



**HAL**  
open science

## 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. 8es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006), Apr 2006, Besançon, France. pp.773-784. halshs-00154100

**HAL Id: halshs-00154100**

**<https://shs.hal.science/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

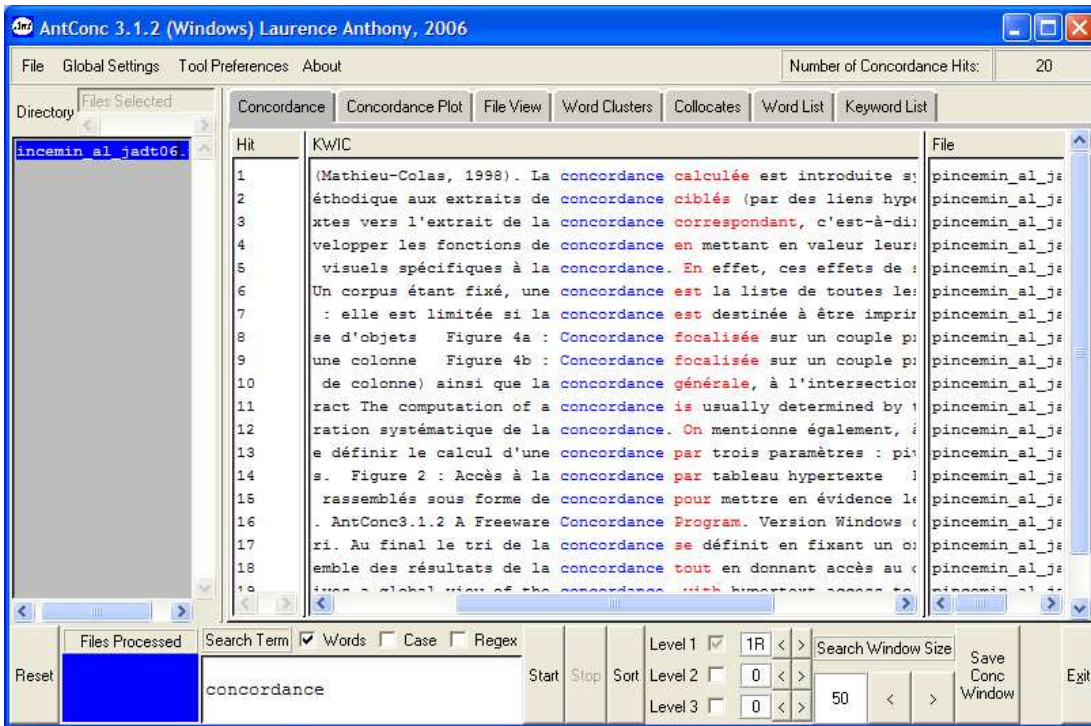
*8èmes Journées internationales d'Analyse statistique des Données Textuelles*

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 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 cursor (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

<i>Left context</i>	shall	- MOT{0,3}	- be .+ed	+ <i>Right context</i>	
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, Frequence	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) :

