Concordanciers: Thème et variations
Bénédicte Pincemin, Fabrice Issac, Marc Chanove, Michel Mathieu-Colas

To cite this version:

HAL Id: halshs-00154100
https://halshs.archives-ouvertes.fr/halshs-00154100
Submitted on 21 Apr 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Concordancers: Theme & Variations

B. Pincemin, F. Issac,
M. Chanove, M. Mathieu-Colas

8èmes Journées internationales d’Analyse statistique des Données Textuelles
JADT 2006, Besançon, 19-21 avril 2006

What is a Concordancer? Or what should it be?

1) Generalization
   - Key features – summary from existing KWIC tools
2) Extension
   1. Emphasis on meaningful specificity of concordancers
3) Specialization
   1. Case of use in a distributional semantics approach
      (Classes d’objets theory, Gaston Gross)
What is a (true) Concordancer?

- **Definition** (and parameters)
  - For a given corpus
  - A list of all occurrences of a word (or linguistic item)
  - Vertically aligned (column), «stacked»
  - Surrounded by their left and right contexts (of a given size)
  - And sorted by a relevant criteria
Parameter #1 : Search object

- Word
- Phrase
- List of items (topic,..)
- Stem
- Annotations (lemma, part-of-speech,..)
- Mixed (as a complex regular expression)
  - Example : CQP (Christ, 1994)

Parameter #2 : Context’s size

- A line
  - Visual stack effect : the contexts are vertically aligned and immediately superposed
- Different focus
  - shorter => lexical phrases, syntactic constructs
  - longer => for some semantic considerations
- Centered or not
Parameter #3 : Sorting order

• Not incidental, but really mandatory feature
  – Visual stack effect :
    • Convergences (and their extent : massive convergences)
    • Divergences

• Classical sorting keys
  – Textual linearity (chronologic order)
  – The search expression (if varying)
  – L1, L2… and R1, R2… (words around the search object, on the left and/or on the right)

• Multiple sort
  – In practical, Contextual key = last key

The best of the concordance : visual effects

• Why ? Heuristic guiding for efficient reading
  – convergences and divergences
  – extent (singularity or repetition)

• How ? Stack effect
  – Vertical alignement
  – Sort that groups similar items together
Consequences on the classical definition - towards a new (but tradition grounded) definition

- Parameter #2 (Context’s size) is undesirable
  - Illusory power
  - Fixed (default) and adjusted to
    - page / window size (corresponding itself to a good look span)
    - reasonable size of characters for a comfortable reading
  - Possibility of a horizontal curser (for screen output)

- New ways to enhance and refine grouping and contrasting visual effects: the zones

**Zones: definition**

- The search object is detailed into adjacent zones
- Each zone is qualified by:
  1) A stack column (or not)
  2) A possibly typographical emphasis (bold characters, choice of a colour)
  3) An eventual sort (and which one: alphabetical, textual, canonical…)
### Zones : example of query

<table>
<thead>
<tr>
<th>Left context</th>
<th>shall</th>
<th>- MOT{0,3}</th>
<th>- be .+ed</th>
<th>+ Right context</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No column</td>
<td>No column</td>
<td>column</td>
<td>column</td>
</tr>
<tr>
<td>2</td>
<td>Normal</td>
<td>Normal</td>
<td>Red + Italics</td>
<td>Green + Bold</td>
</tr>
<tr>
<td>3</td>
<td>No sort</td>
<td>No sort</td>
<td>2, Alphabetical</td>
<td>1, Frequency</td>
</tr>
</tbody>
</table>

### Zones : example of output

| … Such declarations shall | be deposited | by the St… |
| … equally authentic , shall | be deposited | in the ar… |
| … Such gratis personnel shall | be employed | in accorda.. |
| … under 18 years of age shall | not | be employed | in night w.. |
| subject to compulsory education shall | not | be employed | in such wo. |
| … nor life imprisonment […] shall | be imposed | for offence. |
| … was committed . Nor shall | a heavier penalty | be imposed | than the on |
| … was committed . Nor shall | a heavier penalty | be imposed | than the on |
| … Sentence of death shall | not | be imposed | for crimes |
Benefits from Zones

