
</tr>
<tr>
<td>ara</td>
<td>abh</td>
<td>Tajiki Arabic is part of Arabic</td>
</tr>
<tr>
<td>ara</td>
<td>abv</td>
<td></td>
</tr>
<tr>
<td>zho</td>
<td>cjy</td>
<td></td>
</tr>
<tr>
<td>zho</td>
<td>cmn</td>
<td></td>
</tr>
<tr>
<td>zho</td>
<td>cpx</td>
<td></td>
</tr>
<tr>
<td>zho</td>
<td>czh</td>
<td>Huizhou Chinese is part of Chinese</td>
</tr>
</tbody>
</table>
Other informations

- there is a table describing the codes that have been suppressed,
- codes qaa to qtz are reserved for local use; they may be transmitted only if both parties agree to their interpretation,
- ISO 639-3 Registration Authority is SIL International, www.sil.org (Dallas, Texas)
- the code table contains 7642 entries (430 dead languages, 4 special, 63 historic, 114 ancient, 17 artificial)
Dictionaries and Other Lexical Data

- Monolingual Lexical Data
- Bilingual Lexical Data
- Multilingual Lexical Data

Computational Problems posed by ALL Lexical Data

- Denoting languages of the world
- Writing Systems
- Naming a language
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
- RDF and Linked Data
『In the beginning was the byte...』
String Encoding Problems

Initially, each character was represented by a unique code, represented by one byte.

Two main encodings: ASCII and EBCDIC (IBM)

ASCII code only defines 127 characters (using 7 bits). The other 127 characters usually depend on the OS/machine.

<table>
<thead>
<tr>
<th>ANSI extended ASCII</th>
<th>OEM extended ASCII (first IBM PCs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 A B C D E F</td>
<td></td>
</tr>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 A B C D E F</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
</tr>
<tr>
<td>O</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
</tr>
<tr>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
<td>O</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
</tbody>
</table>
“National” encodings also used the 127 higher characters to represent characters of their script\(^9\).
This, plus the fact that everybody called this THE ASCII code...
Hence, begun what can be called:

“the happy mess of encodings...”

\(^9\)the lower 127 characters were not modified because they were necessary to write programs
String Encoding Problems

Some national encodings:

- MacRoman, Windows-1252, Latin1 (≡ ISO-8859-1) for western european languages (French, Italian, ...),
- ISO-8859-2 for Slavic languages written with Latin characters (Czech, Polish, ...),
- ISO-8859-3 for Esperanto, Galician, Maltese, Turkish,
- ISO-8859-4 for Estonian, Lithuanian, Letton...
- ISO-8859-5 or KOI8-R for Cyrillic languages (Russian, Bulgarian, ...)
- ISO-8859-6 for Arabic,
- ISO-8859-7 for modern Greek,
- ISO-8859-8 for Hebrew,
- ISO-8859-9 (≡ Latin 5) almost the same as Latin1, but Islandic characters were removed in favor of Turkish ones,
- ISO-8859-15 (≡ Latin 9) same as Latin 1, plus €, œ, Œ... characters
String Encoding Problems

Scripts with large character sets cannot use this technique. **First strategy**: using an escape sequence that indicates the encoding of following characters.

**e.g.**: ISO-2022-JP: uses 7 bits per byte and 1 to 2 bytes per character. An escape sequence is inserted in the flow to indicate if the following characters use 1 or 2 bytes.

**Second strategy**: using 1 or 2 bytes, the form of which determines the number to be used.

**e.g.**: Shift-JIS: uses 8 bits per byte and 1 or 2 byte for a character. The value of the first byte determines if the character uses 1 or 2 bytes.
String Encoding Problems

All these encodings present a major problem: they are mutually exclusive one of the other!

Hence, it is impossible with a standard treatment\textsuperscript{10}, to mix different scripts...

This is a major problem when dealing with a bilingual dictionary (which is in essence, multilingual).

Hence, an encoding was introduced that is able to represent all currently known scripts: UNICODE.

\textsuperscript{10}For instance, some word processing software could mix encodings and use the font information to infer the proper encoding and display. This led to even more confusion between encoding and font...
What do we code?

How many characters are there below?

\[
\begin{array}{cccc}
\sigma & A \\
FBA6 & \mathcal{A} \\
\sim & A \\
FBA7 & \mathcal{A} \\
\tau & a \\
FBA8 & a \\
\tau & a \\
FBA9 &
\end{array}
\]
Some definitions

**Character**  Abstract entity, considered as atomic, used for writing a language,

**Glyph**  Abstract graphical form taken by a character in a certain context,

**Font**  Set of concrete graphical forms representing a set of glyphs,

**Code Point**  Association between a character and an integer (its code), as defined by a norm.
What do we code?

How many characters are there below?

\[ \sigma \quad A \quad \sigma \quad A \quad a \quad a \]

FBA6  
FBA7  
FBA8  
FBA9
What do we code?

How many characters are there below?

Only one character, four glyphs, 5 code points
What do we code?

How many characters are there below?

1 character, 1 glyph, ≠ fontes
What do we code?

How many characters are there below?

2 characters, 2 glyphs
What do we code?

How many characters are there below?

1 character: A
2 characters: A, capital cyrillic A, ...

FBA6

FBA7

FBA8

FBA9
The ISO 10646 norm defines a set of abstract characters, along with their names in English and French. It assigns code points to these characters. The UNICODE consortium, use these code points and defines the way they are represented in memory along with other elements necessary for the representation of texts.
# Design Principles

<table>
<thead>
<tr>
<th>Characters, not glyphs</th>
<th>The UNICODE standard encodes characters and not glyphs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantics</td>
<td>Characters have a well defined semantics.</td>
</tr>
<tr>
<td>Raw Text</td>
<td>The UNICODE defines the encoding of raw text.</td>
</tr>
<tr>
<td>Logical order</td>
<td>Implicit relation in memory is the logical order.</td>
</tr>
<tr>
<td>Unification</td>
<td>The UNICODE standard unifies identical characters in writing systems, independently of the languages.</td>
</tr>
<tr>
<td>Dynamic Composition</td>
<td>Accentuated forms can be composed dynamically.</td>
</tr>
<tr>
<td>Equivalent sequences</td>
<td>Each static precomposed form has a list of equivalent dynamically composed characters.</td>
</tr>
<tr>
<td>Convertibility</td>
<td>A one-one convertibility is guaranteed between UNICODE and other norms.</td>
</tr>
</tbody>
</table>
Two Antagonistic Principles

One-one convertibility means that characters defined by other encodings must be present in the UNICODE character set.

Hence, UNICODE defines a code point for each form of arabic characters (as these do exist in existing norms).