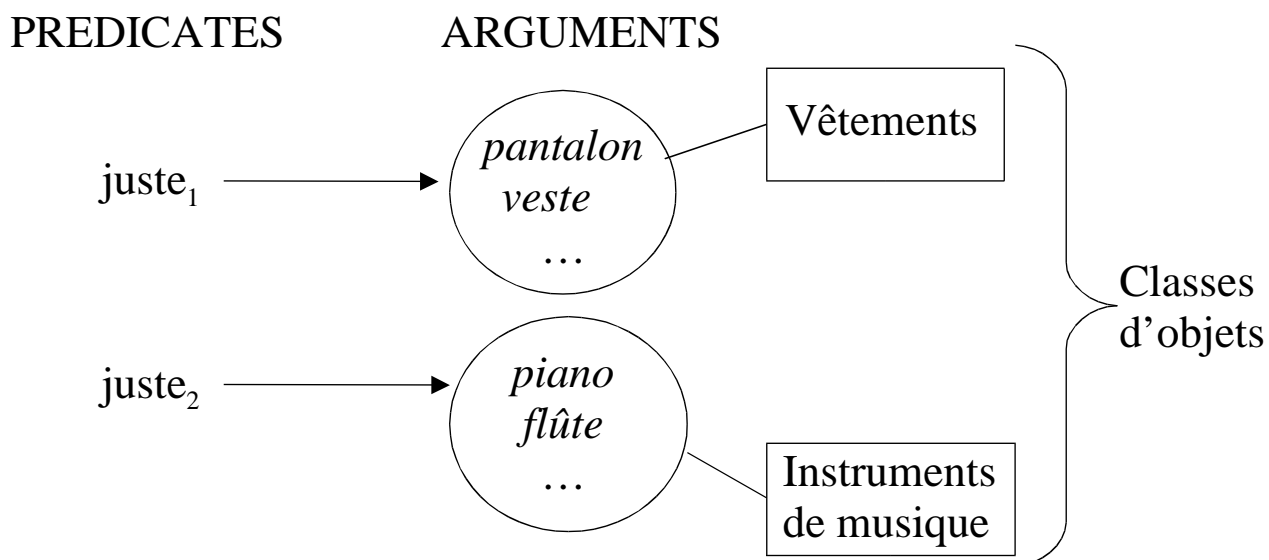
## arguments => predicate

- 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

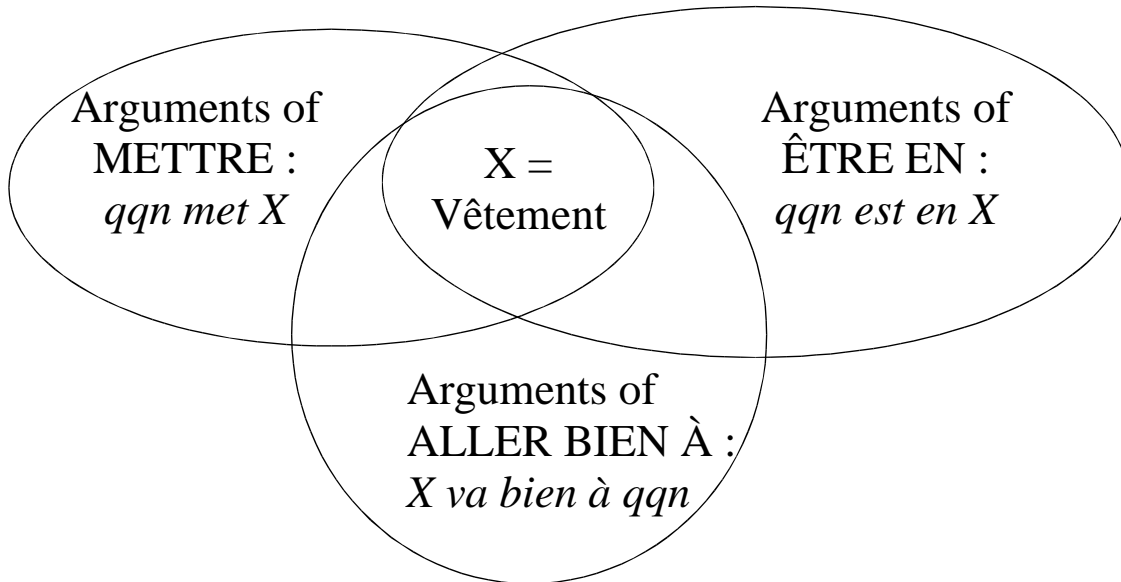
An argument's value is taken from a set called *Classe d'objets*



# 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



## Four ways of exploring a corpus

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

# 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://www. .... ~fa Search

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

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

Position des arguments dans la requete : 3

Position des predicats dans la requete : 1

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

Predicats (separés par le caractère "|") : être|avoir|faire|devoir|pouvoir|falloir|vouloir

effacer les prédicats :  
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			
<u>prendre</u>	888	8	5833	34	310	21	4	2	1	509	7
<u>emprunter</u>	346	8	1867	25	92	25	8	1	4	161	30
<u>ouvrir</u>	263	8	4424	33	103	5	5	1	5	108	3
<u>trouver</u>	89	8	1283	5	18	2	2	6	1	54	1
<u>circuler</u>	83	8	731	26	35	10	3	1	6	1	1
<u>éviter</u>	32	8	1282	3	5	1	2	15	1	4	1
<u>aménager</u>	13	8	405	1	4	1	1	1	2	2	1
<u>sortir</u>	696	7	2433	51	10	8		501	2	6	118
<u>suivre</u>	430	7	3418	16	91	2	2		1	294	24
<u>parcourir</u>	228	7	1519	97	31	5	6		12	71	6
<u>aller</u>	195	7	6474	34	33	4	1	6		115	2
<u>traverser</u>	176	7	1766	96	44	12	13		5	5	1

Document: Done (86.337 secs)

