WP7
Case Study: Patents

Cristina España-Bonet
Lluís Màrquez
Universitat Politècnica de Catalunya, TALP Research Center

– 1st year review –

Luxembourg, March 15th, 2011

Overview

1. WP general view
2. Ongoing work
3. Future work
4. Dissemination

WP general view

Goal

Development of a prototype for translation and retrieval of patents. Test bed for hybrid translation.

WP general view

Goal

Development of a prototype for translation and retrieval of patents. Test bed for hybrid translation.

Especially related to:

WP2 Grammar-based translation method

WP4 Semantic infrastructure for retrieval

WP5 SMT and Hybrid translation systems

WP9 Evaluation

WP general view

Participants & PMs & Tasks

UPC	**15**	Corpus building, hybrid translation, evaluation

WP general view

Participants & PMs & Tasks

UPC	**15**	Corpus building, hybrid translation, evaluation
Ontotext	**15**	Semantic infrastructure, prototype building

WP general view

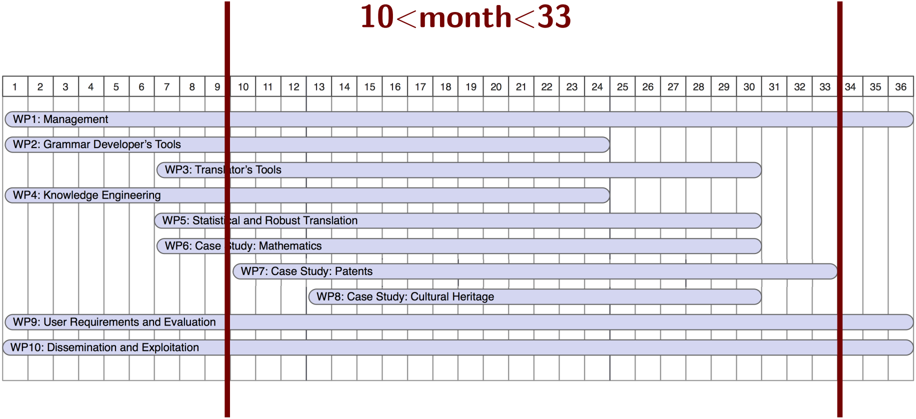
Participants & PMs & Tasks

UPC	**15**	Corpus building, hybrid translation, evaluation
Ontotext	**15**	Semantic infrastructure, prototype building
GOT	**12**	Domain Grammar

WP general view

Timeline

10<month<33



WP1: Management

WP2: Grammar Developer's Tools

WP3: Translator's Tools

WP4: Knowledge Engineering

WP5: Statistical and Robust Translation

WP6: Case Study: Mathematics

WP7: Case Study: Patents

WP8: Case Study: Cultural Heritage

WP9: User Requirements and Evaluation

WP10: Dissemination and Exploitation

WP general view

Milestones & Deliverables

Month 21 — **Month 27** — **Month 33**

D71

Patent MT and retrieval prototype beta.

WP general view

Milestones & Deliverables

Month 21 — **Month 27** — Month 33

D71

Patent MT and retrieval prototype beta.

D72

Patent MT and retrieval prototype.

WP general view

Milestones & Deliverables

Month 21 — Month 27 — Month 33

MS8
Case study complete.

D73
Patent case study final report.

Ongoing work

Overview

1 WP general view

2 Ongoing work
- Corpus
- Compounds tokenizer

3 Future work

4 Dissemination

Ongoing work

Status

WP delayed due to the leave of Matrixware and the search of a **new data provider**.

Meanwhile...
work has started with the patent data given for the **CLEF-IP track** in the CLEF 1010 Conference.

CLEF-IP 2010 Collection

Extract of the MAREC dataset, containing over 2.6 million patent documents pertaining to 1.3 milion patents from the EPO with some content in English, German and French.

Ongoing work

Parallel corpus selection

- Patent documents with **translated claims**.
  (not all of them!)

- IPC classification **A61P**.
  Specific therapeutic activity of chemical compounds or medical preparations.

Ongoing work

Parallel corpus selection

- Patent documents with **translated claims**.
  (not all of them!)
- IPC classification **A61P**.
  Specific therapeutic activity of chemical compounds or medical preparations.

56000 patents out of 1.3 million fulfill these demands.
(279282 aligned parallel fragments)

Ongoing work

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- The use according to claim 7, wherein said cancer diseases comprise bladder, lung, mamma, melanoma and prostate carcinomas.

- A compound according to claim 1 wherein it is (2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide.

