3rd Cameleon Workshop, UFRGS

Written corpora for human beings

Achille Falaise – **LIG-GETALP** & LIDILEM

Motivations

For lexical studies, language learning, translation, etc.

- Dictionaries and grammars are not enough... we need context !

Motivations

For lexical studies, language learning, translation, etc.

- Dictionaries and grammars are not enough... we need context !
  - Textual context
    - “ banks ."

Motivations

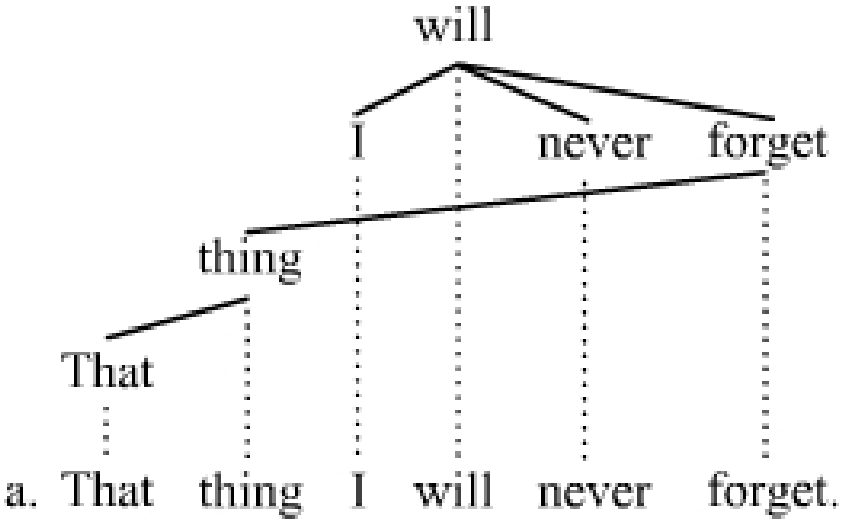
For lexical studies, language learning, translation, etc.

- Dictionaries and grammars are not enough... we need context !
  - Textual context
    - “In this picture, you can see a lot of large rocks on the banks of the river."

Motivations

For lexical studies, language learning, translation, etc.

- Dictionaries and grammars are not enough... we need context !
  - Textual context
  - Syntactic context



Motivations

For lexical studies, language learning, translation, etc.

- Dictionaries and grammars are not enough... we need context !
  - Textual context
  - Syntactic context
  - Structural context

Motivations

For lexical studies, language learning, translation, etc.

- Dictionaries and grammars are not enough... we need context !
  - Textual context
  - Syntactic context
  - Structural context
  - Thematic context
    - Domain, topic, document type, etc.

Consequence : lots of information to store

We keep all these informations

- Text

The role of apoptotic mimicry in host-parasite interplay: is death the only alternative for altruistic behavior?

Consequence : lots of information to store

We keep all these informations

- Text
- Lemmas and POS

```
<TXT>The role of apoptotic mimicry in host-parasite interplay: is death the only alternative for altruistic behavior?
</TXT>
<tokens>
        <t i="1" l="the" f="The" c="Det" p="D"/>
        <t i="2" l="role" f="role" c="Nom?S" p="N"/>
        <t i="3" l="of" f="of" c="Prep" p="O"/>
        <t i="4" l="apoptotic" f="apoptotic" c="Adj" p="A"/>
        <t i="5" l="mimicry" f="mimicry" c="Nom?S" p="N"/>
        [...]
</tokens>
```

Consequence : lots of information to store

We keep all these informations

- Text
- Lemmas and POS
- Syntactic trees

```
<TXT>The role of apoptotic mimicry in host-parasite
interplay: is death the only alternative for altruistic behavior?
</TXT>
<tokens>
        <t i="1" l="the" f="The" c="Det" p="D"/>
        <t i="2" l="role" f="role" c="Nom?S" p="N"/>
        <t i="3" l="of" f="of" c="Prep" p="O"/>
        <t i="4" l="apoptotic" f="apoptotic" c="Adj" p="A"/>
        <t i="5" l="mimicry" f="mimicry" c="Nom?S" p="N"/>
        [...]
</tokens>
<dependances>
        <g r="DET" s="2" c="1"/>
        <g r="PREP" s="2" c="3"/>
        <g r="NOMPREP" s="3" c="5"/>
        <g r="NN" s="5" c="4"/>
        […]
</dependances>
```