But these forms are not characters... but glyphs...
The definition model of characters is based on different levels:

1. abstract character repertory
2. coded character sets [The code points]
3. storage form of the characters [relation between the set of code points and the set of storage units]
4. character serialisation mechanism [correspondence between storage units and serialized sequence of bytes]
5. transfer encodings [reversible transformation of a sequence of bytes]
Representation Levels

Abstrait

Codé

Stocké

UTF-8

UTF-16

UTF-32
### Representation Levels

<table>
<thead>
<tr>
<th>Stocké</th>
<th>UTF-8</th>
<th>UTF-16</th>
<th>UTF-32</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C3 85</td>
<td>00C5</td>
<td>0000 00C5</td>
</tr>
<tr>
<td></td>
<td>E2 84 AB</td>
<td>212B</td>
<td>212B</td>
</tr>
<tr>
<td></td>
<td>F3 B0 80 80</td>
<td>DB80 DC00</td>
<td>000F 0000</td>
</tr>
<tr>
<td></td>
<td>41 CC 8A</td>
<td>0041 030A</td>
<td>0000 0041</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0000 030A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sérialisé</th>
<th>UTF-8</th>
<th>UTF-16BE</th>
<th>UTF-16LE</th>
<th>UTF-32BE</th>
<th>UTF-32LE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C3 85</td>
<td>00 C5</td>
<td>C5 00</td>
<td>00 00 00 C5</td>
<td>C5 00 00 00</td>
</tr>
<tr>
<td></td>
<td>E2 84 AB</td>
<td>21 2B</td>
<td>2B 21</td>
<td>00 00 21 2B</td>
<td>2B 21 00 00</td>
</tr>
<tr>
<td></td>
<td>F3 B0 80 80</td>
<td>DB 80 DC 00</td>
<td>80 DB 00 DC</td>
<td>00 0F 00 00</td>
<td>00 00 0F 00</td>
</tr>
<tr>
<td></td>
<td>41 CC 8A</td>
<td>00 41 03 0A</td>
<td>41 00 0A 03</td>
<td>00 00 00 41</td>
<td>41 00 00 00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>00 00 03 0A</td>
<td></td>
</tr>
</tbody>
</table>
Motivations

UTF-32 uses 32 bits, UTF-16 uses 16 bits (with the possibility of using a pair of short word to represent code points > 0xFFFF). This causes many problems:

```
example in C

char* str = ...; // str = "été", représenté en UTF-16
printf("%d", strlen(str));
```

This program prints 0 or 1, depending on the machine it runs on... The reason is that in C, char means byte and NOT character. Hence, the vast majority of programs do have problems with UNICODE data (SMTP servers, gateways, etc...)
Motivations

UTF-32 uses 32 bits, UTF-16 uses 16 bits (with the possibility of using a pair of short word to represent code points $> 0xFFFF$). This causes many problems:

**example in C**

```c
char* str = ...; // str = "été", représenté en UTF-16
printf("%d", strlen(str));
```

This program prints 0 or 1, depending on the machine it runs on... The reason is that in C, `char` means byte and NOT character.
Motivations

UTF-32 uses 32 bits, UTF-16 uses 16 bits (with the possibility of using a pair of short word to represent code points > 0xFFFF). This causes many problems:

Example in C

```c
char* str = ...; // str = "été", représenté en UTF-16
printf("%d", strlen(str));
```

This program prints 0 or 1, depending on the machine it runs on... The reason is that in C, `char` means `byte` and NOT character.

Hence, the vast majority of programs do have problems with UNICODE data (SMTP servers, gateways, etc...)

**UTF-8**

**Principles**

- Any ASCII character (0x00 \(\leq\) code point \(\leq\) 0x7F) is represented as is,
- Any character \(\geq 0x80\) is represented using the “little train” technique.

### The “little train”

Let \(x\) be a code point using \(n\) significative bits (7 \(>\) \(n\) \(>\) 21)

\[ x = b_n b_{n-1} ... b_1 \]

We scatter these bits in “wagons”, pulled by a strong enough locomotive (110xxxxx, 1110xxxx ou 11110xxx).

- if (7 < \(n\) \(\leq\) 11) \(\rightarrow\) 110xxxxx 10xxxxxx
- if (11 < \(n\) \(\leq\) 16) \(\rightarrow\) 1110xxxx 10xxxxxx 10xxxxxx
- if (16 < \(n\) \(\leq\) 21) \(\rightarrow\) 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
UTF-8

Advantages of this encoding:

- Any ASCII character (0x00 ≤ x ≤ 0x7F) is coded as is, hence it does not break protocols that are based on the ASCII character set,
- Byte 0x00 never appears in a UNICODE UTF-8 stream,
- Repositioning at the beginning of a char is always possible after a move in the byte stream (go back 3 bytes in the worst case),
- There is no difference between in memory storage units and serialized form as endianness is not an issue for bytes.
2 Dictionaries and Other Lexical Data
- Monolingual Lexical Data
- Bilingual Lexical Data
- Multilingual Lexical Data

3 Computational Problems posed by ALL Lexical Data
- Denoting languages of the world
- Writing Systems
- Naming a language
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
- RDF and Linked Data
Searching: some examples

- Naechste vs Nächste ?
- weiss vs weiß ?
- bleu vs Bleu ?
- blau vs Blau ?
Understanding the Problem: a Practical Example

1. Go to page http://de.wiktionary.org/,
2. Search Maß
3. Using a css editor (firebug under firefox will do this easily), change the form of title so that they appear in uppercase add text-transform:uppercase; in the h1 css)
4. You will see that Maß in now rendered as MASS
5. Search MASS, what do you see ?
Understanding the Problem: a Practical Example

1. Go to page
   http://dict.tu-chemnitz.de/dings.cgi?lang=en;service=deen,

2. search Maß

3. search Mass, what do you see?
Sorting: Some Examples

In Swedish: z < ö
In German: ö < z
In a German dictionary: ö is interpreted as o+e: öf < of
In a German phonebook: of < öf
In an Irish phonebook: McAllan < Macbeth
Sometimes: A < a, other times: a < A
In French: cote < côte < coté < côté
2 Dictionaries and Other Lexical Data
- Monolingual Lexical Data
- Bilingual Lexical Data
- Multilingual Lexical Data

3 Computational Problems posed by ALL Lexical Data
- Denoting languages of the world
- Writing Systems
- Naming a language
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
- RDF and Linked Data
Multilinguality in Semantic Web

One may think that “modern” languages are able to handle multilinguality...

For instance, RDF should encode data using different languages.

**Literals**

- no real encoding problem, most representations (n3, TTL, RDF-XML) use UTF-8.
- one may qualify a String with a language, but it uses 2 letters language code.
Multilinguality in Semantic Web

One may think that “modern” languages are able to handle multilinguality...
For instance, RDF should encode data using different languages.

<table>
<thead>
<tr>
<th>Literals</th>
</tr>
</thead>
<tbody>
<tr>
<td>• no real encoding problem, most representations (n3, TTL, RDF-XML) use UTF-8.</td>
</tr>
<tr>
<td>• one may qualify a String with a language, but it uses 2 letters language code.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>• URIs cannot use extended char. You should use en entity (%3C...).</td>
</tr>
<tr>
<td>• IRIs may be used in different serialization forms.</td>
</tr>
</tbody>
</table>
Multilinguality in Semantic Web

One may think that “modern” languages are able to handle multilinguality...
For instance, RDF should encode data using different languages.

**Literals**

- no real encoding problem, most representations (n3, TTL, RDF-XML) use UTF-8.
- one may qualify a String with a language, but it uses 2 letters language code.

**URLs**

- URIs cannot use extended char. You should use en entity (%3C...).
- IRIs may be used in different serialization forms.

**Linked Data**

- HTTP limitations: only accepts URLs...
- hence, the server should transform HTTP URL request and (optionally) transform URLs into IRIs before querying the triple store.
Part II

Example of Multilingual Lexical Databases
4 Electronic Dictionary Research

5 Wordnet and EuroWordnet

6 Wiki projects as Lexical Knowledge

7 BabelNet

8 The Papillon dictionary

9 The LexALP project
Electronic Dictionary Research (EDR) dictionary

One of the biggest lexical database currently available, it started in 1986 and lasted 9 years (1200 men.years for 14 G¥).

