

The ePatent project

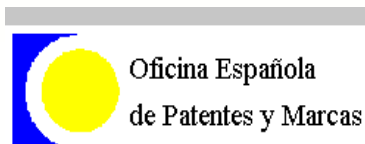
Multilingual access To European Patent databases

Bernard Normier
Lingway
Madrid, May 2002



ePatent framework

- EC eContent program
 - Budget 2.4 M€ started Jan 02. 2 years.
- Partners
 - INPI, France
 - UKPO, UK
 - OEPM, Spain
 - OEPA, Austria
 - Lingway, France (natural language technology)
 - Jouve, France (database and internet infrastructure)



Using NLP to facilitate use of patent database

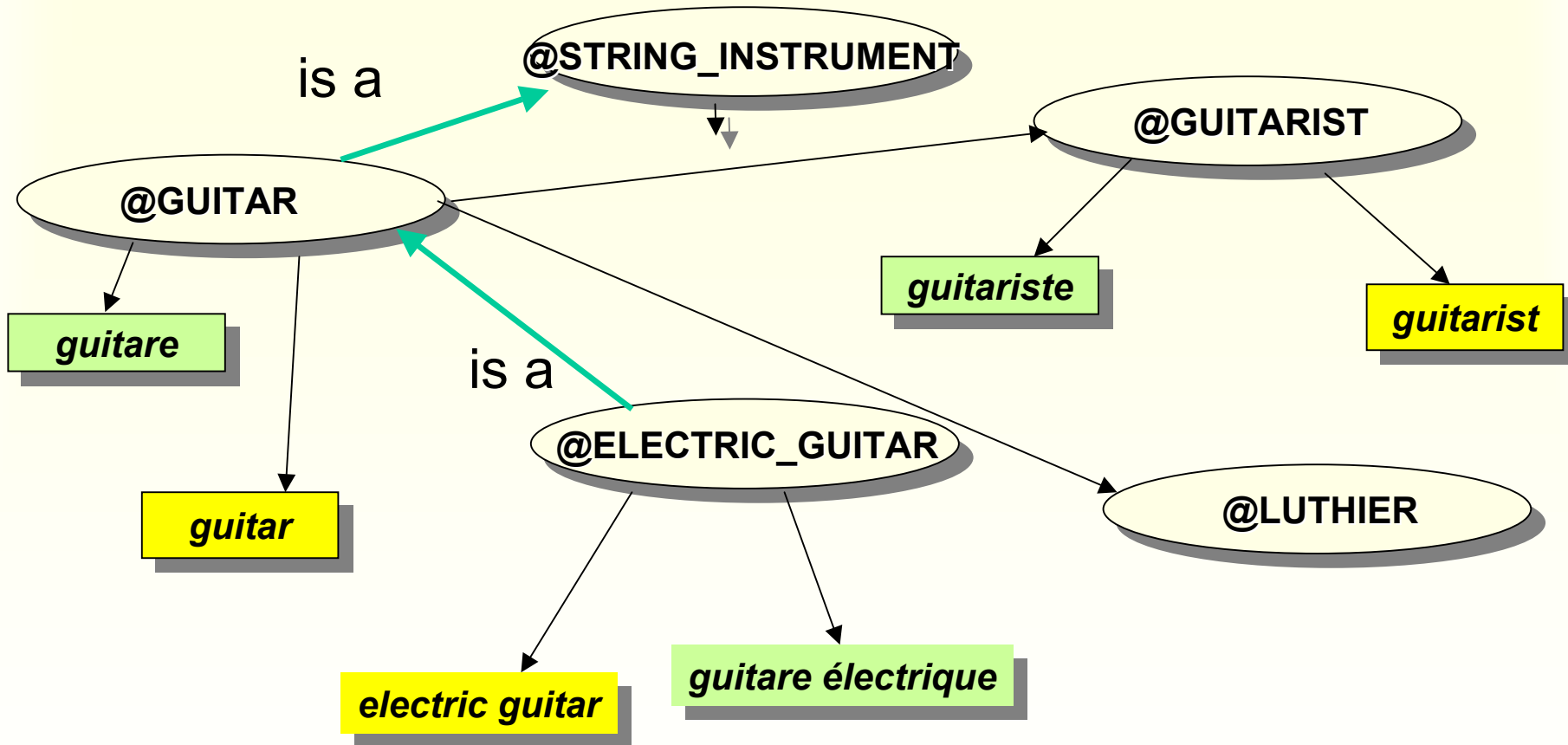
- Patent databases are difficult to use by non-professional users
 - Search using IPC is difficult
 - Search using boolean queries is difficult
 - Multilinguality is a key issue in Europe
- ePatent will offer
 - Easy access for non professional users (SMEs)
 - Search, filtering, reading aids and translation tools