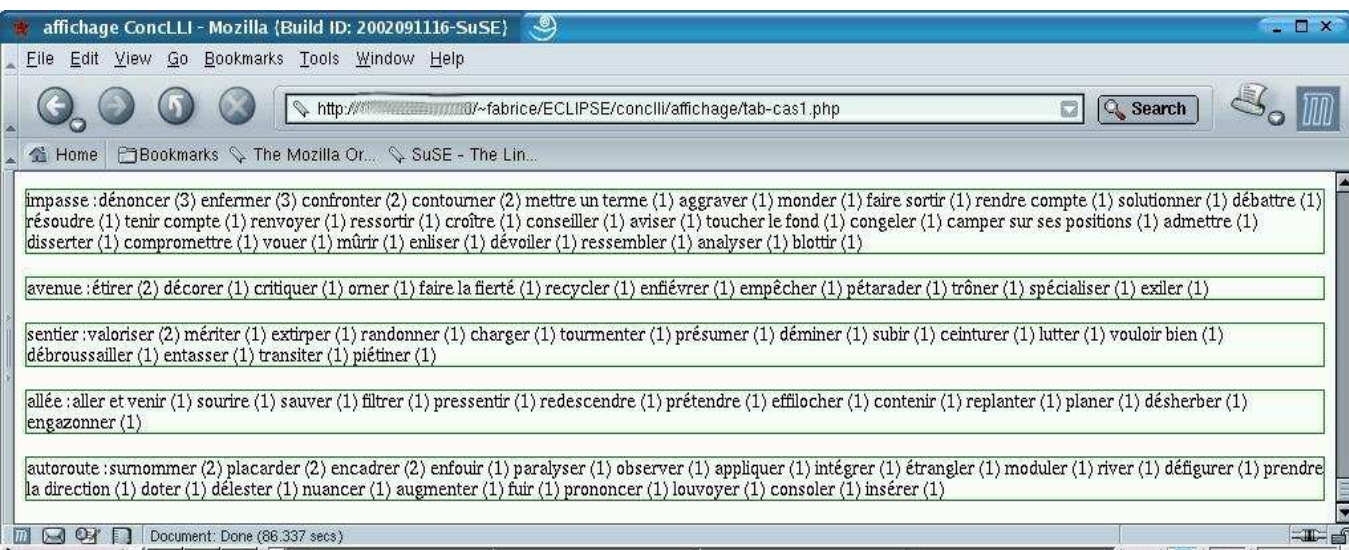
# Specificities of the concordancer

- 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

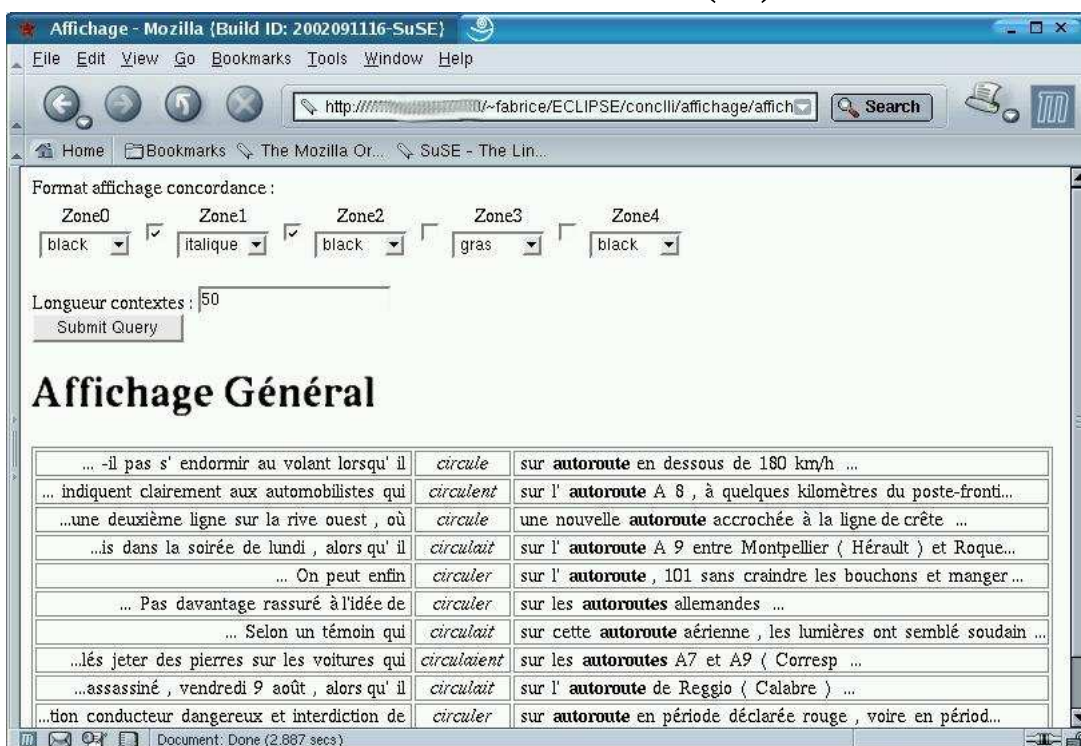
				<u>rue</u>	<u>route</u>	<u>autoroute</u>	<u>avenue</u>	<u>impasse</u>	<u>allée</u>	<u>chemin</u>	<u>sentier</u>
Freq totale				2209	3004	405	231	905	193	3455	357
Freq tab 1				2105	2905	372	213	884	184	3360	336
Nb Total				487	455	160	115	108	93	397	116
Nb tab 1				394	373	131	97	90	84	313	96
Freq corpus				7179	6691	1513	1032	1395	464	6112	879
<u>prendre</u>	888	8	5833	34	310	21	4	2	1	509	7
<u>emprunter</u>	346	8	1867	25	92	25	8	1	4	161	30
<u>ouvrir</u>	263	8	4424	33	103	5	5	1	5	108	3
<u>trouver</u>	89	8	1283	5	18	2	2	6	1	54	1
<u>circuler</u>	83	8	731	26	35	10	3	1	6	1	1
<u>éviter</u>	32	8	1282	3	5	1	2	15	1	4	1
<u>aménager</u>	13	8	405	1	4	1	1	1	2	2	1
<u>sortir</u>	696	7	2433	51	10	8		501	2	6	118
<u>suivre</u>	430	7	3418	16	91	2	2		1	294	24
<u>parcourir</u>	228	7	1519	97	31	5	6		12	71	6
<u>aller</u>	195	7	6474	34	33	4	1	6		115	2
<u>traverser</u>	176	7	1766	96	44	12	13		5	5	1