EDR built a bilingual lexical database with about 300000 entries in Japanese and English.

EDR architecture is somewhat unique as it gathers classical bilingual relations AND a concept dictionary that is used as a pivot.

Other lexical data was also produced, namely, cooccurrence dictionary and corpus (250000 analysed sentences, 10-20M raw sentences).

The data was aimed at machine translation systems.
Electronic Dictionary Research (EDR) dictionary

640000 concepts + description

English Word Dictionary
- 200000 general
- 100000 terms

Japanese Word Dictionary
- 200000 general
- 100000 terms

English Corpus
- 250000 sentences
- English Cooccurrence
- 300000 words

Japanese Corpus
- 250000 sentences
- Japanese Cooccurrence
- 300000 words

400000 bilingual entries
EDR concept dictionary

<table>
<thead>
<tr>
<th>HeadWord</th>
<th>Concept Illustration</th>
<th>Concept Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>plane(ELN1,ECN1)</td>
<td>airplane(vehicle)</td>
<td>A vehicle called airplane</td>
</tr>
<tr>
<td>plane(ELN1,ECN1)</td>
<td>plane(tool)</td>
<td>A carpentry tool called plane</td>
</tr>
</tbody>
</table>

Concepts are classified hierarchically. Moreover, many relations (agent, patient, etc.) are used to define/describe concepts.
Analysis of this dictionary

- Lots of data, directly usable by computer systems,
- Use of a pivot architecture that allows for the easy integration of new languages (see CICC project),
- But creating the concept dictionary was difficult (mainly for methodological problems).
Electronic Dictionary Research

Wordnet and EuroWordnet

Wiki projects as Lexical Knowledge

BabelNet

The Papillon dictionary

The LexALP project
Wordnet

- Wordnet lexical unit is called a **synset**,
- Each synset represents a **concept**,
- A synset is represented by a set of strings, each string denoting a word that bears the concept as one of its meanings.
- Each synset has a definition,
- Synsets are related by several relations (*nyms).

<table>
<thead>
<tr>
<th>POS</th>
<th>Unique Strings</th>
<th>Synsets</th>
<th>Total Wordsense Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>117798</td>
<td>82115</td>
<td>146312</td>
</tr>
<tr>
<td>Verb</td>
<td>11529</td>
<td>13767</td>
<td>25047</td>
</tr>
<tr>
<td>Adjective</td>
<td>21479</td>
<td>18156</td>
<td>30002</td>
</tr>
<tr>
<td>Adverb</td>
<td>4481</td>
<td>3621</td>
<td>5580</td>
</tr>
<tr>
<td>Totals</td>
<td>155287</td>
<td>117659</td>
<td>206941</td>
</tr>
</tbody>
</table>
Wordnet: an example

As such, Wordnet is not a dictionary. However, by providing a search by denotation strings, it behaves as a dictionary. Hence, searching “compose” gives:

**Verb**

- **S:** (v) compose (form the substance of) "Greed and ambition composed his personality"
- **S:** (v) compose, write (write music) "Beethoven composed nine symphonies"
- **S:** (v) write, compose, pen, indite (produce a literary work) "She composed a poem"; "He wrote four novels"
- **S:** (v) compose, compile (put together out of existing material) "compile a list"
- **S:** (v) compose (calm (someone, especially oneself); make quiet) "She had to compose herself before she could reply to this terrible insult"
- **S:** (v) frame, compose, draw up (make up plans or basic details for) "frame a policy"
Relations between synsets in Wordnet

{car; auto; automobile; machine; motorcar}

{cruiser; squad car; patrol car; police car; prowl car}  {cab; taxi; hack; taxicab;}

{motor vehicle; automotive vehicle}

{vehicle}

{conveyance; transport}

hyperonym

meronym

meronym

meronym

meronym
EuroWordnet is a multilingual database build on the wordnet model.

- Relations between synsets are given in different European languages.
- Synsets in European languages are linked by special relations to an ILI, an interlingual index.
- In theory, ILI are interlingual.
- In theory, there is an almost systematic one to one correspondence between ILI and English synset.
Eurowordnet: an example

cours d'eau

ruisseau

ruisselet

{fleuve; rivière}

hypernym

hypernym

hypernym

{stream; watercourse}

river

binnenwater

rivier

fiume

fiume
Eurowordnet: an example
Analysis of Wordnet/EuroWordnet

- Wordnet is not a dictionary. It is a set of concept,
- it does not use any notion of word, hence, you don’t know anything about the word, e.g. its level of language, etc.
- part of speech is defined by the concept, there is no such thing as a verb/noun entity (even if, in English, any noun may be used as a verb...)

- EuroWordnet uses English as a pivot (if not in theory, at least practically), this leads to mistakes in other languages, and, even if mistakes are avoided, this leads to artificial contrastive differences between languages.
- as a positive note: using a pivot allowed for different teams to work dependently on each language; using English as a pivot reduces the need for accurate competence in foreign languages.
4 Electronic Dictionary Research

5 Wordnet and EuroWordnet

6 Wiki projects as Lexical Knowledge

7 BabelNet

8 The Papillon dictionary

9 The LexALP project
Mediawiki promotes several projects with lexical knowledge

- Wikipedia: Community Built Multilingual Encyclopedic Database
  - Pages represent an encyclopedic entry (a concept ?)
  - Disambiguation pages, pages redirects, links to other Wikipedia language editions

- Wiktionary: community built multilingual dictionaries
  - Many different language editions, most of them containing entries in different languages
  - Language editions are poorly structured (different macro-structures, different use of the wiki markup language)

- Omegawiki: A structured dictionary
  - A common dictionary for all languages

- DefinedMeanings: Definitions in several languages

- Expressions: Term in a language associated to a defined meaning

Gilles Sérasset (GETALP-LIG)
Multilingual Text processing & Multilingual
14th October 2015
Mediawiki promotes several projects with lexical knowledge

Wiki Loves Africa: share African cultural fashion and adornment pictures with the world!

Play (theatre)

From Wikipedia, the free encyclopedia
(Redirected from Play (theater))

A play is a form of literature written by a playwright, usually consisting of dialogue between characters, intended for theatrical performance rather than just reading. Plays are performed at a variety of levels, from Broadway, Off-Broadway, regional theater, to Community theatre, as well as University or school productions. There are rare dramatists, notably George Bernard Shaw, who have had little preference whether their plays were performed or read. The term "play" can refer to both the written works of playwrights and to their complete theatrical performance.[1]

Contents

1 Genres
   1.1 Comedy
       1.1.1 Farce
       1.1.2 Satirical
Mediawiki promotes several projects with lexical knowledge

**Play (theatre)**

The following pages redirect to **Play (theatre)** and are grouped by target anchor.

Non-existent anchors are marked with [Invalid]. These redirects can be retargeted. Incorrect redirects are sometimes caused by bots fixing double redirects.