Key technology: natural language processing

- What is NLP
 - Tools to «understand» natural languages
 - Based on dictionaries, semantic networks, grammars
 - Now a mature technology
- Applications
 - Natural language interfaces, mono or multi-lingual
 - Text analysis: indexing, extracting, tagging
 - Translation, multilingual reading aids

A multilingual semantic network



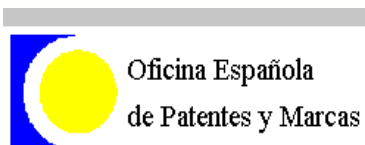
Re-using the French experience of INPI

- Using NL interface on IPC for 5 years
 - Translates NL query to IPC codes
 - Based on a 50.000 words dictionary, 35.000 concepts
 - 30.000 NL queries (in French only) per month
- Extension to 3 new languages
 - English, Spanish and German
- Extension to new features
 - Filtering, Ranking, Translation aids



NL access using IPC to patents

- The language of patents is very specific
 - Distance between the professional language of IP and the general technical and scientific language
- Propose IPC codes through a matching with IPC textual descriptions (Titles, cross-references, key phrases, Notes, Outlines)
- Patents are obtained through the IPC codes



Example IPC NLP access (I)

The screenshot shows a search interface with a search bar containing "appareil à peler des poires" and an "OK" button. Below the search bar, there is a text prompt: "Un nombre important de brevets peut répondre à cette question. Pour sélectionner dans ce(s) domaine(s) les aspects de la technique qui vous intéressent". The search results are displayed under the heading "A-23 : ALIMENTS OU PRODUITS..." and "2 résultats". Two results are listed:

- A-23-N-4/00 - Machines à dépeeler les pépins, caractérisées par...
- A-23-N-4/01 - Machines à dépeeler les légumes pour les légumes bulbes à fleurs;

Three yellow callouts provide additional context:

- A callout pointing to the search bar contains the text: "device to peel peers".
- A callout pointing to the search results heading contains the text: "Foods or foodstuff; their treatment, not covered by other classes".
- A callout pointing to the second search result contains the text: "Machines or apparatus for other treatments of fruits or vegetables for human purposes; machines or apparatus for topping or skimming flowers bulbs".

Example IPC NLP access (II)

In French peler ↔ éplucher

PLUTARQUE

appareil à éplucher des poires

Un nombre important de brevets peut répondre à cette question. Pour préciser votre recherche, sélectionnez un ou plusieurs domaine(s) les aspects de la technique qui vous intéressent.

B-05 : PULVÉRISATION OU ATOMISATION; LEURS APPAREILS 1 résultat

A-23 : ALIMENTS OU PRODUITS ALIMENTAIRES; LEUR TRAITEMENT 1 résultat

A-23-N- Machines à dénoyauter les fruits ou à extraire des fruits les parties contenant les pépins, les pépins, les ossements, les parties dures par leur dispositif d'extraction;

device to skin peers

Foods or foodstuff; their treatment,
not covered by other classes

Example IPC NLP access (III)

poire ↔ fruit

appareil à peler des fruits

OK

Un nombre important de brevets peut répondre à cette question. Pour préciser votre question, précisez le domaine(s) les aspects de la technique qui vous intéressent.

A-23 : ALIMENTS OU PRODUITS...

12 résultats

B-41 : IMPRIMERIE; LIGNARDS...

