Building Multilingual Lexical Databases
a Tutorial in Bangkok

Gilles Sérasset

GETALP-LIG, Université Joseph Fourier, Grenoble, France

11th December 2007
Introduction
Foreword

We are here for a 6 hours tutorial

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 6 hours tutorial

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

Head of the Papillon project since 2000.
Part I

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

3 Computational Problems posed by ALL Lexical Data
   • Representing Lexical Data
   • String Encoding Problems
     • Presentation of the problem
     • What do we code?
     • UNICODE
   • Searching and Sorting
Monolingual Lexical Data
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
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,
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 describe it,
A First Definition

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

\textsuperscript{a}Webster’s New World College Dictionary, Fourth Edition, 2002

1. a collection of \textit{words},
2. each word comes with a \textit{set of information} that describe it,
3. there is an implictie \textit{access method} to each word.
Examples of monolingual dictionary

**compose** |ˈkʌmpɔz|
---
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 |kəmˈpɔz|**
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 <kɪmˈpɔːz>** 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."
A 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).
A 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 are no definition or description of each word (this is left to the dictionary),
A 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**, 
2. there are no definition or description of each 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 does 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 does 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?
Bilingual Lexical Data
Bilingual Dictionary: an example

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

**Écouter**: US - UK

---

**Principal Translations/Principales traductions**

- **compose** (write)  
  vtr écrire ⇒
- **compose** (music, literature)  
  vtr composer ⇒ (musique, littérature)
- **compose** (put together)  
  vtr composer ⇒
- **compose** (technical)  
  vtr constituer ⇒ (composer)

---

**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?
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>иностранец</td>
<td>外侨</td>
<td>أجنبي؛ غريب； دخيل</td>
</tr>
<tr>
<td>Judicial Support</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>Division</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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>جزمة خدماتية</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) المتة</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 الأمين 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; nuclear weapon</td>
<td>bombe nucléaire; arme nucléaire</td>
<td>bomba nuclear; arma nuclear</td>
<td>ядерное оружие</td>
<td>核弹；核武器</td>
<td>قنبلة نووية: سلاح نووي</td>
</tr>
</tbody>
</table>

*Table showing multilingual terms and translations.*
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 composer /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 do 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 do 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 has to be handled...
Dictionaries and Other Lexical Data

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

Computational Problems posed by ALL Lexical Data

- Representing Lexical Data
- String Encoding Problems
  - Presentation of the problem
  - What do we code?
  - UNICODE
- Searching and Sorting
«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 define 127 characters (using 7 bits). The other 127 characters usually depended on the OS/machine.

- **ANSI extended ASCII**
- **OEM extended ASCII (first IBM PCs)**
“National” encodings also used the 127 higher characters to represent characters of their script\(^1\). This, plus the fact that everybody called this THE ASCII code... Hence, begun what can be cold:

“The happy mess of encodings...”

\(^1\)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 (Hungarian, Polish, ...),
- ISO-8859-3 for Esperanto, Galician, Maletese, Turkish,
- ISO-8859-4 for Estonian, Lithuanian, Litton...
- 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

Language 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 following characters use 1 or 2 bytes.

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

**e.g.**: S-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 do present a major problem: they are exclusive one of the other!

Hence, it is impossible without a non-standard treatment\(^2\), to mix different languages...

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

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

\(^2\)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?

\[
\begin{align*}
\delta & \quad A \\
FBA6 & \quad A \\
\sim & \quad A \\
FBA7 & \quad a \\
FBA8 & \quad a \\
FBA9 & \quad a
\end{align*}
\]
Some definitions

**Character**  Abstract entity, considered as atomic, use 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?

\[
\begin{align*}
\text{FBA6} & \quad A \\
\text{FBA7} & \quad A \\
\text{FBA8} & \quad a \\
\text{FBA9} & \quad a
\end{align*}
\]
What do we code?

How many characters are there below?

Un seul caractère, quatre Glyphes, 5 points de codes
What do we code?

How many characters are there below?

1 caractère, 1 glyphs, ≠ fontes
What do we code?

How many characters are there below?
What do we code?

How many characters are there below?