- The pharmaceutical composition according to claim 1 or 2, wherein said platinum anticancer agent is selected from at least one of the complexes having structures of: **IMAGE**.

Ongoing work

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- **The use according to claim 7, wherein** `said cancer diseases comprise bladder, lung, mamma, melanoma and prostate carcinomas.`

- **A compound according to claim 1 wherein** `it is (2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide.`

- `The pharmaceutical` **composition according to claim 1 or 2, wherein said** `platinum anticancer agent is selected from at least one of the complexes having structures of: **IMAGE**.`

Ongoing work

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- The use according to claim 7, wherein said cancer diseases comprise **bladder, lung, mamma, melanoma and prostate carcinomas**.

- A compound according to claim 1 wherein it is (2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide.

- The pharmaceutical composition according to claim 1 or 2, wherein said **platinum anticancer agent** is selected from at least one of the complexes having structures of: **IMAGE**.

Ongoing work

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- The use according to claim 7, wherein said cancer diseases comprise bladder, lung, mamma, melanoma and prostate carcinomas.
- A compound according to claim 1 wherein it is **(2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide**.
- The pharmaceutical composition according to claim 1 or 2, wherein said platinum anticancer agent is selected from at least one of the complexes having structures of: **IMAGE**.

Ongoing work

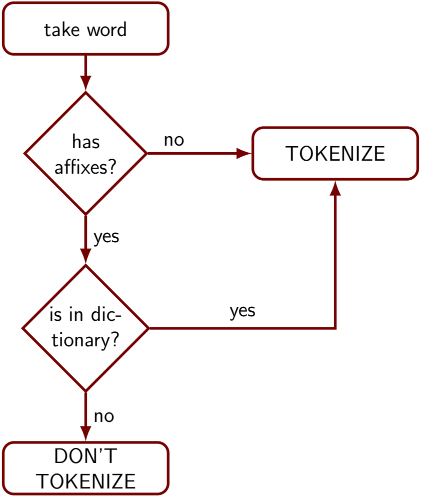
Chemical domain issues

The main issue is the **treatment of chemical compounds**.

- **Compound detector**
  Based on affix detection.
- **Compound tokenizer**
  Based on the detector and a regular tokenizer.
- **Compound translator**
  Two separate approaches: SMT and GF.

Ongoing work

Compound tokenizer (non-tokenizer!)



Ongoing work

Compound tokenizer (non-tokenizer!)

Elements that appear in the **list of affixes**

Prefixes `Meth-, Eth-, Prop-, Pentadec-, imido-, selenocarboxy-, hydroxy-, Propion-, Arachid-...`

Sufixes `-ol, -one, -al, -aldehyde, -oic, -oate, -oxy, -sulfonic, -nitrile, -amine, -isocyanide...`

(English & German: 142 elements, French: 148 elements)

Ongoing work

Compound tokenizer (non-tokenizer!)

Elements that appear in the **list of affixes**

Prefixes `Meth-`, `Eth-`, `Prop-`, `Pentadec-`, `imido-`, `selenocarboxy-`, `hydroxy-`, `Propion-`, `Arachid-...`

Sufixes `-ol`, `-one`, `-al`, `-aldehyde`, `-oic`, `-oate`, `-oxy`, `-sulfonic`, `-nitrile`, `-amine`, `-isocyanide...`

(English & German: 142 elements, French: 148 elements)

Need to check against a **dictionary** (English).

Ongoing work

Compound detection from the tokenizer

The method works better as a tokenizer than as a compound detector, it beds for **high recall** instead of precision.

Actual missclassifications:

- Proper names: `Hôpit`**al**
- Words which are not in the dictionary: `Extracorpore`**al**
- Groups: `-internation`**al**
- Typos: `compar`**oate**

Ongoing work

Compound detection from the tokenizer

The method works better as a tokenizer than as a compound detector, it beds for **high recall** instead of precision.