1 résultat

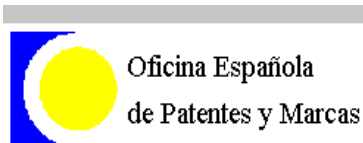
device to peel fruits

A-23 : ALIMENTS OU PRODUITS ALIMENTAIRES; LEUR TRAITEMENT, NON

- A-23-N-7/00** - Epluchage ou pelage des légumes ou des fruits;
- A-23-N-12/00** - Machines à nettoyer, blanchir, sécher, griller ou torréfier les fruits ou les légumes, p. ex. le café, le cacao, les noix;
- A-23-N** - MACHINES OU APPAREILS POUR TRAITER LES RÉCOLTES DE FRUITS, DE LÉGUMES OU DE BULBES A FLEURS EN GRANDES QUANTITÉS, NON PRÉVUS AILLEURS; ÉPLUCHAGE DE LÉGUMES OU PELAGE DE FRUITS EN GRANDES QUANTITÉS; APPAREILS POUR LA PRÉPARATION DE LA NOURRITURE POUR ANIMAUX;
- A-23-N-17/00** - Appareils spécialement adaptés pour la préparation de la nourriture des animaux;
- A-23-N-15/00** - Machines ou appareils destinés à d'autres traitements des fruits ou des légumes pour les besoins de l'homme; Machines ou appareils pour écimer ou peler les bulbes à fleurs;
- A-23-N-11/00** - Séparation de la sève des tiges;
- A-23-N-5/00** - Machines à écaler, cerner ou casser les noix;
- A-23-N-4/00** - Machines à dénoyauter les fruits ou à extraire des fruits les parties contenant les pépins, caractérisées par leur dispositif d'extraction;

Classical full text access to patents

- Key words access
- Search for patents, through a boolean query and the full text engine ranking mechanism, by matching Titles and Abstracts
- No benefit of the IPC expertise
- Classical problems of full text engines : noise and silence

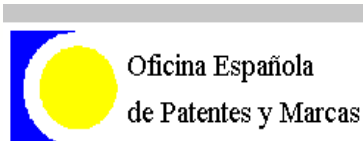


Enhancing full text with NLP

- Analyse a natural language query
- Use the semantic network to expand to related words
- Generate the appropriate boolean query to the text engine
- Reduce by a significant factor both noise and silence
- Allows a better highlighting of relevant words in text

Filtering approach

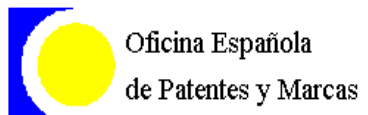
- Analyse the document structure and “rhetorical” markers in the text
 - Done by an automatic XML tagging with “semantic tags”
- In batch on large volumes (XML retro-conversion)
- In real time on small sets (result of a previous query)
- Rank relevant documents according to the query, and highlights relevant text fragments



Exemple

```
brevet.doc - Microsoft Word
File Edit View Insert Format Tools Table Window Help

- <brevet>
  <doc Number="000526">### DoCuMeNtNuMbEr ###
  000526</doc>
  - <section type="InventionField">
    <title>TECHNICAL FIELD</title>
    - <paragraph>
      <sentence type="InventionField" value="an apparatus for
      guiding a vehicle along a predetermined path"
      confidence="1.0">The present invention relates to an
      apparatus for guiding a vehicle along a predetermined
      path.</sentence>
      <sentence type="InventionField" value="an apparatus for
      mounting on an implement which is being towed
      behind a tractor or the like" confidence="1.0">More
      particularly, the present invention relates to an
      apparatus for mounting on an implement which is
      being towed behind a tractor or the like.</sentence>
      <sentence type="InventionProperties" value="has a wheel
      mechanism for following a precut trench and a trench
      mechanism for simultaneously forming a trench for
      the wheel mechanism to follow on a subsequent path
      by the tractor/implement" confidence="0.7">The
      guiding apparatus has a wheel mechanism for
      following a precut trench and a trench mechanism for
      simultaneously forming a trench for the wheel
      mechanism to follow on a subsequent path by the
      tractor/implement.</sentence>
    </paragraph>
  </section>
  - <section type="InventionBackground">
    <title>BACKGROUND OF THE INVENTION</title>
    - <paragraph>
      <sentence type="PreviousStateOfArt" value="Systems for
      remotely guiding vehicles by operator control or in
      accordance with an automatic guide mechanism
      which follows a pre-established path"
      confidence="1.0">Systems for remotely guiding
      vehicles by operator control or in accordance with an
      automatic guide mechanism which follows a pre-
      established path have been long known in the
      art.</sentence>
    </paragraph>
  </section>
</brevet>
```