![Diagram showing characters and their corresponding codes: FBA6, FBA7, FBA8, FBA9, and A. The diagram also includes a note: 2 caractères (A majuscule latin, A majuscule cyrillique, ...)
ISO 10646 norm defines a set of abstract characters, along with their name in English and French. It assigns code points to these characters. The UNIOCODE consortium, use these code points and defines the way they are represented in memory along with other element necessary to the representation of texts.
Conception Principles

<table>
<thead>
<tr>
<th>Characters, not glyphs</th>
<th>The UNICODE standard encodes characters and not glyphs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic</td>
<td>Characters have a well defined semantic.</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>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 bijective convertibility is guarantied between UNICODE and other norms.</td>
</tr>
</tbody>
</table>
Two Antagonist Principles

A bijective 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…
Representation Levels

The definition model of characters is based on different levels:

1. abstract character repertoire
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 [correspondance between storage units and serialized sequence of bytes]
5. transfer encodings [reversible transformation of a sequence of bytes]
<table>
<thead>
<tr>
<th>Representation Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stocké</strong></td>
</tr>
<tr>
<td>UTF-8</td>
</tr>
<tr>
<td>C3 85</td>
</tr>
<tr>
<td>E2 84 AB</td>
</tr>
<tr>
<td>F3 B0 80 80</td>
</tr>
<tr>
<td>41 CC 8A</td>
</tr>
<tr>
<td>UTF-16</td>
</tr>
<tr>
<td>00C5</td>
</tr>
<tr>
<td>212B</td>
</tr>
<tr>
<td>DB80 DC00</td>
</tr>
<tr>
<td>0041 030A</td>
</tr>
<tr>
<td>UTF-32</td>
</tr>
<tr>
<td>0000 00C5</td>
</tr>
<tr>
<td>212B</td>
</tr>
<tr>
<td>000F 0000</td>
</tr>
<tr>
<td>0000 0041</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>UTF-32LE</td>
</tr>
<tr>
<td>C5 00 00 00</td>
</tr>
<tr>
<td>2B 21 00 00</td>
</tr>
<tr>
<td>00 0F 00 00</td>
</tr>
<tr>
<td>41 00 00 0A</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Sérialisé</strong></td>
</tr>
<tr>
<td>UTF-8</td>
</tr>
<tr>
<td>C3 85</td>
</tr>
<tr>
<td>E2 84 AB</td>
</tr>
<tr>
<td>F3 B0 80 80</td>
</tr>
<tr>
<td>41 CC 8A</td>
</tr>
<tr>
<td>UTF-16BE</td>
</tr>
<tr>
<td>00 00C5</td>
</tr>
<tr>
<td>21 2B</td>
</tr>
<tr>
<td>DB 80 DC 00</td>
</tr>
<tr>
<td>00 41 03 0A</td>
</tr>
<tr>
<td>UTF-16LE</td>
</tr>
<tr>
<td>C5 00 00</td>
</tr>
<tr>
<td>2B 21</td>
</tr>
<tr>
<td>80 DB 00 DC</td>
</tr>
<tr>
<td>41 00 0A 03</td>
</tr>
<tr>
<td>UTF-32BE</td>
</tr>
<tr>
<td>00 00 00 C5</td>
</tr>
<tr>
<td>00 00 21 2B</td>
</tr>
<tr>
<td>00 0F 00 00</td>
</tr>
<tr>
<td>00 00 00 41</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>UTF-32LE</td>
</tr>
<tr>
<td>C5 00 00 00</td>
</tr>
<tr>
<td>2B 21 00 00</td>
</tr>
<tr>
<td>00 00 0F 00</td>
</tr>
<tr>
<td>41 00 00 00</td>
</tr>
</tbody>
</table>
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 implies many problems:

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

**example in C**

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

This program print 0 or 1, depending on the machine on which it runs...
The reason is that in C, `char` means byte NOT character.
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 implies many problems:

**example in C**

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

This program print 0 or 1, depending on the machine on witch it runs... The reason is that in C, `char` means byte 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 ≤ code point ≤ 0x7F) is represented as is,
- Any character ≥ 0x80 is represented using the “little train” technique.

Le “little train”

Let \( x \) be a code point using \( n \) significative bits (7 > \( n \) > 21)
\[
x = b_n b_{n-1} \ldots b_1
\]
We scatter these bits in “wagons”, pulled by a strong enough locomotive
(110xxxxx, 1110xxxx ou 11110xxx).
- if \((7 < n \leq 11)\) \(\rightarrow 110xxxxx \quad 10xxxxxx\)
- if \((11 < n \leq 16)\) \(\rightarrow 1110xxxx \quad 10xxxxxx \quad 10xxxxxx\)
- if \((16 < n \leq 21)\) \(\rightarrow 11110xxx \quad 10xxxxxx \quad 10xxxxxx \quad 10xxxxxx\)
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 appear 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.
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 blau, then search Blau, You’ll see that case difference is not taken into account by search engine,
3. Search Maß
4. 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)
5. You will see that Maß in now rendered as MASS
6. Search MASS, what do you see ?
7. As another illustration, search Wörterbuch then Woerterbuch
Understanding the Problem: a Practical Example

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

2. search blau, then Blau, You’ll see that case difference is not taken into account by search engine,

3. search Maß

4. search MASS, what do you see ?

5. As another illustration, search Wörterbuch then Woerterbuch
Understanding the Problem: a Practical Example

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

2. search blau, then Blau, You’ll see that case difference is not taken into account by search engine,

3. search Maß

4. search MASS, what do you see ?