**No anchor or section**

- Play (theater)
- Play (theatrical)
- Playgoer
- Playlet
- Playlets
- Script (performing arts)
- Script (theater)
- Stage play
- Stage plays
- Stageplay
- Stageplays
- Theater play
- Theater/play
- Theatre play
- Theatre/play
Mediawiki promotes several projects with lexical knowledge
Mediawiki promotes several projects with lexical knowledge

- **Wikipedia**: Community Built Multilingual Encyclopedic Database
  - Pages represent an encyclopedic entry (a concept?)
  - Disambiguation pages, pages redirects, links to other Wikipedia language editions

- **Wiktionary**: community built multilingual dictionaries
  - Many different language editions, most of them containing entries in different languages
  - Language editions are poorly structured (different macro-structures, different use of the wiki markup language)
Mediawiki promotes several projects with lexical knowledge

Pronunciation [edit]

- enPR: plâ, IPA(key): /pleɪ/
- Audio (US) [Audio Player]  
- Rhymes: -eɪ

Verb [edit]

play (third-person singular simple present plays, present participle playing, simple past and past participle played)

1. (intransitive) To act in a manner such that one has fun; to engage in activities expressly for the purpose of recreation or entertainment. [quotations ▼]  
   They played long and hard.

2. (ergative) To perform in (a sport); to participate in (a game).

   He plays on three teams. Who’s playing now? play football; play sports; play games

   1. (transitive) To compete against, in a game. [quotations ▼]

3. (intransitive) To take part in amorous activity; to make love, fornicate; to have sex. [quotations ▼]

4. (transitive) To act as the indicated role, especially in a performance. [quotations ▼]

   He plays the King, and she's the Queen. No part of the brain plays the role of permanent memory.

5. (heading, transitive, intransitive) To produce music or theatre.

   1. (intransitive, of a music) To produce music. [quotations ▼]

   2. (intransitive, chiefly of a person) To produce music using a musical instrument.
Mediawiki promotes several projects with lexical knowledge

<table>
<thead>
<tr>
<th>Translations</th>
<th>edit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>playful activity</strong></td>
<td></td>
</tr>
<tr>
<td><strong>literary composition</strong></td>
<td></td>
</tr>
<tr>
<td><strong>theatrical performance</strong></td>
<td></td>
</tr>
</tbody>
</table>

- American Sign Language: A@Left1Chesthigh-A@Right1Chesthigh Upanddown
- Arabic: مسرح (masrah), مسرحية (masrabiyya)
- Armenian: հերկայակ'վիմ (herkayak'um)
- Catalan: obra
- Chinese: 戏剧 (xìjù), 話劇, 话剧 (huàjù)
- Czech: hra (cs)
- Danish: stykke (da), spil
- Dutch: optreden (nl) n, schouwspel (nl) n, toneelstuk (nl) n, stuk (nl) n
- Erzya: нәлкума (nalkuma)
- Esperanto: teatra
- Ewe: fefe
- Finnish: näytelmä (fi)
- French: pièce de théâtre (fr) f
- German: Schauspiel (de) n
- Hebrew: מִשְׁאֶה (mišāq)
- Hungarian: színdarab (hu), darab (hu)
- Indonesian: pentas (id), drama (id)
- Irish: dráma m
- Italian: dramma (it) m
- Japanese: 演劇 (enjugu, engeki), 戯曲 (giyoku, gikyoku)
- Korean: 연극 (ko) (yeon-geuk), 희곡 (ko) (huigok)
- Kurdish: şano (ku) f, pîyes (ku) f, tiyatro (ku) f
- Latin: fabula (la) f
- Macedonian: представа (predstava)
- Malayalam: സ്കൂസ്പിൽ (ml) (nātakam)
- Norwegian: Bokmål: skuespill n, Nynorsk: skodespel n
- Persian: پیس (pies)
- Polish: sztuka (pl) f
- Portuguese: peça f
- Russian: пьеса (ru) f (p'jesa), спектакль (ru) m (speektákl')
- Serbo-Croatian: Cyrillic: представа f, Roman: prèdstava f
- Slovak: hra f
- Slovene: igra (sl) f
- Spanish: obra teatral f
- Swahili: tamthilia (sw)
- Swedish: pjäs (sv) c, skådespel (sv) n, teaterpjäs (sv) c, teaterstyrke (sv) n, drama (sv) n
- Tagalog: dula (t)
- Telugu: సినిమాస్థానం (te) (nātakam)
- Turkish: oyun (tr), tiyatro oyunu
- Welsh: drama (cy) f
Mediawiki promotes several projects with lexical knowledge

**play**

<table>
<thead>
<tr>
<th>Sommaire</th>
<th>[masquer]</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Anglais</td>
<td></td>
</tr>
<tr>
<td>[x] Étymologie</td>
<td></td>
</tr>
<tr>
<td>[+] Nom commun</td>
<td></td>
</tr>
<tr>
<td>[+] Verbe</td>
<td></td>
</tr>
<tr>
<td>[x] Prononciation</td>
<td></td>
</tr>
<tr>
<td>[x] Voir aussi</td>
<td></td>
</tr>
</tbody>
</table>

**Anglais**  [modifier l' modifier le wikicode]

**Étymologie**  [modifier l' modifier le wikicode]

*Étymologie manquante ou incomplète*. Si vous la connaissez, vous pouvez l'ajouter en cliquant ici.

Du moyen anglais.

**Nom commun**  [modifier l' modifier le wikicode]

<table>
<thead>
<tr>
<th>play</th>
<th>pler</th>
</tr>
</thead>
<tbody>
<tr>
<td>play</td>
<td>plays</td>
</tr>
</tbody>
</table>

1. Pièce de théâtre, pièce.

2. Activité de jouer.

- *Not all play is learning and not all learning is play.* — (Jan van Gilis, *Several Perspectives on Children's Play: Scientific Reflections for Practitioners*, 1 janvier 2007)

- The elementary school places an emphasis on learning, but also on play activities.
Mediawiki promotes several projects with lexical knowledge

- **Wikipedia**: Community Built Multilingual Encyclopedic Database
  - Pages represent an encyclopedic entry (a concept ?)
  - Disambiguation pages, pages redirects, links to other Wikipedia language editions
- **Wiktionary**: Community built multilingual dictionaries
  - Many different language editions, most of them containing entries in different languages
  - Language editions are poorly structured (different macro-structures, different use of the wiki markup language)
- **Omegawiki**: A structured dictionary
  - A common dictionary for all languages
  - **DefinedMeanings**: Definitions in several languages
  - **Expressions**: Term in a language associated to a defined meaning
Mediawiki promotes several projects with lexical knowledge

**play**

---

<table>
<thead>
<tr>
<th>Sommaire [masquer]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anglais</td>
</tr>
<tr>
<td>Étymologie</td>
</tr>
<tr>
<td>Nom commun</td>
</tr>
<tr>
<td>Verbe</td>
</tr>
<tr>
<td>Prononciation</td>
</tr>
<tr>
<td>Voir aussi</td>
</tr>
</tbody>
</table>

---

**Anglais** [modifer | modifier le wikicode]

---

**Étymologie** [modifer | modifier le wikicode]

Étymologie manquante ou incomplète. Si vous la connaissez, vous pouvez l'ajouter en cliquant ici.

Du moyen anglais.