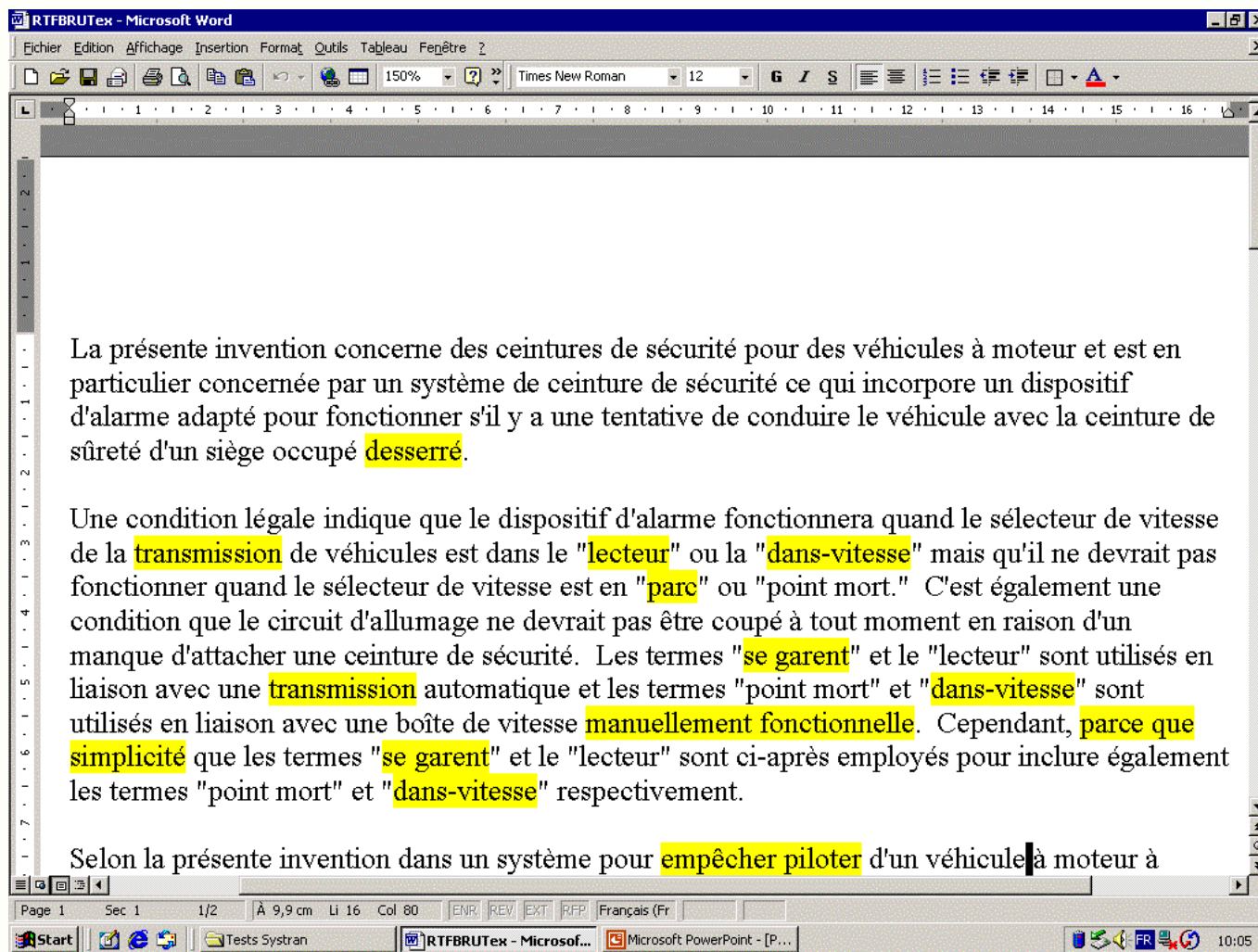


Reading aids and translation

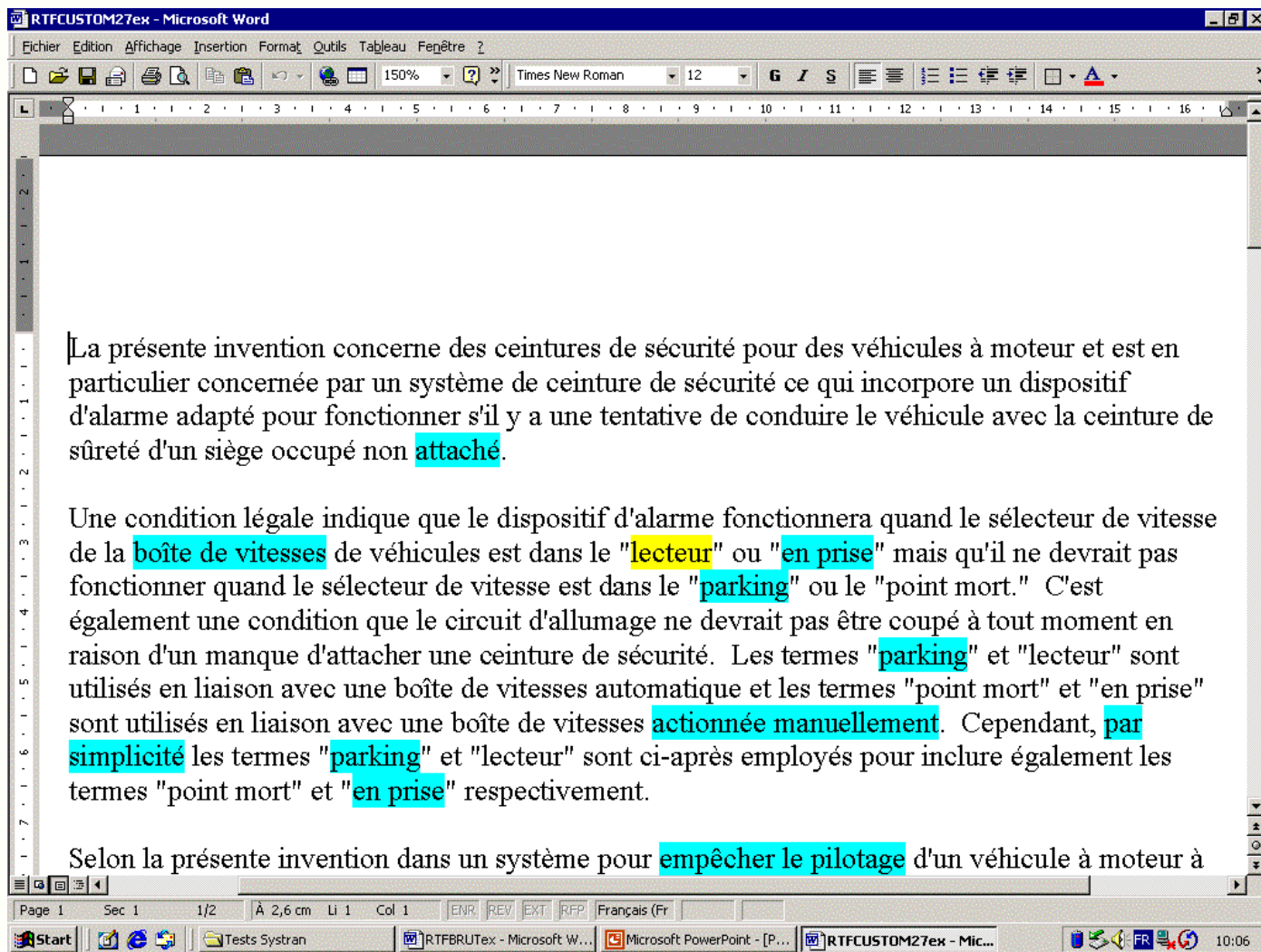
- Help the user with foreign language
- On line dictionaries
- Term identification and translation
- “draft” translation
 - (not yet decided in ePatent)
 - Enrich existing MT systems with Patent terminology