5. As another illustration, search Wörterbuch then Woerterbuch

6. Now, select the English definition dictionary,

7. Search ae then ä

8. Search maß

You’ll see that search is adapting itself to the resource.
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
Formalizing these Problems

Let \( x \) and \( y \) be two Strings:

- **Searching:** \( x = y \)?
  
  ex: Maß = Mass?

- **Sorting:** \( x < y \)?
  
  ex: coté < côté?

- **Searching a sub string:** \( \exists v, w \mid x = vyw \)?
  
  ex: "churro" contient "c"?
Searching and sorting operations are not functions of a string, but rather functions on the interpretation of the String in a certain language. For instance, French names should be sorted using the German phonebook order if users of the phonebook are German.

You need to be able to declare, when comparing, the order you want to use.

The exact choice of order depends on:

- the language of the text
- the language of the user
- the particular application
Collation: definition

Collating:

1. (transitive) To examine diverse documents et cetera in order to discover similarities and differences.

2. (transitive) To assemble something in a logical sequence.

1922, Virginia Woolf, Jacob’s Room, Vintage Classics, paperback edition, page 101 Detest your own age. Build a better one. And to set that on foot read incredibly dull essays upon Marlowe to your friends. For which purpose one must collate editions in the British Museum.

3. (transitive) To sort multiple copies of printed documents into sequences of individual page order, one sequence for each copy, especially before binding.

In our case, collation defines equality and order of 2 texts.

Warning: an collation algorithm defines the order. It does not sort.
Comparing at different levels

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>Base Character</td>
<td>role &lt; roles &lt; rule</td>
</tr>
<tr>
<td>L2</td>
<td>Accents</td>
<td>role &lt; rôle &lt; roles</td>
</tr>
<tr>
<td>L3</td>
<td>Case</td>
<td>role &lt; Role &lt; rôle</td>
</tr>
<tr>
<td>L4</td>
<td>Punctuation</td>
<td>role &lt; “rôle” &lt; Role</td>
</tr>
</tbody>
</table>
Specific Problems

Unicode characters that are considered as canonically equivalent are sorted the same way.

<table>
<thead>
<tr>
<th>Å</th>
<th>U+212B ANGSTROM SIGN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>U+00C5 LATIN CAPITAL LETTER A WITH RING ABOVE</td>
</tr>
<tr>
<td>A + °</td>
<td>U+0041 LATIN CAPITAL LETTER A, U+030A COMBINING RING ABOVE</td>
</tr>
</tbody>
</table>

Context should be taken into account

Contractions  
H < Z, but  
CH > CZ

Expansions  
OE < Æ < OF

French uses reverse level 2 string

Normal Accent Ordering  
cote < coté < côte < côté

French Accent Ordering  
cote < côte < coté < côté
### Parameters of the Collation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>the main parameter</td>
</tr>
<tr>
<td>Force</td>
<td>the number of level considered during collation</td>
</tr>
<tr>
<td>Case ordering</td>
<td>uppercase before lowercase ?</td>
</tr>
<tr>
<td>User rules</td>
<td>used to specify application specific rules (e.g. “&amp;” ≡ “and”)</td>
</tr>
<tr>
<td>Parameter merging</td>
<td>merges collation parameters of different orders (e.g. latin script sorted as French, Thai words as Thai)</td>
</tr>
<tr>
<td>Order between scripts Numbers</td>
<td>e.g. Latin &lt; Greek &lt; Cyrillic vs. Greek &lt; Cyrillic &lt; Latin use numerical order instead of lexicographical order (e.g. Figure 2b &lt; Figure 10a).</td>
</tr>
</tbody>
</table>
Overview of the Unicode Collation Algorithm

Characters (or character sequences) are assigned one (or several) collation elements.

A **collation element** is an ordered list of 3 (or more) 16 bits weights.

For each collation element \( X \), we note \( X_1, X_2, X_3, X_4, \ldots \) this list of weights.

<table>
<thead>
<tr>
<th>Character</th>
<th>Collation Element</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0300 &quot;&quot;</td>
<td>[0000.0021.0002]</td>
<td>COMBINING GRAVE ACCENT</td>
</tr>
<tr>
<td>0061 &quot;a&quot;</td>
<td>[06D9.0020.0002]</td>
<td>LATIN SMALL LETTER A</td>
</tr>
<tr>
<td>0062 &quot;b&quot;</td>
<td>[06EE.0020.0002]</td>
<td>LATIN SMALL LETTER B</td>
</tr>
<tr>
<td>0063 &quot;c&quot;</td>
<td>[0706.0020.0002]</td>
<td>LATIN SMALL LETTER C</td>
</tr>
<tr>
<td>0043 &quot;C&quot;</td>
<td>[0706.0020.0008]</td>
<td>LATIN CAPITAL LETTER C</td>
</tr>
<tr>
<td>0064 &quot;d&quot;</td>
<td>[0712.0020.0002]</td>
<td>LATIN SMALL LETTER D</td>
</tr>
</tbody>
</table>