Actual missclassifications:

- Proper names: `Hôpit`**al**
- Words which are not in the dictionary: `Extracorpore`**al**
- Groups: `-internation`**al**
- Typos: `compar`**oate**

103,272 (compounds + noise)

Ongoing work

Corpus

Provisional **tokenized** parallel corpus in the chemical domain

SET	Segments	EN tok	DE tok	FR tok
Training	279,282	7,954,491	7,346,319	8,906,379
Development	993	29,253	26,796	33,825
Test	1,008	31,239	28,225	35,263

IPC A61P

Future work

Overview

1. WP general view
2. Ongoing work
3. Future work
4. Dissemination

Future work

Planning

Near future work

- Related to the **corpus**
- Related to the **domain grammar**
- Related to the **knowledge infrastructure**

Further work

- **Prototype** building

Future work

Related to the corpus

- Modify corpus according to the provided data
- Prepare it for the interaction with WP5, more cleaning needed
- Automatic detection and extraction of compounds

Future work

Related to the domain grammar & knowledge infrastructure

Domain grammar

- Creation of a modular GF grammar for patents
- Compounds module & General structures module

Knowledge infrastructure

- Semantic representation for patents

Dissemination

Overview

1. WP general view
2. Ongoing work
3. Future work
4. Dissemination

MOLTO

Dissemination

WP dissemination

Recently start WP

No related publications yet

Little research within the WP.
Few publications expected
(Resources at LREC)

WP7
Case Study: Patents

Cristina España-Bonet
Lluís Màrquez

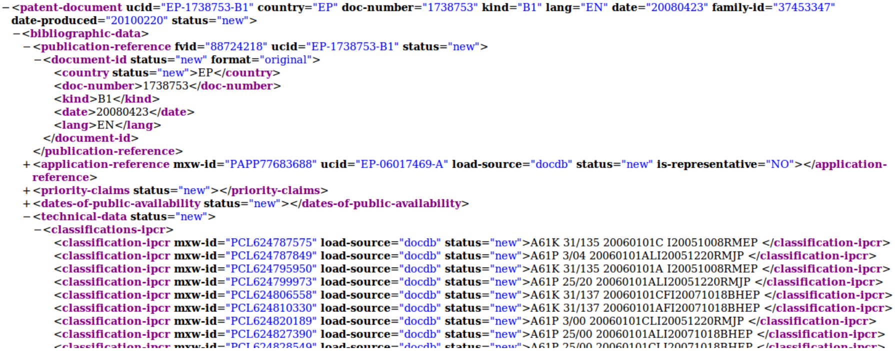
Universitat Politècnica de Catalunya, TALP Research Center

– 1st year review –

Luxembourg, March 15th, 2011

A Patent document

Patent document, **IPC** classification.

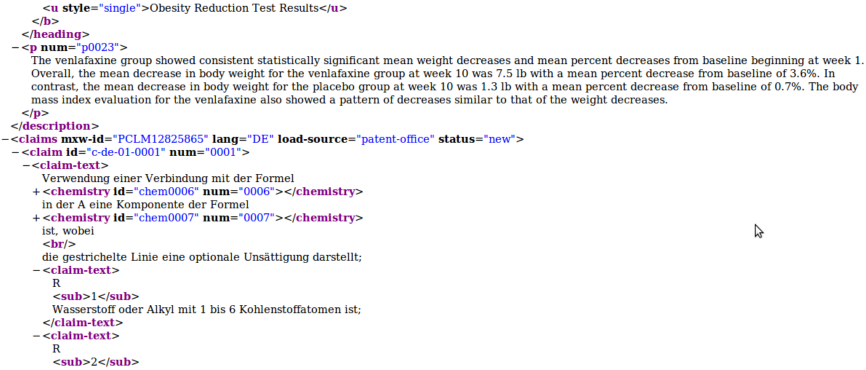


```
–<patent-document ucid="EP-1738753-B1" country="EP" doc-number="1738753" kind="B1" lang="EN" date="20080423" family-id="37453347"
 date-produced="20100220" status="new">
  –<bibliographic-data>
    –<publication-reference fvid="88724218" ucid="EP-1738753-B1" status="new">
      –<document-id status="new" format="original">
          <country status="new">EP</country>
          <doc-number>1738753</doc-number>
          <kind>B1</kind>
          <date>20080423</date>
          <lang>EN</lang>
        </document-id>
      </publication-reference>
    +<application-reference mxw-id="PAPP77683688" ucid="EP-06017469-A" load-source="docdb" status="new" is-representative="NO"></application-
      reference>
    +<priority-claims status="new"></priority-claims>
    +<dates-of-public-availability status="new"></dates-of-public-availability>
    –<technical-data status="new">
      –<classifications-ipcr>
          <classification-ipcr mxw-id="PCL624787575" load-source="docdb" status="new">A61K 31/135 20060101C I20051008RMEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624787849" load-source="docdb" status="new">A61P 3/04 20060101ALI20051220RMJP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624795950" load-source="docdb" status="new">A61K 31/135 20060101A I20051008RMEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624799973" load-source="docdb" status="new">A61P 25/20 20060101ALI20051220RMJP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624806558" load-source="docdb" status="new">A61K 31/137 20060101CFI20071018BHEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624810330" load-source="docdb" status="new">A61K 31/137 20060101AFI20071018BHEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624820189" load-source="docdb" status="new">A61P 3/00 20060101CLI20051220RMJP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624827390" load-source="docdb" status="new">A61P 25/00 20060101ALI20071018BHEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624828549" load-source="docdb" status="new">A61P 25/00 20060101CLI20071018BHEP </classification-ipcr>
```