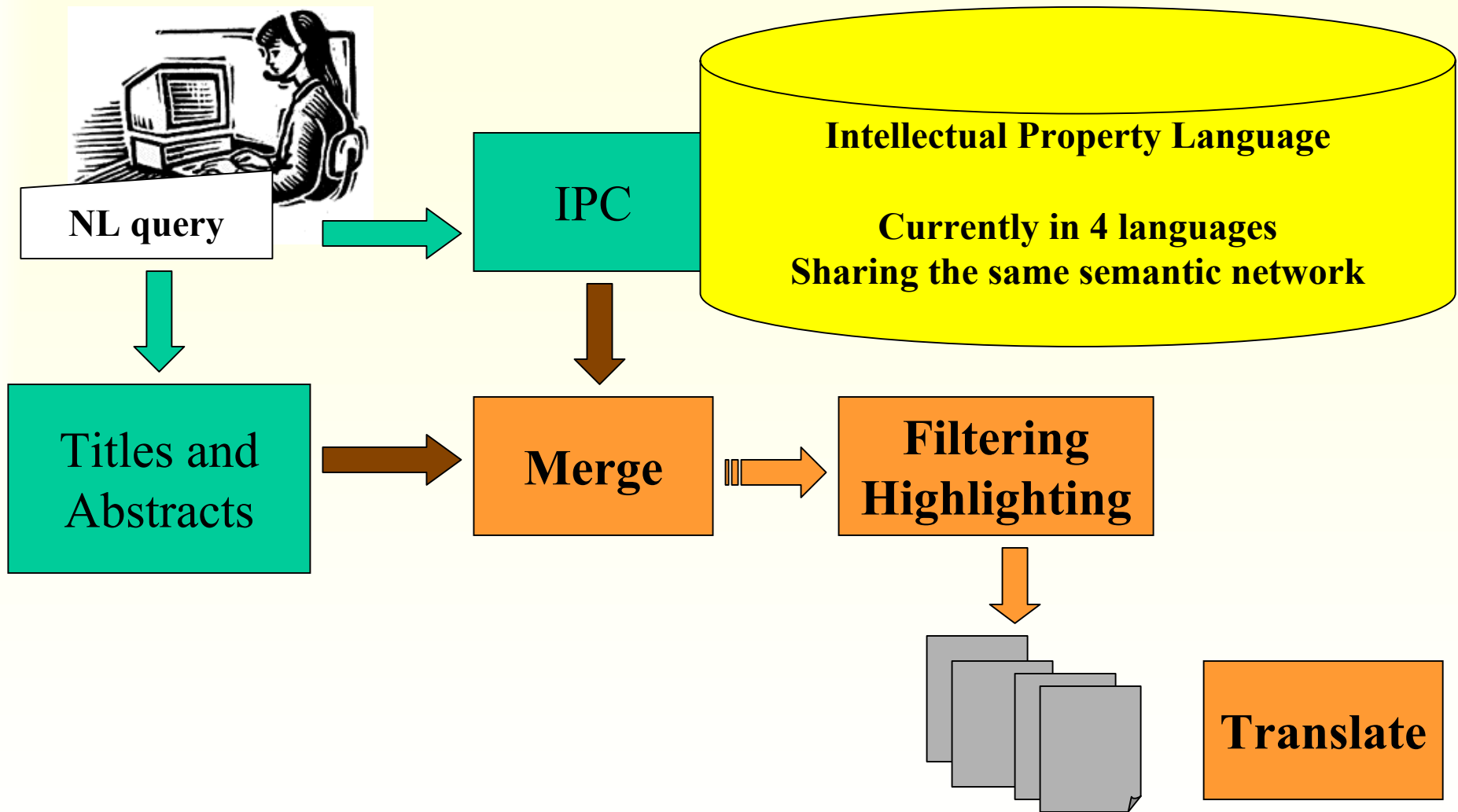
Machine translation before terminology adaptation



Machine translation after terminology adaptation



Shared multilingual dictionaries



Conclusion

- NLP is a mature technology
 - Performance and quality are acceptable
 - Dictionaries can be shared between several tasks
 - Semantic networks can be reused between languages
- NLP is a key technology for patent databases
 - To give an easy access to non-specialist
 - To develop high value added services
 - To target new market niches, in particular in SMEs