Note: the grave accent is **ignorable** at level 1 (it is equal to 0x0000 at level 1).
<table>
<thead>
<tr>
<th>String</th>
<th>cáb</th>
</tr>
</thead>
<tbody>
<tr>
<td>normalised string</td>
<td>ca´b</td>
</tr>
<tr>
<td>Collation Elements</td>
<td>[0706.0020.0002], [06D9.0020.0002], [0000.0021.0002], [06EE.0020.0002]</td>
</tr>
<tr>
<td>Sort key</td>
<td>0706 06D9 06EE 0000 0020 0021 0020 0020 0000 0002 0002 0002 0002</td>
</tr>
</tbody>
</table>
## A Comparison Example

<table>
<thead>
<tr>
<th>String</th>
<th>Sort Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>cab</td>
<td>0706 06D9 06EE 0000 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>Cab</td>
<td>0706 06D9 06EE 0000 0020 0020 0020 0000 0008 0002 0002</td>
</tr>
<tr>
<td>cáb</td>
<td>0706 06D9 06EE 0000 0020 0020 0021 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>cabb</td>
<td>0706 06D9 06EE 06EE 0000 0020 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>dab</td>
<td>0712 06D9 06EE 0000 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
</tbody>
</table>
### A Comparison Example

<table>
<thead>
<tr>
<th>String</th>
<th>Sort Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>cab</td>
<td>0706 06D9 06EE 0000 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>Cab</td>
<td>0706 06D9 06EE 0000 0020 0020 0020 0000 0008 0002 0002</td>
</tr>
<tr>
<td>cáb</td>
<td>0706 06D9 06EE 0000 0020 0020 0021 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>cabb</td>
<td>0706 06D9 06EE 06EE 0000 0020 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>dab</td>
<td>0712 06D9 06EE 0000 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
</tbody>
</table>

Difference between uppercase and lowercase is handled at level 3.
A Comparison Example

<table>
<thead>
<tr>
<th>String</th>
<th>Sort Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>cab</td>
<td>0706 06D9 06EE 0000 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>Cab</td>
<td>0706 06D9 06EE 0000 0020 0020 0020 0000 0008 0002 0002</td>
</tr>
<tr>
<td>cáb</td>
<td>0706 06D9 06EE 0000 0020 0020 0021 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>cabb</td>
<td>0706 06D9 06EE 06EE 0000 0020 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>dab</td>
<td>0712 06D9 06EE 0000 0020 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
</tbody>
</table>

Accents are handled at level 2
### A Comparison Example

<table>
<thead>
<tr>
<th>String</th>
<th>Sort Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>cab</td>
<td>0706 06D9 06EE 0000 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>Cab</td>
<td>0706 06D9 06EE 0000 0020 0020 0020 0000 0008 0002 0002</td>
</tr>
<tr>
<td>cáb</td>
<td>0706 06D9 06EE <strong>0000</strong> 0020 0020 0021 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>cabb</td>
<td>0706 06D9 06EE <strong>06EE</strong> 0000 0020 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
<tr>
<td>dab</td>
<td>0712 06D9 06EE 0000 0020 0020 0020 0000 0002 0002 0002</td>
</tr>
</tbody>
</table>

Level separator, which is strictly lower than any valid weight, allows for the comparison of a word and its prefix.
Collation Elements can be used for string/substring searching. Indeed, they reflect UNICODE canonical equivalence, as well as language specific character interpretations (e.g. ß = ss in German).

Ignoring level 3 gives a case insensitive search,
Ignoring level 2 gives a diacritic insensitive search.
Collation and Searching

But things are not so easy:

1. the length of a pattern and the match may differ (e.g. aa should match å in Danish),
2. due to ignorable characters (at a different level), a match can occur at different position (e.g. abc can match abc, abc- or -abc- in a local than ignores dashes),
3. a match boundary should (or should not) appear inside a contraction (e.g. c should (or should not ?) match churro in Spanish ?)
4. same goes for expansions (e.g. should ba match bæ?)
5. if match boundary can occur inside contraction or expansion, should c, (ç) match in c^, (c)
Part II

Example of Multilingual Lexical Databases
4 Electronic Dictionary Research

5 Wordnet and EuroWordnet

6 The Papillon dictionary

7 The LexALP project

8 Summary
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 M¥).

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, coocurrence dictionary and corpus (250000 analysed sentences, 10-20M raw sentences).