---

**Nom commun** [modifer | modifier le wikicode]

**play \( \text{\textit{pler}} \)**

1. Pièce de théâtre, pièce.
2. Activité de jouer.

- *Not all play is learning and not all learning is play.* — (Jan van Gil, *Several Perspectives on Children's Play: Scientific Reflections for Practitioners*, 1 janvier 2007)
- *The elementary school places an emphasis on learning, but also on play activities.*
Mediawiki promotes several projects with lexical knowledge

DefinedMeaning: 852347 - act

**Definition**

<table>
<thead>
<tr>
<th>Language</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Castilian</td>
<td>Participar como actor en una obra teatral.</td>
</tr>
<tr>
<td>English</td>
<td>To perform a theatrical role.</td>
</tr>
<tr>
<td>French</td>
<td>Participer comme acteur dans un film ou au théâtre.</td>
</tr>
<tr>
<td>Italian</td>
<td>Partecipare come attore in un ruolo teatrale o cinematografico.</td>
</tr>
<tr>
<td>Turkish</td>
<td>Bir aktör olarak film veya sahne gösterisine katılmak.</td>
</tr>
</tbody>
</table>

**Synonyms and translations**

<table>
<thead>
<tr>
<th>Expression</th>
<th>Language</th>
<th>Spelling</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bavarian</td>
<td>spöin</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>Bavarian</td>
<td>spuin</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>Castilian</td>
<td>actuar</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>Castilian</td>
<td>escenificar</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>Castilian</td>
<td>interpretar</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>Castilian</td>
<td>declamar</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>≈ Castilian</td>
<td>protagonizar</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>≈ Castilian</td>
<td>recitar</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>acteren</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>spelen</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>act</td>
<td>show ▼</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>do</td>
<td>show ▼</td>
<td></td>
</tr>
</tbody>
</table>
4. Electronic Dictionary Research

5. Wordnet and EuroWordnet

6. Wiki projects as Lexical Knowledge

7. BabelNet

8. The Papillon dictionary

9. The LexALP project
Overview of the BabelNet data set

- One of the results of the MultiJEDI ERC Grant (Europe)
- R. Navigli, Sapienza University of Rome
- A HUGE Multilingual Network of Words/Concepts
- Structure inspired by Wordnet
- Most data aggregated from other existing lexical data
  - Originally, Wordnet + Wikipedia (+ MT systems)
  - Open Multilingual Wordnet, OmegaWiki, Wiktionary, Wikidata, GeoNames, Verbnet, Wikiquotes, WoNeF, Microsoft multilingual terminology
Structure of the BabelNet data set

- Babel synsets (a set of words, like wordnet synsets, but in different languages),
- Defined by aggregating Wordnet synsets and wikipedia pages
- Then, missing translations are computed using an MT System
- data is now available as Lexical Linked Data (Lemon Model, see later)
Structure of the BabelNet data set

(a) Excerpt of the WordNet graph centered on the synset \textit{play}_n^1.

(b) Excerpt of the Wikipedia graph centered on the Wikipage \textit{Play (theatre)}. 
Structure of the BabelNet data set

Fig. 2. An illustrative overview of BabelNet (we label nodes with English lexicalizations only): unlabeled edges are obtained from links in the Wikipages (e.g., *play* (theatre) links to *Musical theatre*), whereas labeled ones from WordNet (e.g., *play*\textsubscript{fr} \textit{has-part} *stage direction*\textsubscript{fr}).
Fig. 4. Translating Babel synsets based on a Machine Translation system. In order to fill lexical gaps (i.e., missing translations, typically for resource-poor languages), sense-annotated data are collected from SemCor and Wikipedia, and their most frequent translations are included as additional lexicalizations in the network.
## Structure of the BabelNet data set

### BabelNet 3.5: General statistics

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of languages</td>
<td>272</td>
</tr>
<tr>
<td>Total number of Babel synsets</td>
<td>13,801,844</td>
</tr>
<tr>
<td>Total number of Babel senses</td>
<td>119,036,997</td>
</tr>
<tr>
<td>Total number of concepts</td>
<td>6,066,256</td>
</tr>
<tr>
<td>Total number of Named Entities</td>
<td>7,735,588</td>
</tr>
<tr>
<td>Total number of lexico-semantic relations</td>
<td>380,239,084</td>
</tr>
<tr>
<td>Total number of glosses (textual definitions)</td>
<td>40,634,604</td>
</tr>
<tr>
<td>Total number of images</td>
<td>10,767,833</td>
</tr>
<tr>
<td>Total number of Babel synsets with at least one domain</td>
<td>1,558,806</td>
</tr>
<tr>
<td>Total number of compounds</td>
<td>743,296</td>
</tr>
<tr>
<td>Total number of Babel synsets with at least one picture</td>
<td>2,948,668</td>
</tr>
<tr>
<td>Total number of RDF triples</td>
<td>1,951,194,299</td>
</tr>
</tbody>
</table>
Electronic Dictionary Research

Wordnet and EuroWordnet

Wiki projects as Lexical Knowledge

BabelNet

The Papillon dictionary

The LexALP project
Overview of the Papillon Project

- Quickly extended itself to a larger consortium
  - with partners in Australia, Canada, China, France, Germany, Japan, Malaysia, Thailand, Vietnam...
  - …and accepting any new partner motivated enough to join the adventure...
- Aims at the development of a rich, “open source”, multilingual lexical database
Architecture of the Data

Macrostructure: An acception based multilingual lexical database

Français
Riz (plante monocotylédone)
Riz (grain)

Anglais
Rice (food grain)
Rice (seeds)

Japonais
御飯
米
稲

Malais
padi (unharvested grain)
nasi (cooked)
beras (uncooked)
Architecture of the Data
A microstructure inspired by Mel’čuk’s ECD and Polguère’s DICO

regretter,
v.tr.
sentiment LA personne X ~ SON action Y

GOVERNMENT PATTERN

X = I    Y = II
    1. N
1. N
    2. de V-inf

LEXICAL FUNCTIONS

QSyn : se repentir
S0 : regret#1
Able2 : (Que l’on peut R.) regrettable
Magn : (Intensément) beaucoup

Y étant grave, Magn : amèrement, cruellement; _se mordre les doigts_

EXAMPLES

1. C’est une décision qu’il va regretter cruellement.
2. Il ne regrette pas d’avoir investi 4 000 F dans ce nouveau programme.
gros fumeur lit. trans. big smoker
actual trans. heavy smoker
An MT system has to identify that *gros fumeur* is a collocation, something between a free combination and a full idiom
Mel’čuk’s ECD and Lexical Functions

Following Meaning-Text terminology, we call collocation a linguistic expression made up of two components:

- the base of the collocation: a full lexical unit which is “freely” chosen by the speaker on the basis of its meaning (e.g. ‘smoker’ → smoker);
- the collocate: a lexical unit or a multilexical expression which is chosen in a (partially) arbitrary way to express a given meaning and/or a grammatical structure contingent upon the choice of the base (e.g. ‘intense’ → heavy).
Rich bilingual dictionary
  - fumeur ⇔ smoker
  - gros fumeur ⇔ heavy smoker

Minimal bilingual dictionary
  - fumeur ⇔ smoker
  + rich monolingual dictionaries
    - intensification(fumeur) = gros
    - intensification(smoker) = heavy
Collocations are numerous and various in nature.

