



## Concordanciers : Thème et variations

Bénédicte Pincemin, Fabrice Issac, Marc Chanove, Michel Mathieu-Colas

### ► To cite this version:

Bénédicte Pincemin, Fabrice Issac, Marc Chanove, Michel Mathieu-Colas. Concordanciers : Thème et variations. Viprey, Jean-Marie, et al. 8es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006), Apr 2006, Besançon, France. Presses Universitaires de Franche-Comté, 2, pp.773-784, 2006. <halshs-00154100>

**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

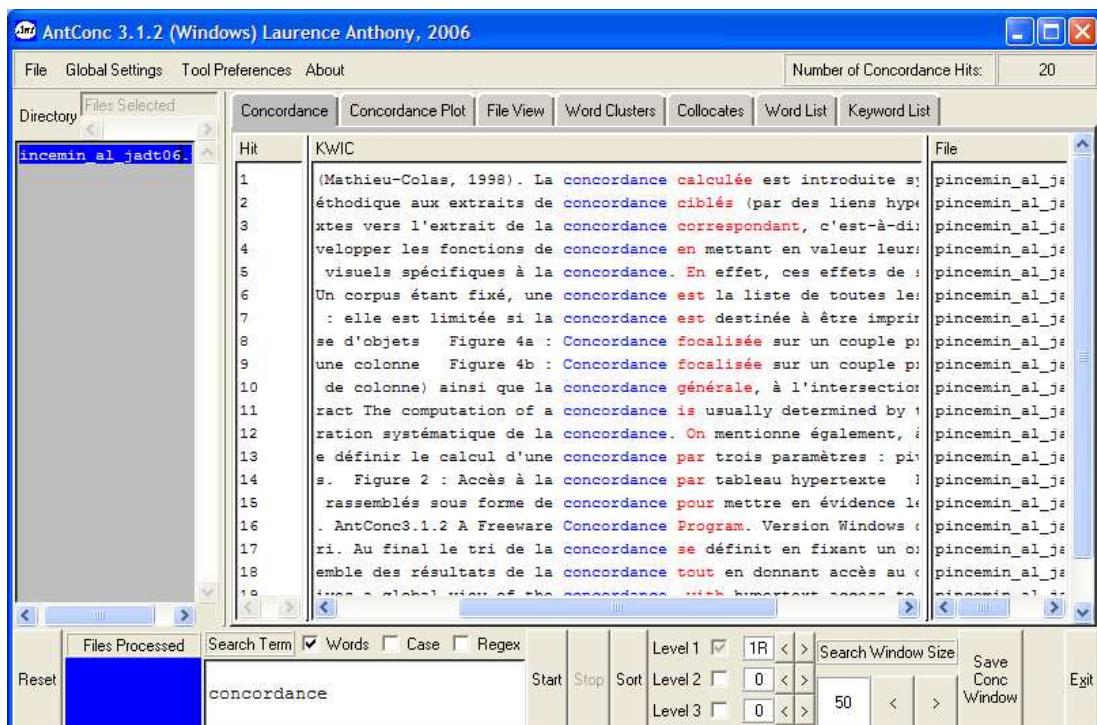
*8èmes Journées internationales d'Analyse statistique des Données Textuelles*

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)

# Example : AntConc



## 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 alignment
  - 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

Left context		shall	- MOT{0,3}	- be .+ed	Right context
1	No column	No column	column	column	No column
2	Normal	Normal	<i>Red + Italic</i>	<b>Green + Bold</b>	Normal
3	No sort	No sort	2, Alphabetical	1, Frequency	3, Alphabetical

# Zones : example of output

... Such declarations shall		<b>be deposited</b>	by the St...
... equally authentic , shall		<b>be deposited</b>	in the ar...
...	...	...	...
... Such gratis personnel shall		<b>be employed</b>	in accorda..
... under 18 years of age shall	<i>not</i>	<b>be employed</b>	in night w..
subject to compulsory education shall	<i>not</i>	<b>be employed</b>	in such wo.
...	...	...	...
... nor life imprisonment [...] shall		<b>be imposed</b>	for offence.
... was committed . Nor shall	<i>a heavier penalty</i>	<b>be imposed</b>	than the on
... was committed . Nor shall	<i>a heavier penalty</i>	<b>be imposed</b>	than the on
... Sentence of death shall	<i>not</i>	<b>be imposed</b>	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 dependancies
- Predicates are defined by their argumental pattern, syntactically **and semantically** :
  - Conduire<sub>1</sub> (hum, hum, loc) : *Pat conduit son petit frère à l'école*
  - Conduire<sub>2</sub> (hum, transport) : *Pat conduit une décapotable*
  - Conduire<sub>3</sub> (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*