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

- English Word Dictionary
  - 200000 general
  - 100000 terms
  - 250000 sentences
  - English Cooccurrence 300000 words

- Japanese Word Dictionary
  - 200000 general
  - 100000 terms
  - 250000 sentences
  - Japanese Cooccurrence 300000 words

- 640000 concepts + description
- 400000 bilingual entries
EDR concept dictionary

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

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

5 Wordnet and EuroWordnet

6 The Papillon dictionary

7 The LexALP project

8 Summary
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><strong>155287</strong></td>
<td><strong>117659</strong></td>
<td><strong>206941</strong></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
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 correspondance between ILI and English synset.
Eurowordnet: an example

- {stream; watercourse}
  - river
    - eq_synonym
      - {cours d'eau; ruisseau}
      - eq_synonym
        - fleuve; rivière
        - eq_synonym
          - ruisseau
          - ruisselet
            - eq_synonym
              - fiume
              - fiume
            - eq_synonym
              - binnenwater
            - eq_near_synonym
              - rivier
Eurowordnet: an example
Analysis of Wornet/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.
Electronic Dictionary Research

Wordnet and EuroWordnet

The Papillon dictionary

The LexALP project

Summary
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
- 米
- 稻

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 repen
tir
S0 : regret#1
Able2 : (Que l’on peut R.) regrettable
Magn : (Intensément) beaucoup
Y étant grave, Magn : amèrement, cruellement ; _se mordre les doigs_

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.
Mel’čuk’s ECD and Lexical Functions: Motivations

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).
Mel’čuk’s ECD and Lexical Functions

Rich bilingual dictionary
- fumeur $\Leftrightarrow$ smoker
- gros fumeur $\Leftrightarrow$ heavy smoker

Minimal bilingual dictionary
- fumeur $\Leftrightarrow$ 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

\[
\text{Magn} \quad \text{(rain)} \quad = \quad \text{torrential} \\
\text{function} \quad \text{(base)} \quad = \quad \text{collocate}
\]
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

QSyzn : 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.
Electronic Dictionary Research

Wordnet and EuroWordnet

The Papillon dictionary

The LexALP project

Summary
LexALP Project Motivations

- The Alpine Convention gathers European countries who agrees 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)

Gilles Sérasset (GETALP-LIG)
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

[Image of multilingual lexicon]
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