E.g.: COLÈRE ‘anger’

- colère aveugle/noire, lit. ‘blind/black anger’
- colère sourde/froide, lit. ‘deaf/cold anger’
- fou/ivre de colère, lit. ‘mad/drunken of anger’
- rouge/blanc de colère, lit. ‘red/white of anger’

etc.
Mel’čuk’s ECD and Lexical Functions: Collocation and Semantic Derivation

Intensification of rain:
- torrential (collocate)
- downpour (semantic derivation)
- torrential rain, downpour

Semantic derivations
- (quasi)synonymy/antonymy
- verbal, nominal, adjectival or adverbial derivations
- name of a participant or circonstant, e.g. crime is linked to author [of a crime] or criminal, victim, instrument [of a crime], etc.

Both types of lexical relation could and should be encoded by the same conceptual device.
Mel’čuk’s ECD and Lexical Functions: Collocations as Functions

Concept of LF: Zholkovskij & Mel’čuk 1965
Base-collocate relations are oriented

\[
\begin{align*}
\text{Magn} & \quad \text{(rain)} & = & \quad \text{torrential} \\
\text{function} & \quad \text{(base)} & = & \quad \text{collocate}
\end{align*}
\]
Architecture of the Data
A microstructure inspired by Mel’čuk’s ECD and Polguère’s DICO

regretter ,
v.tr.
sentiment LA personne X ~ SON action Y

GOVERNMENT PATTERN
X = I   Y = II
  1. N
  1. N
     2. de V-inf

LEXICAL FUNCTIONS
QSyn : se repenir
S0 : regret#1
Able2 : ( Que l’on peut R. ) regrettable
Magn : ( Intensément ) beaucoup
Y étant grave, Magn : amèrement , cruellement ; _se mordre les doigts_

EXAMPLES
1. C’est une décision qu’il va regretter cruellement.
2. Il ne regrette pas d’avoir investi 4 000 F dans ce nouveau programme.
4 Electronic Dictionary Research

5 Wordnet and EuroWordnet

6 Wiki projects as Lexical Knowledge

7 BabelNet

8 The Papillon dictionary

9 The LexALP project
The Alpine Convention gathers European countries who agree on certain rules on
- environment, spatial planning, transport infrastructure, ...
- for the Alpine Space (from Monaco to Slovenia)
LexALP Project Motivations

- The Alpine Convention gathers European countries who agree on certain rules on
  - environment, spatial planning, transport infrastructure, …
  - for the Alpine Space (from Monaco to Slovenia)
- But these rules need to be carefully worded
  - and negotiators often disagree on the wording of documents
  - (half the time of meetings spent in discussions about the minutes)
LexALP Project Motivations

- The Alpine Convention gathers European countries who agree on certain rules on
  - environment, spatial planning, transport infrastructure, ... 
  - for the Alpine Space (from Monaco to Slovenia)
- But these rules need to be carefully worded
  - and negotiators often disagree on the wording of documents
  - (half the time of meetings spent in discussions about the minutes)
- Moreover, these rules should be implemented by participating states
  - hence, adequate terminology should be used in each country
  - to correctly refer to existing laws
Examples of Legal Language Problems

- Different terms for the same meaning
  - chien drogue is used in European texts...
  - ...when texts from France use chien renifleur
Examples of Legal Language Problems

- **Different terms for the same meaning**
  - chien drogue is used in European texts...
  - ...when texts from France use chien renifleur

- **Different meanings for the same term**
  - Landeshauptmann in Bolzano (Italy)...
  - ...has much less power than an Austrian Landeshauptmann
Examples of **Legal** Language Problems

- **Different terms for the same meaning**
  - chien drogue is used in European texts...
  - ...when texts from France use chien renifleur

- **Different meanings for the same term**
  - Landeshauptmann in Bolzano (Italy)...
  - ...has much less power than an Austrian Landeshauptmann

- **Superficial translation leads to legal differences**
  - elezione suppletiva is commonly held whenever an elected deputy or senator either resigns or dies
  - Ersatzwahlen are very rare, as in Germany the first non-elected candidate is called to parliament
A Legal Language Harmonisation System
For Environment and Spatial Planning within the Multilingual Alps

- Describe legal terminology of the Alpine Convention (in its four languages)
- Harmonise it
A Legal Language Harmonisation System
For Environment and Spatial Planning within the Multilingual Alps

- Describe legal terminology of the Alpine Convention (in its four languages)
- Harmonise it

→ For use by negotiators and translators of the Alpine Convention
A Legal Language Harmonisation System
For Environment and Spatial Planning within the Multilingual Alps

- Describe legal terminology of the Alpine Convention (in its four languages)
- Harmonise it

→ For use by negotiators and translators of the Alpine Convention

- Link it to the equivalent/near-equivalent terms in the national legal systems
A Legal Language Harmonisation System
For Environment and Spatial Planning within the Multilingual Alps

- Describe legal terminology of the Alpine Convention (in its four languages)
- Harmonise it

→ For use by negotiators and translators of the Alpine Convention

- Link it to the equivalent/near-equivalent terms in the national legal systems

→ For use by jurist who will implement the law in other legal systems
LexALP Bibliographic Database

- Legal documents are collected (in raw text format)
- Along with their classification:
  - title of the document
  - short title of the document
  - abbreviation of the document
  - language
  - legal system
  - legal hierarchy (e.g. national, regional, supranational)
  - legal text type
  - translation status (original or translation)
  - subfield
Spatial planning and sustainable development

1. Conservation of nature and landscape protection
   1.1 Conservation of nature
   1.2 Landscape protection
   1.3 Protected areas
   1.4 Protection of flora and fauna
   1.5 Environmental protection
   1.6 Water protection
   1.7 Environmental impact assessment
   1.8 Natural disasters

2. Transport
   2.0 Transport policy
   2.1 Rail transport
   2.2 Road transport
   2.3 Air transport
   2.4 Sea transport
   2.5 Inland navigation
   2.6 Combined transport
   2.7 Passenger transport
   2.8 Transport of goods
   2.9 Transport networks
   2.10 Transport safety
   2.11 Contract of carriage

3. Regional economic development
   3.1 Regional policy
   3.2 Industry
   3.3 Trade
   3.4 Handicraft
   3.5 Tourism
   3.6 Employment
   3.7 Co-operation
   3.8 Energy

4. Rural areas
   4.1 Agriculture
   4.2 Forests
   4.3 Zootechnics
   4.4 Hunting, fisheries and fish farming
   4.5 Natural hazards

5. Urban areas
   5.1 Town planning
   5.2 Urban areas
   5.3 Building
LexALP Bibliographic Database

- Legal documents are collected (in raw text format)
- Along with their classification:
  - title of the document
  - short title of the document
  - abbreviation of the document
  - language
  - legal system
  - legal hierarchy (e.g. national, regional, supranational)
  - legal text type
  - translation status (original or translation)
  - subfield
  - + additional fields depending on the legal system
**Title:**

created the 10.03.2006 by GrP_EURAC

<table>
<thead>
<tr>
<th>abbreviation:</th>
<th>déc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>language:</td>
<td>fra</td>
</tr>
<tr>
<td>translation status:</td>
<td>OL</td>
</tr>
<tr>
<td>legal system:</td>
<td>EU</td>
</tr>
<tr>
<td>legal hierarchy:</td>
<td>supranational</td>
</tr>
<tr>
<td>legal text type:</td>
<td>décision</td>
</tr>
<tr>
<td>subfield1:</td>
<td>3.5 Tourism</td>
</tr>
<tr>
<td>official date:</td>
<td>13.07.1992</td>
</tr>
<tr>
<td>official journal date:</td>
<td>13.08.1992</td>
</tr>
<tr>
<td>official journal nb:</td>
<td>L 231</td>
</tr>
<tr>
<td>number:</td>
<td>92/421</td>
</tr>
<tr>
<td>identification number:</td>
<td>88836</td>
</tr>
<tr>
<td>show:</td>
<td>* doc in cache</td>
</tr>
<tr>
<td></td>
<td>* doc in www</td>
</tr>
</tbody>
</table>
LexALP Bibliographic Database