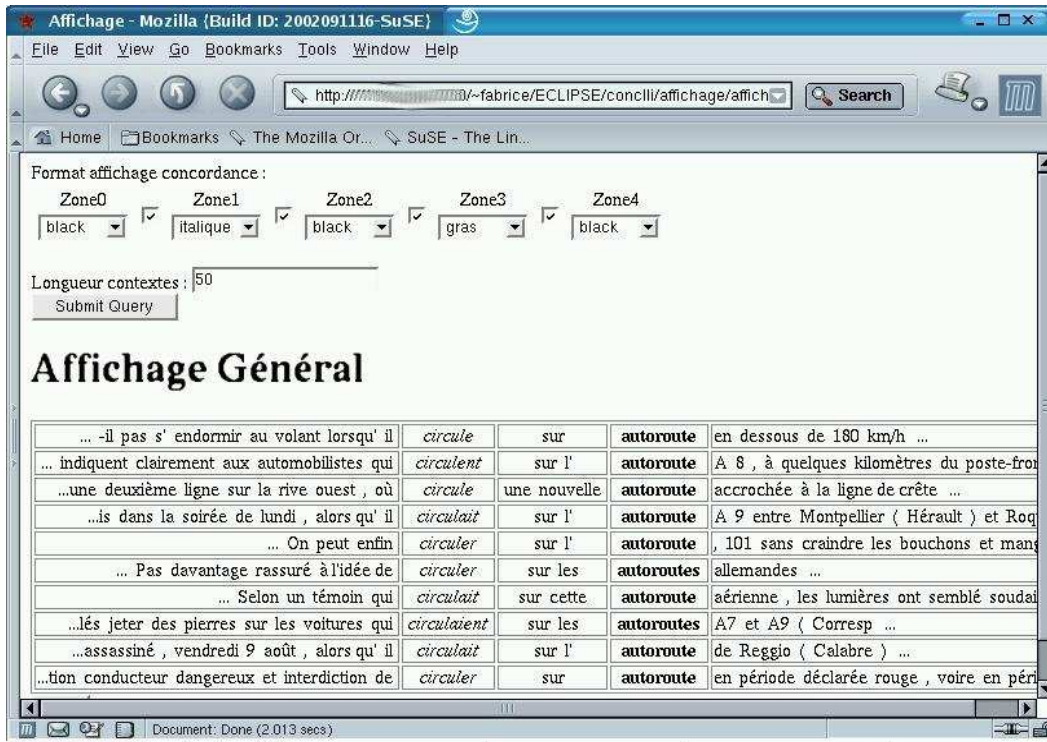
# Lists (out of table) : predicates found with only one argument



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



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



## 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