abbreviation: déc.
language: fra
translation status: OL
legal system: EU
legal hierarchy: supranational
legal text type: décision
subfield1: 3.5 Tourism
official date: 13.07.1992
official journal date: 13.08.1992
official journal nb: L 231
number: 92/421
identification number: 88836
show: * doc in cache
* doc in www
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 |
```
## 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 Fortwicklung des betrieblichen Gesundheitsschutzes 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 Gesundheitsschutzes 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 $\neq$ principe de précaution
  - ex: chien renifleur $\neq$ chien drogue
- Lexies are linked by interlingual relations called “axies”
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** trafi intra-alpin
- **fr** transport intra-alpin
- **fr** circulation intra-alpine
- **de** inneralpiner Verkehr
- **sl** znotrajalpski promet
LexALP Microstructure

### LexALP Edition Interface

**legalSystem:** AC circulation intra-alpine  
**geographical-code:** INT  
**Usage:** frequency: infrequent  
is a technical term?  

<table>
<thead>
<tr>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transport</td>
<td></td>
</tr>
</tbody>
</table>

#### Domains:

<table>
<thead>
<tr>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>trafic intra-alpin</td>
<td></td>
</tr>
</tbody>
</table>

#### Related terms:

<table>
<thead>
<tr>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>trafic intra-alpin</td>
<td>[termref: fra.trafic_intra-alpin.1010] Synonym</td>
</tr>
<tr>
<td>transport intra-alpin</td>
<td>[termref: fra.transport_intra-alpin.1] Synonym</td>
</tr>
</tbody>
</table>
### Definition (created by:)

[Trafic constitué de trajets ayant leur point de départ et/ou d’arrivée à l’intérieur de l’espace alpin.

### Sources:

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

### Contexts:

<table>
<thead>
<tr>
<th>Les Parties contractantes s’engagent à mener une politique</th>
</tr>
</thead>
<tbody>
<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**: trasporto intraalpino
- **Italian**: traffico intraalpino
- **French**: trafic intra-alpin
- **French**: transport intra-alpin
- **French**: circulation intra-alpine
- **German**: inneralpiner Verkehr
- **Slovenian**: znotrajnalpski 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

→ 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 intra-alpino
- traffico intra-alpino

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

Italian

principio di precauzione

French

principe de précaution

Slovene

nacelo preventive

German

Vorsorgeprinzip

principe de précaution
Electronic Dictionary Research

Wordnet and EuroWordnet

The Papillon dictionary

The LexALP project

Summary
What have we learned until then?

- many different lexical resources do exist nowadays,
- many more will be built in the future,
What have we learned until then?

- many different lexical resources do exist nowadays,
- many more will be built in the future,
- each lexical database defines its own macro-structure
  - monolingual: usually 1 volume for 1 language,
  - bilingual: usually 2 independent volumes ($l_1 \rightarrow l_2 + l_2 \rightarrow l_1$)
  - multi-bilingual: either same as bilingual, but with more independent volumes, or a set of monolingual volume plus a set of bilingual links volumes
  - fork dictionaries: either a unique volume or monolingual volume + translation links volumes, only one source language,
  - multilingual: using monolingual volumes + one central, pivot volume;
What have we learned until then?

- many different lexical resources do exist nowadays,
- many more will be built in the future,
- each lexical database defines its own macro-structure
  - monolingual: usually 1 volume for 1 language,
  - bilingual: usually 2 independent volumes ($l_1 \rightarrow l_2 + l_2 \rightarrow l_1$)
  - multi-bilingual: either same as bilingual, but with more independent volumes, or a set of monolingual volume plus a set of bilingual links volumes
  - fork dictionaries: either a unique volume or monolingual volume + translation links volumes, only one source language,
  - multilingual: using monolingual volumes + one central, pivot volume;
- each lexical database defines its own micro-structure
  - Here, there is no real limit to the imagination of lexicologists...
Part III

Building MLDB: a mixture of Linguistic and Computer Problems
Analysing the problem

Jibiki: an open source platform for MLDB collaborative building

Services provided and how they work
Analysing the problem: economic aspects

- Building a rich bilingual dictionary is expensive...
Analysing the problem: economic aspects

- Building a rich bilingual dictionary is expensive...
  - ex: EDR: $\geq 350000$ words in Japanese-English
  - $\rightarrow 14305$ M¥
  - $\rightarrow$ VERY expensive
Analysing the problem: economic aspects

- Building a rich bilingual dictionary is expensive...
  - ex: EDR: > 350000 words in Japanese-English
  - → 14305 M¥
  - → VERY expensive

- How can we reduce the costs?
  - Share the costs → using a general structure, adaptable to many applications
  - Reduce the costs → reusing existing data whenever possible, implementing adequate tools
  - Federate the costs → allow anyone to contribute and use the data
Analysing the Problem: competences

- Dictionaries play a central role in human and artificial Language Processing tasks
- But building a dictionary is highly difficult
Analysing the Problem: competences

- Dictionaries play a central role in human and artificial Language Processing tasks
- But building a dictionary is highly difficult
- It requires linguistic skills
  - to define a valid structure
  - and produce correct entries
Analysing the Problem: competences

• Dictionaries play a central role in human and artificial Language Processing tasks

• But building a dictionary is highly difficult

• It requires linguistic skills
  ▶ to define a valid structure
  ▶ and produce correct entries

• It requires rigorous organisational skills
  ▶ to manage different lexicographers
  ▶ with different skills
Analysing the Problem: competences

- Dictionaries play a central role in human and artificial Language Processing tasks.
- But building a dictionary is highly difficult.
- It requires linguistic skills:
  - to define a valid structure
  - and produce correct entries
- It requires rigorous organisational skills:
  - to manage different lexicographers
  - with different skills
- It requires computational skill:
  - to ensure well-formedness/validity of the entries
  - to avoid losing data by mistake
Analysing the Problem: competences

- Dictionaries play a central role in human and artificial Language Processing tasks
- But building a dictionary is highly difficult
- It requires linguistic skills
  - to define a valid structure
  - and produce correct entries
- It requires rigorous organisational skills
  - to manage different lexicographers
  - with different skills
- It requires computational skill
  - to ensure well-formedness/validity of the entries
  - to avoid losing data by mistake

→ The linguistic skill will always be necessary
Analysing the Problem: competences

- Dictionaries play a central role in human and artificial Language Processing tasks
- But building a dictionary is highly difficult
- It requires linguistic skills
  - to define a valid structure
  - and produce correct entries
- It requires rigorous organisational skills
  - to manage different lexicographers
  - with different skills
- It requires computational skill
  - to ensure well-formedness/validity of the entries
  - to avoid losing data by mistake

→ But the other skills should be optional
Analysing the Problem: requirements

- For this, we need a dictionary building system
