FastKwic, an “intelligent“ concordancer using FASTR

Veronika Lux-Pogodalla\textsuperscript{12}  Dominique Besagni\textsuperscript{1}  Karën Fort\textsuperscript{1}

\textsuperscript{1}INIST-CNRS, 2 allée de Brabois, 54500 Vandoeuvre-lès-nancy
\textsuperscript{2}ATILF-CNRS, 44 avenue de la Libération, 54000 Nancy

May, 2010
1 Outline

2 Introduction

3 FastKwic

4 Implementation in the TermSciences website

5 Conclusion
**Contexe**

- TermSciences: more than 540,000 terms BUT no definitions.
- Large collection of bibliographical records at INIST.

⇒ Wish for a concordancer.
Elements for the specification

- No complex request language.
- Mono- and multi-word terms ...
- ... occurring in texts with several variations.
Term variations (Jacquemin, Royauté)

Example

- gamma-linolenic acid / gamma linolenic acid / \(\gamma\)-Linolenic acid.
- structural gene / structural erm gene.
- structural gene / structural and regulatory genes.
- resistance mechanism / mechanism of claritomycin resistance.

Major types

- typographical variations (ex. with or without -).
- morphological variations (ex. plural).
- syntactic variations (ex. insertion, coordination, permutation).
Package features

- Two UTF-8 compliant Perl modules.
- Depends on several external freely available tools/resources
  - FASTR [Jacquemin 1997, Jacquemin et al. 1997],
  - TreeTagger [Schmid 1997],
  - Flemm [Namer 2000]
- Freely available on http://www.cnrtl.fr/outil/s/.
Usage
Usage

- **Resource compilation.**
Usage

- Resource compilation.
- Indexing.

- TreeTagger + Flemm (for French)
- POS tagged lemmatized corpus
- compiled terminology (PATR II rules)
- meta-rules for term variation modelling
- FASTR indexing + document transformation
- List of:
  - term$_{i}$
  - variants of term$_{i}$
  - occurrences of term$_{i}$ and its variants in the corpus
  - position in text
Usage

- **Resource compilation.**
- **Indexing.**
- **Production of a concordancer.**

```xml
<?xml version='1.0' encoding='UTF-8'?>
<Concordancer>
  <Term>
    <TotalNumber>2</TotalNumber>
    <Preferential>
      <String>Gene amplification</String>
      <Number>2</Number>
      <Occurrences>
        <Occurrence>
          <Reference>000007</Reference>
          <Position>1:32</Position>
          <Transform>XX,25,Perm</Transform>
          <Context><b>Amplification of the MYC gene is</b> associated with dmi</Context>
        </Occurrence>
        <Occurrence>
          <Reference>000008</Reference>
          <Position>1:38</Position>
          <Transform>XX,15,Ins</Transform>
          <Context><b>This gene facilitated amplification of</b> a 407-bp DNA fragme</Context>
        </Occurrence>
      </Occurrences>
    </Preferential>
  </Term>
</Concordancer>
```
Limitations

- Use of TreeTagger with limited context ⇒ errors in the POS tagging.
- Particular linguistic entities (**1,3,4-thiadiazole(2-amino)**) and their variants not taken into account.
Resources, corpus

- Terminological resource = more than 540,000 terms ⇒
  - FASTR: only for terms with standard linguistic patterns.
  - IRC3(Royauté): for geographical names, drug names, chemical compounds, etc.

- Corpus = 30,744 records for French, 398,952 for English.

- Result of indexing put in MySQL database and accessed from http://www.termosciences.fr/-/Index/Search/Concordancer/
Result example

1. Check terms in the left part.
2. Search for terms and their variants into the PASCAL 2005 concordancer.

Display results for domain:

- All results

Display results for variant:

- All results

8 results of about 27745370

1. de l’intérêt du conte pour le développement affectif et social d’enfants aveugles
2. est son lien intense avec le développement affectif et les apprentissages. Les au
3. du toutus, mais aussi sur le développement psychomoteur et affectif de l’enfant. Cet article dres
4. mique et multicouche d’information du développement social et éducatif
5. alage de quelques repères du développement psychoaffectif de l’enfant, nous établissons
6. r histoire infantile de son développement psychoaffectif et de ses relations familiales
7. istics, spatial cognition and affective development - illustrate this complex pro
8. ackground. Research indicates that affective aspects of development provide a basis for autonomous learn
Achievements and perspectives

- FastKwic freely available for the community
  http://www.cnrtl.fr/outils/.
- Possible improvements:
  - integration of new languages
  - linking FastKwic to a terminology extraction tool (ex. ACABIT)
and you, my lady, are the living image of him!" "THANK YOU, my lady." "The duke will never be dead if you want. I'm not doing anything important." "THANK YOU," said Karen. "You're very kind. I've go sign at the outskirts of the village that said, "THANK YOU," "What are you grinning at?" Mary asked time, sir. That's about half an hour from now." "THANK YOU." Carl changed the time on his watch. "A "Dunno about tea; fuckin good at makin a noise." "THANK YOU for sharing that with us, Gav. I shall r , get out. I'm a busy man." "Goodbye," said Kee. "THANK YOU for talking to me. But don't forget, Mr and a strong wind was blowing. She turned round. "THANK YOU, gentlemen. I will have to talk to them. , "Shall I put this straight into your basket?" "THANK YOU so much!" She would never be able to exp d: "What shall I say? Nobody could foresee this. "THANK YOU for this wonderful night, an emotional W . "Dunno about tea; fuckin good at makin a noise." "THANK YOU for sharing that with us, Gav. I shall r , he has no passport. I am a p ing was exquisitely ornamented with tiny pearls. "THANK YOU, my lord," she said, her eyes shining wi ,ived you would call on us." Joan hid her surprise. "THANK YOU for the letter, Your Grace -- 'twas kind ,en me so much happiness. I buy her jewels to say "THANK YOU"." In May 1972 the Duke became ill. Wh