PREDICATES

ARGUMENTS

juste<sub>1</sub>

pantalon  
veste  
...

Vêtements

juste<sub>2</sub>

piano  
flûte  
...

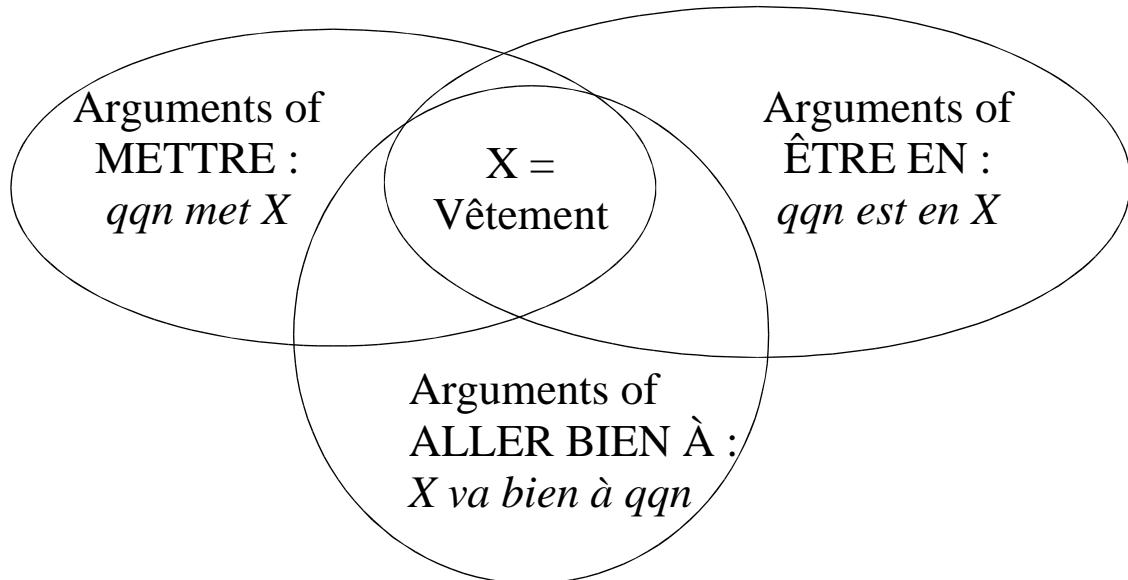
Instruments  
de musique

Classes  
d'objets

# *Classes d'objets* Theory (3/3) :

(appropriate) predicates => arguments' classes

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



## Four ways of exploring a corpus

Looking for →	Syntactic characterization	Class composition
Building classes of ↓		
arguments	<b>Given</b> = <i>classe d'objets</i> <b>Looking for</b> = appropriate predicates	Given = appropriate predicates Looking for = elements of the <i>classe d'objets</i>
predicates	Given = class of predicates Looking for = <i>classes d'objets</i> as defining arguments	<i>Given</i> = argumental pattern (with classes d'objets) <i>Looking for</i> = 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

affichage ConclLI - Mozilla (Build ID: 2002091116-SuSE)

File Edit View Go Bookmarks Tools Window Help

http://.../SuSE-fal

Search

Home Bookmarks The Mozilla Or... SuSE - The Lin...