- which allows team work
- which provides integrity of data (backup/history/rollback...)
- which is customisable
  - to accept different micro-structures (structure of each entry)
  - to accept different macro-structure (monolingual, bilingual, multi-bilingual, pivot based, ...)
- with minimal knowledge of computer science
Analysing the Problem: requirements

- For this, we need a dictionary building system
- which allows team work
- which provides integrity of data (backup/history/rollback...)
- which is customisable
  - to accept different micro-structures (structure of each entry)
  - to accept different macro-structure (monolingual, bilingual, multi-bilingual, pivot based, ...)
- with minimal knowledge of computer science

→ No system was able to fulfil all these requirements
Analysing the Problem: requirements

- For this, we need a dictionary building system
- which allows team work
- which provides integrity of data (backup/history/rollback...)
- which is customisable
  - to accept different micro-structures (structure of each entry)
  - to accept different macro-structure (monolingual, bilingual, multi-bilingual, pivot based, ...)
- with minimal knowledge of computer science

→ We developed the Jibiki platform
Analysing the problem

Jibiki: an open source platform for MLDB collaborative building

Services provided and how they work
The Jibiki platform
A community web site development platform offering several services

- Dictionary access for many different dictionaries
- Dictionary edition for selected dictionaries
- Community services
  - Mailing list archive
  - Documentation pool shared by partners
Implementation

Papillon Application (java + enhydra)

presentation layer
- HTML
- CSS
- javascript
- CGI
- WML
- xhtml
- chtml

business layer
- Data validation
- Contributions management
- Requests management
- Users/Groups
- Mailing list archive
- Information sharing

data layer
- JDBC
- Lexie
- axie
- Dico
- Volume
- Historique
- Utilisateur
- Message
- Information

serveur HTTP (apache)

Relational database (PostgreSQL)
XML-UTF8

Gilles Sérasset (GETALP-LIG)
Analysing the problem

Jibiki: an open source platform for MLDB collaborative building

Services provided and how they work
We used the Jibiki plateform to set-up LexALP website
it offers access to the term bank...
and allows for dictionary edition
The LexALP website

Simple dictionary access

[Image of a web interface for LexALP with a search bar labeled 'Word: trafic', 'Source: French', 'Target: All', and a 'Go' button.]
The LexALP website
Simple dictionary access

**TRAIFIC INTRA-ALPIN, n.m.** [HARMONISED/FINALISED]

*TRANSPORT INTRA-ALPIN [kwic] (Synonym)*
*CIRCULATION INTRA-ALPIN [kwic] (Synonym)*

[T]rafic constitué de trajets ayant leur point de départ et/ou d'arrivée à l'intérieur de l'espace alpin.

*Source:* Prot. Transp., art. 2
*Context:* Des projets routiers à grand débit pour le trafic intra-alpin peuvent être réalisés, si [...] les besoins en matière de transports ne peuvent être satisfaits, ni par une meilleure utilisation des capacités routières et ferroviaires existantes, ni par l'extension ou la construction d'infrastructures ferroviaires ou fluvio-maritimes, ni par l'amélioration d'un transport combiné, ni par d'autres mesures relatives à l'organisation des transports [...].

**INNERALPINER VERKEHR, n.m.** [HARMONISED/FINALISED]

Verkehr mit Ziel und Quelle im Alpenraum inklusive Verkehr mit Ziel oder Quelle im Alpenraum.

*Source:* Prot. Verk., Art. 2/Cieri
*Context:* Die Vertragsparteien verpflichten sich zu einer nachhaltigen Verkehrspolitik, die [...] den inneralpinen und alpenquerenden Verkehr durch Steigerung der Effektivität und Effizienz der Verkehrssysteme und durch Förderung umwelt- und ressourcenschonenderer Verkehrsträger unter wirtschaftlich tragbaren Kosten gewährleistet [...].

**TRAFFICO INTRAALPINO, n.m.** [HARMONISED/FINALISED]

*TRASPORTO INTRAALPINO [kwic] (Synonym)*

Trafico o trasporto con origine e destinazione all'interno del territorio alpino incluso il traffico o trasporto con origine o destinazione nel territorio alpino.

*Source:* Prot. Trasp., art. 2/Cieri
*Context:* Le Parti contraenti si impegnano ad attuare una politica sostenibile dei trasporti tesa a [...] ridurre gli effetti negativi e i rischi derivanti dal traffico intraalpino e transalpino ad un livello che sia tollerabile per l'uomo, la fauna e la flora e il loro habitat [...].

**ZNOTRAJALPSKI PROMET, n.m.** [HARMONISED/UNPROCESSED]

Promet s ciljem in izvorom v alpskem prostoru (notranji promet), vključno s prometom s ciljem ali izvorom v alpskem prostoru.

*Source:* Prot. Promet, čl. 2
*Context:* Cestni projekt višega reda za znotrajalpski promet se lahko udejanji le takrat, če [...] potreb po prevoznih zmogljivostih ni moč zadovoljiti z boljšo izkoriščenostjo obstoječih cestnih in železniških zmogljivosti, z rekonstrukcijo ali novogradnjo železniških in plovnih infrastruktur ter z izboljšanjem kombiniranega prometa kot tudi z nadaljnjimi prometno-organizacijskimi ukrepi [...].
The LexALP website

Edit terms and interlingual links

**LexALP edition interface**

**legalSystem:** AC

**circulation intra-alpine**

**n.f.**

**[REJECTED]**

**FINALISED**

**geographical-code:** INT

**Usage:** frequency:

**infrequent**

**is a technical term?**

**Domains:**

- **Transport**

**Related terms:**

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

- **[harmonised ?]** transport intra-alpin
  - **[termref]:** fra.transport_intra-alpin.1
  - **Synonym**

**Definition (created by):**

[Trafic constitué de trajets ayant leur point de départ et/ou d’arrivée à l’intérieur de l’espace alpin.

**Sources:**

- Prot. Transp., art. 2

**Contexts:**

- Les Parties contractantes s’engagent à mener une politique

**Notes:**

Gilles Sérasset (GETALP-LIG)  
Building Multilingual Lexical Databases  
11th December 2007

110 / 116
What is the necessary work for this?
Providing dictionary access

1. Decide on the structure of your lexical database
   - Monolingual, bilingual, multi-bilingual, interlingual
   - Here, we chose an interlingual approach
   - with 5 monolingual dictionaries and 1 pivot
What is the necessary work for this?

Providing dictionary access

1. Decide on the structure of your lexical database
   - Monolingual, bilingual, multi-bilingual, interlingual
   - Here, we chose an interlingual approach
   - with 5 monolingual dictionaries and 1 pivot

2. Then, decide on the structure of your dictionary
   - Your only constraint is: use an XML structure
What is the necessary work for this?

Providing dictionary access

<volume volume-language="fra">
  <entry lang="fra" status="UNKNOWN"
    process_status="UNPROCESSED" legalSystem="AC"
    id="fra.principe_de_précaution.123425.e">
    <term>principe de précaution</term>
    <grammar>n.m.</grammar>
    <domain>environnement</domain>
    <usage frequency="UNKNOWN" geographical-code="" technical="false"/>
    <definition>...</definition>
    ...
    <note></note>
    <admin>
      <created-by>...</created-by>
      <creation-date>...</creation-date>
      ...
    </admin>
  </entry>
</volume>
What is the necessary work for this?

Providing dictionary access

1. Decide on the structure of your lexical database
   - Monolingual, bilingual, multi-bilingual, interlingual
   - Here, we chose an interlingual approach with 5 monolingual dictionaries and 1 pivot

2. Then, decide on the structure of your dictionary
   - Your only constraint is: use an XML structure

3. Identify your pieces of informations
   - by associating elements in your structures with elements of a simple structure...
   - ...called the Common Dictionary Markup structure
   - association is done with simple XPaths
What is the necessary work for this?

Providing dictionary access