Consequence : lots of information to store

We keep all these informations

- Text
- Lemmas and POS
- Syntactic trees
- Structural information

```
<head>
<SEQ id="f-1475-9292-2-6-eti.xml-2666-1">
<TXT>The role of apoptotic mimicry in host-parasite
interplay: is death the only alternative for altruistic behavior?
</TXT>
<tokens>
        <t i="1" l="the" f="The" c="Det" p="D"/>
        <t i="2" l="role" f="role" c="Nom?S" p="N"/>
        <t i="3" l="of" f="of" c="Prep" p="O"/>
        <t i="4" l="apoptotic" f="apoptotic" c="Adj" p="A"/>
        <t i="5" l="mimicry" f="mimicry" c="Nom?S" p="N"/>
        [...]
</tokens>
<dependances>
        <g r="DET" s="2" c="1"/>
        <g r="PREP" s="2" c="3"/>
        <g r="NOMPREP" s="3" c="5"/>
        <g r="NN" s="5" c="4"/>
        […]
</dependances>
</SEQ>
</head>
```

Consequence : lots of information to store

We keep all these informations

- Text
- Lemmas and POS
- Syntactic trees
- Structural information
- Document metadata (document type, topic, author, etc.)

But what about the users ?

Linguists, language learners, translators...

- Not computer scientists
  - Easily disturbed by complex interfaces
  - Do not read manuals

ScienQuest : corpora for linguists and language learners

- Spiral development model
  - Advices from users
  - Development
  - Test with users
- Multiple corpora
  - English scientific texts (15M words)
  - English learners reports (1M words)
  - French scientific texts (5M words)
  - French press (200M words)
  - German press (200M words)
  - Spanish press (200M words)
  - English press (200M words)
  - Russian litterature (500k words)

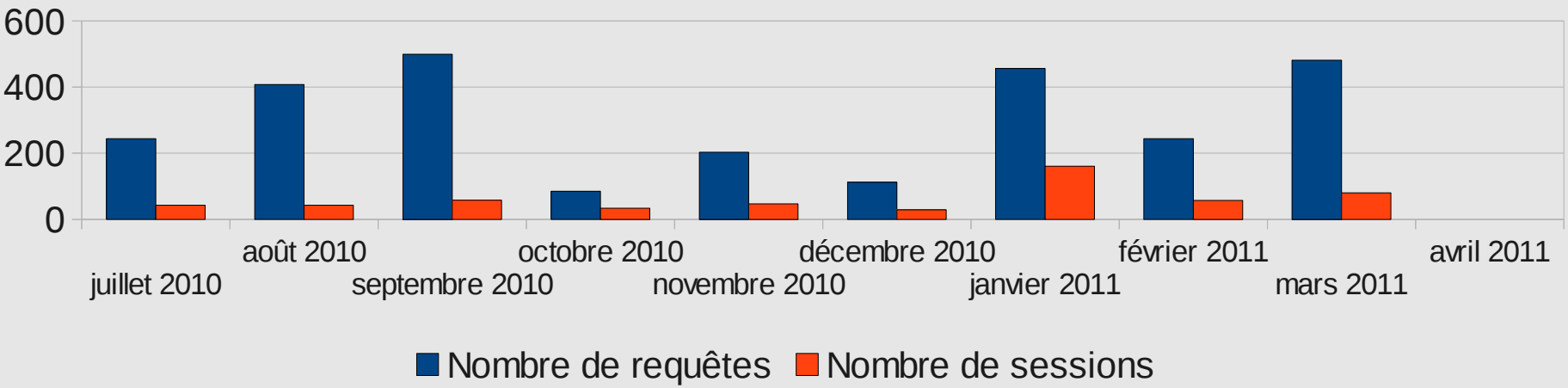
ScienQuest : corpora for linguists and language learners

Demo

http://scientext.msh-alpes.fr

ScienQuest : results

- Since June 2010, 9,021 requests (1,474 sessions)



- Assisted search : 75% (of which 47% with syntax)
- Semantic search : 23%
- Advanced search : 2%

AXiMAG : corpora *by* and for translators

- Interactive Multilingual Access Gateways (website translation)
- 2 aspects :
  - Collaborative (crowdsourcing) aligned copora building
  - Multilingual access to websites

AxiMAG : corpora *by* and for translators

Demo

http://service.aximag.fr/xwiki/bin/view/imag/LICIA-fr-FR

AXiMAG : future

- Real world (& real users) test and evaluation
  - ex. LICIA Website
- Prototype operationalisation
- AXiMAG society

Thanks !

Questions, comments, corpora ?

http://scientext.msh-alpes.fr

http://service.aximag.fr/xwiki/bin/view/imag/LICIA-fr-FR