Requête (indiquer avec le mot clef "ARG" l'emplacement des arguments) :  
<m !="PRED" c="Vm">[">"]<.">["<"]<m ##MOT[1,2]#<m !="ARG">[">"]<.">["<"]<m>

Position des arguments dans la requête : 3

Position des prédictats dans la requête : 1

Arguments (séparés par le caractère "|") : rue|route|autoroute|avenue|impasse|allée|chemin|s

Prédicats (séparés par le caractère "|") : être|avoir|faire|devoir|pouvoir|falloir|vouloir

effacer les prédictats :

oui  non

Seuil de regroupement : 3

Submit Query

	rue	route	autoroute	avenue	impasse	allée	chemin	sentier
Freq totale	2209	3004	405	231	905	193	3455	357
Freq tab 1	2105	2905	372	213	884	184	3360	336
Nb Total	487	455	160	115	108	93	397	116
Nb tab 1	394	373	131	97	90	84	313	96
Freq corpus	7179	6691	1513	1032	1395	464	6112	879
prendre	888	8	5833	34	310	21	4	2
emprunter	346	8	1867	25	92	25	8	1
ouvrir	263	8	4424	33	103	5	5	1
trouver	89	8	1283	5	18	2	2	6
circuler	83	8	731	26	35	10	3	1
éviter	32	8	1282	3	5	1	2	15
aménager	13	8	405	1	4	1	1	1
sortir	696	7	2433	51	10	8	501	2
suivre	430	7	3418	16	91	2	2	6
parcourir	228	7	1519	97	31	5	12	71
aller	195	7	6474	34	33	4	1	6
traverser	176	7	1766	96	44	12	13	5

# 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 *classes d'objets* theory) :
  - A relevant predicate can be used with all the elements of the *classe d'objets*
- Visual stack effect

	rue	route	autoroute	avenue	impasse	allée	chemin	sentier
Freq totale	2209	3004	405	231	905	193	3455	357
Freq tab 1	2105	2905	372	213	884	184	3360	336
Nb Total	487	455	160	115	108	93	397	116
Nb tab 1	394	373	131	97	90	84	313	96
Freq corpus	7179	6691	1513	1032	1395	464	6112	879
prendre	888	8	5833	34	310	21	4	2
emprunter	346	8	1867	25	92	25	8	1
ouvrir	263	8	4424	33	103	5	5	108
trouver	89	8	1283	5	18	2	6	1
circuler	83	8	731	26	35	10	3	1
éviter	32	8	1282	3	5	1	2	4
aménager	13	8	405	1	4	1	1	2
sortir	696	7	2433	51	10	8	501	2
suivre	430	7	3418	16	91	2	2	294
parcourir	228	7	1519	97	31	5	6	12
aller	195	7	6474	34	33	4	1	71
traverser	176	7	1766	96	44	12	13	6

# Lists (out of table) : predicates found with only one argument

A screenshot of a Mozilla browser window titled "affichage ConcLLI - Mozilla {Build ID: 2002091116-SuSE}". The address bar shows "http://.../~fabrice/ECLIPSE/conclli/affichage/tabcas1.php". The page content displays several lists of predicates:

- Impasse : dénoncer (3) enfermer (3) confronter (2) contourner (2) mettre un terme (1) aggraver (1) monder (1) faire sortir (1) rendre compte (1) solutionner (1) débattre (1) résoudre (1) tenir compte (1) renvoyer (1) ressortir (1) croître (1) conseiller (1) aviser (1) toucher le fond (1) congeler (1) camper sur ses positions (1) admettre (1) dissenter (1) compromettre (1) vouer (1) mûrir (1) enliser (1) dévoiler (1) ressembler (1) analyser (1) blottir (1)
- avenue : étirer (2) décorer (1) critiquer (1) orner (1) faire la fierté (1) recycler (1) enfêvrir (1) empêcher (1) pétarader (1) trôner (1) spécialiser (1) exiler (1)
- sentier : valoriser (2) mériter (1) extirper (1) randonner (1) charger (1) tourmenter (1) présumer (1) déminier (1) subir (1) ceinturer (1) lutter (1) vouloir bien (1) débroussailler (1) entasser (1) transiter (1) piétiner (1)
- allée : aller et venir (1) sourire (1) sauver (1) filtrer (1) pressentir (1) redescendre (1) prétendre (1) efflocher (1) contenir (1) replanter (1) planer (1) désherber (1) engazonner (1)
- autoroute : surnommer (2) placarder (2) encadrer (2) enfouir (1) paralyser (1) observer (1) appliquer (1) intégrer (1) étrangler (1) moduler (1) river (1) défigurer (1) prendre la direction (1) doter (1) délester (1) nuancer (1) augmenter (1) fuir (1) prononcer (1) louoyer (1) consoler (1) insérer (1)