```xml
<volume volume-language="fra">
  <entry lang="fra" status="UNKNOWN"
         process_status="UNPROCESSED" legalSystem="AC"
         id="fra.principe_de_précaution.123425.e">
    <term>principe de précaution</term>
    <grammar>n.m.</grammar>
    <domain>environnement</domain>
    <usage frequency="UNKNOWN" geographical-code="" technical="false"/>
    <definition>...</definition>
    ...
    <note></note>
    <admin>
      <created-by>...</created-by>
      <creation-date>...</creation-date>
    </admin>
  </entry>
</volume>
```

```plaintext
cdm-headword = /volume/entry/term/text()
cdm-pos = /volume/entry/pos/text()
cdm-id = /volume/entry@id
```
What is the necessary work for this?

Providing dictionary access

1. Decide on the structure of your lexical database
   - Monolingual, bilingual, multi-bilingual, interlingual
   - Here, we chose an interlingual approach
   - with 5 monolingual dictionaries and 1 pivot

2. Then, decide on the structure of your dictionary
   - Your only constraint is: use an XML structure

3. Identify your pieces of informations
   - by associating elements in your structures with elements of a simple structure...
   - ...called the Common Dictionary Markup structure
   - association is done with simple XPaths

4. Provide a presentation form for your users
   - by way of a simple XML stylesheet (producing xhtml from the xml structure)
What is the necessary work for this?

Providing dictionary Editing

1. Define the XML schema of your structure
What is the necessary work for this?

Providing dictionary Editing

1. Define the XML schema of your structure
2. Provide an empty structure bearing default values
What is the necessary work for this?

Providing dictionary Editing

1. Define the XML schema of your structure
2. Provide an empty structure bearing default values
3. Let the system generate a default interface
What is the necessary work for this?
Providing dictionary Editing

1. Define the XML schema of your structure
2. Provide an empty structure bearing default values
3. Let the system generate a default interface
4. Customise it to fit your user needs
   - you can hide the information you don’t want your contributors to see
Part IV

A Step by Step tutorial on jibiki
Part V

An Autopsy of Existing Dictionaries Build with Jibiki
Making Structures Inter-operate