- Zones are especially efficient to (visually) group and sort tokens selected by a pattern with contextual conditions and (very) variable realizations
- Compared to the state-of-art:
  - As powerful as every kind of sort in existing KWIC concordancers
  - Allows sorting on distant words, with better control (not only the number of words)
- Multiplied and characterized visual stack effects

A concordancer for distributional semantics

- Context: *Classes d’objets* theory
- Goal: efficient use of corpora in order to build, complete or correct the linguistic description
- Concordancers are already used (and useful) for these tasks, but:
  - Massive outputs
  - Difficulty to focus on contextual dependancies (variability)
Classes d’objets Theory (1/3) : arguments => predicate

- Language (and especially semantics) is described through the predicate – argument dependencies
- Predicates are defined by their argumental pattern, syntactically and semantically:
  - Conduire₁ (hum, hum, loc): *Pat conduit son petit frère à l’école*
  - Conduire₂ (hum, transport): *Pat conduit une décapotable*
  - Conduire₃ (voie, locatif): *Ce sentier conduit à la mer*
- Linguistical vs ontological approach of semantic

Classes d’objets Theory (2/3) : arguments are structured in classes

An argument’s value is taken from a set called *Classe d’objets*

<table>
<thead>
<tr>
<th>PREDICATES</th>
<th>ARGUMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>juste₁</td>
<td>pantalon, veste, ...</td>
</tr>
<tr>
<td></td>
<td>Vêtements</td>
</tr>
<tr>
<td></td>
<td>Classes d’objets</td>
</tr>
<tr>
<td>juste₂</td>
<td>piano, flûte, ...</td>
</tr>
<tr>
<td></td>
<td>Instruments de musique</td>
</tr>
</tbody>
</table>
Classes d’objets Theory (3/3) :
(appropriate) predicates => arguments’ classes

A few appropriate predicates (faisceau de prédicats appropriés) can select all the elements of a class, and only them

Arguments of METTRE:
$qqn \text{ met } X$

Arguments ofÊTRE EN:
$qqn \text{ est en } X$

Arguments of ALLER BIEN À:
$X \text{ va bien à } qqn$

Four ways of exploring a corpus

<table>
<thead>
<tr>
<th>Looking for</th>
<th>Syntactic characterization</th>
<th>Class composition</th>
</tr>
</thead>
</table>
| Building classes of arguments | Given = classe d’objets
Looking for = appropriate predicates | Given = appropriate predicates
Looking for = elements of the classe d’objets |
| predicates | Given = class of predicates
Looking for = classes d’objets as defining arguments | Given = argumental pattern (with classes d’objets)
Looking for = class of predicates |
The KWAC-LLI prototype

- Corpus = Newspaper (Le Monde), morphosyntactically tagged (Cordial)
- Classe d’objets = communication routes (voies de communication, Mathieu-Colas, 1998)
- Goal = to find new appropriate predicates
Specificities of the concordancer

- Synthetic table
  - Plus some results as lists, when more suited
  - Avoids the output overflow: mediates and organizes the results
- Results are ordered according to the linguistic principle (in the \textit{classes d’objets} theory):
  - A relevant predicate can be used with all the elements of the \textit{classe d’objets}
- Visual stack effect

\begin{table}
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline
 & \textit{rue} & \textit{route} & \textit{autoroute} & \textit{avenue} & \textit{impasse} & \textit{allée} & \textit{chemin} & \textit{sentier} \\
\hline
\textit{Freq totale} & 2209 & 3004 & 405 & 231 & 905 & 193 & 3455 & 357 \\
\textit{Freq tab 1} & 2105 & 2905 & 372 & 213 & 884 & 184 & 3360 & 336 \\
\textit{Nb Total} & 487 & 455 & 160 & 115 & 108 & 93 & 397 & 116 \\
\textit{Nb tab 1} & 394 & 373 & 131 & 97 & 90 & 84 & 313 & 96 \\
\textit{Freq corpus} & 7179 & 6691 & 1513 & 1032 & 1395 & 464 & 6112 & 879 \\
\hline
\end{tabular}
\end{table}
Lists (out of table) : predicates found with only one argument

KWAC-LLI : concordance lines with zones (1)
Main ideas

- A concordance is more than a set of contexts, because of its heuristic **visual effects**: vertical alignment and sort order
- **Zones** to develop and refine querying possibilities
- KWAC-LLI for distributional semantics, with a synthetic table