Extra slides

A Patent document

Description, **claims**.



```
        <u style="single">Obesity Reduction Test Results</u>
      </b>
    </heading>
  – <p num="p0023">
      The venlafaxine group showed consistent statistically significant mean weight decreases and mean percent decreases from baseline beginning at week 1.
      Overall, the mean decrease in body weight for the venlafaxine group at week 10 was 7.5 lb with a mean percent decrease from baseline of 3.6%. In
      contrast, the mean decrease in body weight for the placebo group at week 10 was 1.3 lb with a mean percent decrease from baseline of 0.7%. The body
      mass index evaluation for the venlafaxine also showed a pattern of decreases similar to that of the weight decreases.
    </p>
  </description>
– <claims mxw-id="PCLM12825865" lang="DE" load-source="patent-office" status="new">
  – <claim id="c-de-01-0001" num="0001">
    – <claim-text>
        Verwendung einer Verbindung mit der Formel
      + <chemistry id="chem0006" num="0006"></chemistry>
        in der A eine Komponente der Formel
      + <chemistry id="chem0007" num="0007"></chemistry>
        ist, wobei
        <br/>
        die gestrichelte Linie eine optionale Unsättigung darstellt;
      – <claim-text>
          R
          <sub>1</sub>
          Wasserstoff oder Alkyl mit 1 bis 6 Kohlenstoffatomen ist;
        </claim-text>
      – <claim-text>
          R
          <sub>2</sub>
```

Extra slides

Language domain and genre, other characteristics

Claims have also **long sentences** and **missing information**.

Excerpt 2

- Use of compounds of formula I **IMAGE** wherein R1 signifies substituted C1-C4-alkylene, whereby the substituents are selected from the group comprising unsubstituted aryloxy or aryloxy mono- to penta-substituted by R5, and unsubstituted pyridyloxy or pyridyloxy mono- to tetra-substituted by R5, whereby the substituents may be the same as one another or different if the number thereof is greater than 1; R2 signifies unsubstituted phenyl or phenyl mono- to penta-substituted by R5, or unsubstituted pyridyl or pyridyl mono- to tetra-substituted by R5; R3 is methyl; R4 signifies hydrogen, C1-C6-alkyl or halogen-C1-C6-alkyl; R5 signifies C1-C6-alkyl, C1-C6-alkoxy, halogen-C1-C6-alkyl, halogen-C1-C6-alkoxy, C2-C6-alkenyl, halogen-C2-C6-alkenyl, C2-C6-alkinyl, halogen-C2-C6-alkinyl, C3-C8-cycloalkyl, C1-C6-alkylcarbonyl, halogen-C1-C6-alkylcarbonyl, C1-C6-alkoxycarbonyl, halogen-C1-C6-alkoxycarbonyl, C1-C6-alkylsulfonyl, C1-C6-alkylsulfinyl, halogen, cyano or nitro; A signifies C(R6)(R7), CH=CH or C=C; R6 and R7 either, i ndependently of one another, signify hydrogen, halogen, C1-C6-alkyl, C1-C6-alkoxy, halogen-C1-C6-alkyl, halogen-C1-C6-alkoxy or C3-C6-cycloalkyl; or together signify C2-C6-alkylene; R8 and R9 are hydogen; m and n, independently...of one other, are 0 or 1; and optionally enantiomers thereof, with the proviso that if m is 0 then R1 is retained; in the preparation of a pharmaceutical composition for the control of endoparasitic helminths in warm-blooded productive livestock and domestic animals.

Extra slides

Compound tokenizer demands

Regular tokenizer

8-difluoro-2- [3-fluoro-4 - [(L-lysyl) amino] phenyl]
-7-methyl-4H-1-benzopyran-4-one

- Parenthesis and square brackets are separated.
- Punctuation is separated.

Desired tokenizer

8-difluoro-2-[3-fluoro-4-[(L-lysyl)amino]phenyl]-7-methyl-4H-1-benzopyran-4-one