→ this leads to a bibliographic database that contains documents of the corpus, but also documents that are not available in the corpus
LexALP Document Structure

- The documents of each legal system are automatically annotated by the means of several scripts.
- The annotation is done on a structural level taking into account textual subdivisions (sentences etc.) as well as subdivisions particular to legal documents (chapters, articles, ...).

```xml
<div type="section" id="FR_U3_ENVIR-PL-L4-T2-44-14-11.xml.b.c0.se3">
  <p id="FR_U3_ENVIR-PL-L4-T2-44-14-11.xml.b.c0.se3.p0">
    <title id="FR_U3_ENVIR-PL-L4-T2-44-14-11.xml.b.c0.se3.p0.ti1">
      Article L420-3
    </title>
  </p>
  <p id="FR_U3_ENVIR-PL-L4-T2-44-14-11.xml.b.c0.se3.p1">
    <s id="FR_U3_ENVIR-PL-L4-T2-44-14-11.xml.b.c0.se3.p1.s1">
      Constitue un acte de chasse tout acte volontaire lié à la recherche, à la poursuite ou à l’attente du gibier ayant pour but ou pour résultat la capture ou la mort de celui-ci.
    </s>
  </p>
</div>
```
LexALP Corpus Structure

```
corpus_structure
ID    segment_type  segment_id  starting_point  ending_point  document_id

corpus_words
ID    word  position  document_id

document_info
ID    full_title  abbreviation  language  legal_system  ...

corpus_alignment
ID    DE_document_id  FR_document_id  IT_document_id  SI_document_id
```
LexALP Corpus Search

Search the corpus:

WORD: ~ schutz
LANGUAGE: deu
SEGMENT TYPE: s
LIMIT HITS: 5

document:
LEGAL SYSTEM: DE
LEGAL HIERARCHY: all
LEGAL TEXT TYPE:
SUBFIELD 1: Conservation of nature and landscape protection
SUBFIELD 2: all
SUBFIELD 3: all

KWIC search:

WORD: ita
LANGUAGE: show up to 5 context words
show up to 10 hits

SHOW KWIC
Output of your search for the word 'schutz':

<table>
<thead>
<tr>
<th>N</th>
<th>bib infos</th>
<th>segment</th>
<th>key</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DETAILS</td>
<td>In den Fällen des Satzes 1 hat der Arbeitgeber den betreffenden Beschäftigten Schutzkleidung und Atemschutzgeräte zur Verfügung zu stellen, die sie während der gesamten Dauer der erhöhten Exposition tragen müssen.</td>
<td>art. 11</td>
</tr>
<tr>
<td>2</td>
<td>DETAILS</td>
<td>(1) Im Rahmen der nach § 3 des Arbeitsschutzgesetzes zu treffenden Maßnahmen hat der Arbeitgeber für eine angemessene arbeitsmedizinische Vorsorge zu sorgen.</td>
<td>art. 15</td>
</tr>
<tr>
<td>3</td>
<td>DETAILS</td>
<td>5. die Fortentwicklung des betrieblichen Gesundheitschutzes bei Tätigkeiten mit Gefahrstoffen auf der Grundlage gewonnener Erkenntnisse.</td>
<td>art. 15</td>
</tr>
<tr>
<td>4</td>
<td>DETAILS</td>
<td>(2) Unbeschadet des § 22 des Arbeitsschutzgesetzes ist der zuständigen Behörde auf ihr Verlangen Folgendes mitzuteilen:</td>
<td>art. 19</td>
</tr>
<tr>
<td>5</td>
<td>DETAILS</td>
<td>3. die nach § 13 des Arbeitsschutzgesetzes verantwortlichen Personen,</td>
<td>art. 19</td>
</tr>
<tr>
<td>6</td>
<td>DETAILS</td>
<td>5. die geplanten Sicherheitsmaßnahmen zur Gewährleistung des Gesundheitschutzes und der Sicherheit der betroffenen Beschäftigten,</td>
<td>art. 20</td>
</tr>
<tr>
<td>7</td>
<td>DETAILS</td>
<td>(1) Zur Beratung in allen Fragen des Arbeitsschutzes zu Gefahrstoffen wird beim Bundesministerium für Wirtschaft und Arbeit der Ausschuss für Gefahrstoffe (AGS) gebildet, in dem fachkundige Vertreter der Arbeitgeberverbände, der Gewerkschaften, der Länderbehörden, der Träger der gesetzlichen Unfallversicherung und weitere fachkundige Personen, insbesondere der Wissenschaft, in angemessener Zahl vertreten sein sollen.</td>
<td>art. 21</td>
</tr>
<tr>
<td>8</td>
<td>DETAILS</td>
<td>d) das Gefährdungspotential der Untersuchungstechnik für den Beschäftigten. Bei der Wahrnehmung seiner Aufgaben berücksichtigt der Ausschuss für Gefahrstoffe die allgemeinen Grundsätze des Arbeitsschutzes nach § 4 des Arbeitsschutzgesetzes.</td>
<td>art. 21</td>
</tr>
<tr>
<td>9</td>
<td>DETAILS</td>
<td>(6) Die Geschäfte des Ausschusses führt die Bundesanstalt für Arbeitsschutz und Arbeitsmedizin.</td>
<td>art. 21</td>
</tr>
</tbody>
</table>
The LexALP term bank

- The term bank is a collection of “lexies”
- Each lexie represents a word sense (or acceptation) of the domain
  - ex: principe de précaution ≠ principe de précaution
  - ex: chien renifleur ≠ chien drogue
- Lexies are linked by interlingual relations called “axies”
Building the LexALP Term Bank
Monolingual step

1. Develop a quadrilingual term bank
   - French, German, Italian and Slovenian
   - based on Alpine Convention texts
Building the LexALP Term Bank
Monolingual step

- **it** trasporto intraalpino
- **it** traffico intraalpino
- **fr** trafic intra-alpin
- **fr** transport intra-alpin
- **fr** circulation intra-alpine
- **de** inneralpiner Verkehr
- **sl** znotrajalpski promet
LexALP Microstructure

LexALP edition interface

<table>
<thead>
<tr>
<th>legalSystem</th>
<th>circulation intra-alpine</th>
<th>n.f.</th>
<th>REJECTED</th>
<th>FINALISED</th>
</tr>
</thead>
<tbody>
<tr>
<td>geographical-code</td>
<td>INT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usage</td>
<td>frequency</td>
<td>infrequent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>is a technical term</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Domains:

- Transport

### Related terms:

- trafic intra-alpin
  - [harmonised ?]  
  - {termref: fra.trafic_intra-alpin.1010}  
  - Synonym

- transport intra-alpin
  - [harmonised ?]  
  - {termref: fra.transport_intra-alpin.1}  
  - Synonym
LexALP Microstructure

| Definition (created by: | | ) |
|------------------------|------------------|
| Trafic constitué de trajets ayant leur point de départ et/ou d'arrivée à l'intérieur de l'espace alpin. |