Document: Done (86.337 secs)

# KWAC-LLI : concordance lines with zones (1)

A screenshot of a Mozilla browser window titled "Affichage - Mozilla {Build ID: 2002091116-SuSE}". The address bar shows "http://.../~fabrice/ECLIPSE/conclli/affichage/affich". The page content includes configuration options and a list of concordance lines.

Format affichage concordance :

Zone0	Zone1	Zone2	Zone3	Zone4
black	italique	black	gras	black

Longueur contextes : 50

Submit Query

### Affichage Général

... -il pas s' endormir au volant lorsqu' il	circule	sur <b>autoroute</b> en dessous de 180 km/h ...
... indiquent clairement aux automobilistes qui	circulent	sur l' <b>autoroute A 8</b> , à quelques kilomètres du poste-fronti...
...une deuxième ligne sur la rive ouest , où	circule	une nouvelle <b>autoroute</b> accrochée à la ligne de crête ...
...is dans la soirée de lundi , alors qu' il	circulait	sur l' <b>autoroute A 9</b> entre Montpellier ( Hérault ) et Roque...
... On peut enfin	circuler	sur l' <b>autoroute</b> , 101 sans craindre les bouchons et manger ...
... Pas davantage rassuré à l'idée de	circuler	sur les <b>autoroutes</b> allemandes ...
... Selon un témoin qui	circulait	sur cette <b>autoroute</b> aérienne , les lumières ont semblé soudain ...
...lés jeter des pierres sur les voitures qui	circulaient	sur les <b>autoroutes</b> A7 et A9 ( Corresp ...
...assassiné , vendredi 9 août , alors qu' il	circulait	sur l' <b>autoroute</b> de Reggio ( Calabre ) ...
...ton conducteur dangereux et interdiction de	circuler	sur <b>autoroute</b> en période déclarée rouge , voire en périod...

Document: Done (2.887 secs)

# KWAC-LLI : concordance lines with zones (2)

The screenshot shows a Mozilla browser window titled "Affichage - Mozilla {Build ID: 2002091116-SuSE}". The address bar shows the URL <http://fabrice/ECLIPSE/conclli/affichage/affichage>. The main content area displays a search interface for "Format affichage concordance" with five zones (Zone0 to Zone4) each having a dropdown menu and a checked checkbox. Below this is a text input for "Longueur contextes : 50" and a "Submit Query" button. The results section is titled "Affichage Général" and contains a table with several rows of search results. The table has columns for context snippets, words, and their parts-of-speech (e.g., circule, sur, autoroute). The results describe various incidents involving cars and roads, such as people throwing stones at vehicles and driving on autoroutes.

Context Snippet	Word	POS	Context Description
... -il pas s' endormir au volant lorsqu' il	circule	sur	autoroute en dessous de 180 km/h ...
... indiquent clairement aux automobilistes qui	circulent	sur l'	autoroute A 8 , à quelques kilomètres du poste-fro
...une deuxième ligne sur la rive ouest , où	circule	une nouvelle	autoroute accrochée à la ligne de crête ...
...is dans la soirée de lundi , alors qu' il	circulait	sur l'	autoroute A 9 entre Montpellier ( Hérault ) et Roq
... On peut enfin	circuler	sur l'	autoroute , 101 sans craindre les bouchons et mang
... Pas davantage rassuré à l'idée de	circuler	sur les	autoroutes allemandes ...
... Selon un témoin qui	circulait	sur cette	autoroute aérienne , les lumières ont semblé soudai
...lés jeter des pierres sur les voitures qui	circulaient	sur les	autoroutes A7 et A9 ( Corresp ...
...assassiné , vendredi 9 août , alors qu' il	circulait	sur l'	autoroute de Reggio ( Calabre ) ...
...tion conducteur dangereux et interdiction de	circuler	sur	autoroute en période déclarée rouge , voire en péri

## 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