<table>
<thead>
<tr>
<th>Sources:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prot. Transp., art. 2</td>
</tr>
<tr>
<td>url:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contexts:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Les Parties contractantes s'engagent à mener une politique</td>
</tr>
<tr>
<td>url: <a href="http://www.convenzioni.eu">http://www.convenzioni.eu</a></td>
</tr>
</tbody>
</table>
Building the LexALP Term Bank

Multilingual steps

- Develop a quadrilingual term bank
  - French, German, Italian and Slovenian
  - based on Alpine Convention texts
Building the LexALP Term Bank
Multilingual steps

- **Italian (it)**: trasporto intraalpino
- **German (de)**: inneralpiner Verkehr
- **French (fr)**: trafic intra-alpin, transport intra-alpin, circulation intra-alpine
- **Slovenian (sl)**: znotrajalpski promet

Multilingual Text processing & Multilingual Lexical Databases
14th October 2015 122 / 134
Building the LexALP Term Bank

Multilingual steps

- Develop a quadrilingual term bank
  - French, German, Italian and Slovenian
  - based on Alpine Convention texts

- Link equivalents on the Alpine Convention level

→ We use "axies" (with an e) for the multilingual interoperability
Building the LexALP Term Bank
Multilingual steps

Italian
trasporto intraalpino
traffico intraalpino

German
inneralpiner Verkehr

French
trafic intra-alpin
transport intra-alpin
circulation intra-alpine

Slovene
znotrajalpski promet
Building the LexALP Term Bank

Multilingual steps

- Develop a quadrilingual term bank
  - French, German, Italian and Slovenian
  - based on Alpine Convention texts
- Link equivalents on the Alpine Convention level
- Harmonize the terms at the Alpine Convention level
Building the LexALP Term Bank

Multilingual steps

Italian
- trasporto intraalpino
- traffico intraalpino

Slovene
- znotrajalpski promet

German
- inneralpiner Verkehr

French
- trafic intra-alpin
- transport intra-alpin
- circulation intra-alpine

Slovene
- znotrajalpski promet
Building the LexALP Term Bank

Multilingual steps

- Develop a quadrilingual term bank
  - French, German, Italian and Slovenian
  - based on Alpine Convention texts
- Link equivalents on the Alpine Convention level
- Harmonize the terms at the Alpine Convention level
- Link these terms with related terms defined in national legal texts
  - Austria, France, Germany, Italy, Slovenia, Switzerland (Monaco, Liechtenstein)
  - based on national/regional legal texts
Building the LexALP Term Bank

Multilingual steps

Italian
principio di precauzione

Slovene
nacelo preventive

German
Vorsorgeprinzip

French
principe de précaution

principe de précaution

principe de précaution

principe de précaution
Part III

Ontologies and the Lexicon
Introduction

Lexical Systems (Polguère)

Ontolex/Lemon Model

DBnary
Structuring the dictionary as a graph

1. No more "entry" (but you need to type your nodes somehow)
Structuring the dictionary as a graph

1. No more "entry" (but you need to type your nodes somehow)
2. May be represented in RDF
Structuring the dictionary as a graph

1. No more "entry" (but you need to type your nodes somehow)
2. May be represented in RDF
3. Leads to the Lexical Web
Structuring the dictionary as a graph

1. No more "entry" (but you need to type your nodes somehow)
2. May be represented in RDF
3. Leads to the Lexical Web
4. Properties of arcs may be used to rationalize the MLDB data
   ▶ you can still provide a "natural" way to present your data
Introduction

Lexical Systems (Polguère)

Ontolex/Lemon Model

DBnary
Example: Mel’čuk ECD as a Lexical System (A. Polguère)
Example: Mel’čuk ECD as a Lexical System (A. Polguère)

```
Ce qu’on dit pour appeler ~
  If
  value
    « Minet ! »
  value
    « Minou ! »
  value
    « Petit ! »
```

```
CHAT#
  arg
```

```
Ce qu’on dit pour appeler ~(CHAT#)
```
Introduction

Lexical Systems (Polguère)

Ontolex/Lemon Model

DBnary
The Ontolex/Lemon Model (current w3c discussion group)
The Ontolex/Lemon Model (current w3c discussion group)
Introduction

Lexical Systems (Polguère)

Ontolex/Lemon Model

DBnary
Example: Wiktionary data as an RDF lexical resource (The DBnary project)

{{also|chát}}
==English==
{{wikipedia}}

===Pronunciation===
* {{IPA1/tʃæt/}}
* {{audio|en-us-chat.ogg|Audio (US)}}
* {{rhymes|æt}}

===Etymology 1===
Abbreviation of '[[chatter]]'.
{{ref|bird and louse also from chatter??}}

====Verb====
{{en-verb/chat/ting}}
[[image:Wikimania 2009 - Chatting (3).jpg|thumb|Two people '"chatting"'. (1) (2)]]

# To be [[engage]]d in informal [[conversation]].
#: "'She '"chatted''' with her friend in the cafe.'"
#: '"I like to '"chat''' over a coffee with a friend.'"
# To talk more than a few words.
#: "'I met my old friend in the street, so we '"chatted''' for a while.'"
# To exchange [[text]] or [[voice]] [[message]]s in [[real time]] through a [[computer network]], ..
#: '"Do you want to '"chat''' online later?''

====Translations====
{{trans-toplbe engaged in informal conversation}}
* Bulgarian: {{tlbg|бъбря}}
* [[Catalan]]: {{tlcalxerrar}}
* Danish: {{+da|snakke}} {{+da|sludre}}
Example: Wiktionary data as an RDF lexical resource (The DBnary project)
Example: Wiktionary data as an RDF lexical resource (The DBnary project)
Example: Wiktionary data as an RDF lexical resource (The DBnary project)
Example: Wiktionary data as an RDF lexical resource (The DBnary project)
DBnary as it is today

see http://kaiko.getalp.org/about-dbwnary
Current developments in DBnary

To be done (Master and PhDs still possible)

- Assigning translations sources to their correct LexicalSense
- Assigning translations targets to their correct LexicalSense
- Same for Lexico-semantic relations
- Aligning dbnary lexical senses with sense from other resources.

What do we need for this?

- Word Sense Disambiguation
- Cross lingual semantic similarity
Current developments in DBnary

To be done (Master and PhDs still possible)

- Assigning translations sources to their correct LexicalSense
- Assigning translations targets to their correct LexicalSense
- Same for Lexico-semantic relations
- Aligning dbnary lexical senses with sense from other resources.

What do we need for this?

- Word Sense Disambiguation
- Cross lingual semantic similarity
Conclusion

Lexical Resources are interesting

- Lexical Resources are now a key element in Semantic Analysis application.
- After losing some interest, they are now regaining attention.
- As we are now approaching the limits of statistical/learning based methods.
- Lexical Resources are a key element for future hybrid approaches...
Conclusion

Lexical Resources are interesting

- Lexical Resources are now a key element in Semantic Analysis application.
- After losing some interest, they are now regaining attention.
- As we are now approaching the limits of statistical/learning based methods.
- Lexical Resources are a key element for future hybrid approaches...

But still hard to produce

- Wiktionary/Wikipedia have changed the story.
- But data quality is variable.
- Lexical Architecture is still a debate (and English tends to take a central role...).
- We still need good data and non ethnologically centered multilingual data.