# Analysis of Language Resources

Ninth Lecture

Hiroyuki Akama

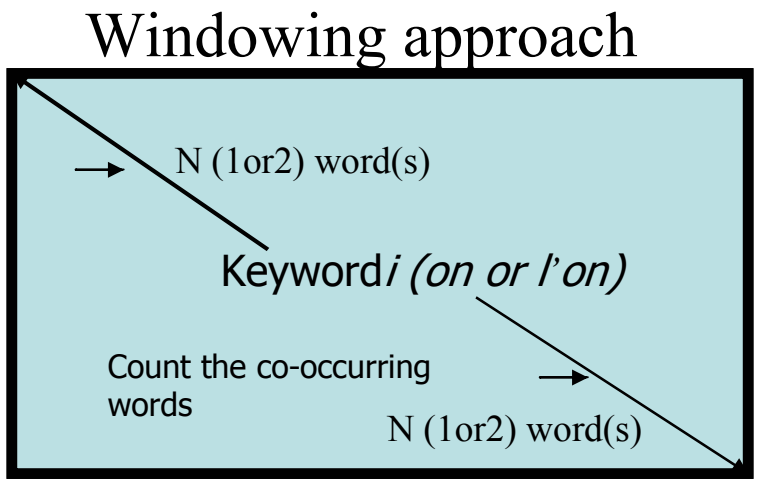
# From the researches made in Akama Laboratory

- Biblical Research:
- *The Frame Problem in Text Analysis*
- *Development of NLP based Synoptic Software for Text Analysis* →Previous time
  - (With Miyake, Nakagawa and Makoshi)
- French Grammar:
- *Probabilistic Language Processing in the Form of a Decision Tree*
  - (With, Shimizu and Shimizu)

# Review

- Lexical co-occurrence data, obtained by a windowing method, would be an important key to the secrets of the language activities and the discursive practices.

We will treat here not only content-sensitive words but also noise words in the interest of the probabilistic model of speech.



# Data Mining of Lexical Information

- Example: Computer-assisted research to find out by a probabilistic method based on the corpus linguistics and the data-mining some practical criteria for choosing ***“on” or “l’on”,*** which are semantically and grammatically equivalent as **French impersonal subject pronoun**.
- Data Gathering: The "window" is set for all the instances of ***“on” or “l'on”*** to obtain information about the neighboring words (the features of the 3 or 5-gram instances centering “on” or “l’on”) .
- Computation: Determine from the viewpoint of information theory (using C5.0 of R. Quinlan) the most appropriate branching diagram (tree) for classifying all the cases of ***“on” and “l’on”.***

# Whether you use “on” or “l’on”

- That is up to you, your preference…, but,…
- Your decision making for this alternative depends unconsciously upon several heterogeneous factors, which are

  1. social-individual factor,
  2. purely phonological factor → Vaugelas
  3. (new discovery) lexical and etymological factor

     → We will attribute it to Damourette & Pichon

# Fixed Phonetic Rules?

- Euphonic Rule : Constraints that are based on the easiness to pronounce, write, hear and read, depending on the neighboring sounds or spellings
- According to Vaugelas et al.,

1) Use "l'on" to avoid a "hiatus" (gap), if the preceding word ends with a vowel sound.

  - ○: **“ou l’on”,**
  - ×: **“ou on” [u-o] (meaning “or one <verb>”)**

2) Use "on" to avoid a "cacophonie" (cacophony), if the next word begins with "l" sound.

  - ○: **“on le dit”,**
  - ×: **“l’on le dit” (meaning “one says that”)**

# Various Factors

- Damourette & Pichon suggested some heterogeneous parameters as historical conditions, intelligence of speakers and diachronic status of words, etc.

1)"L'on" is preferred in the regions where they wrote "en" in place of "on".

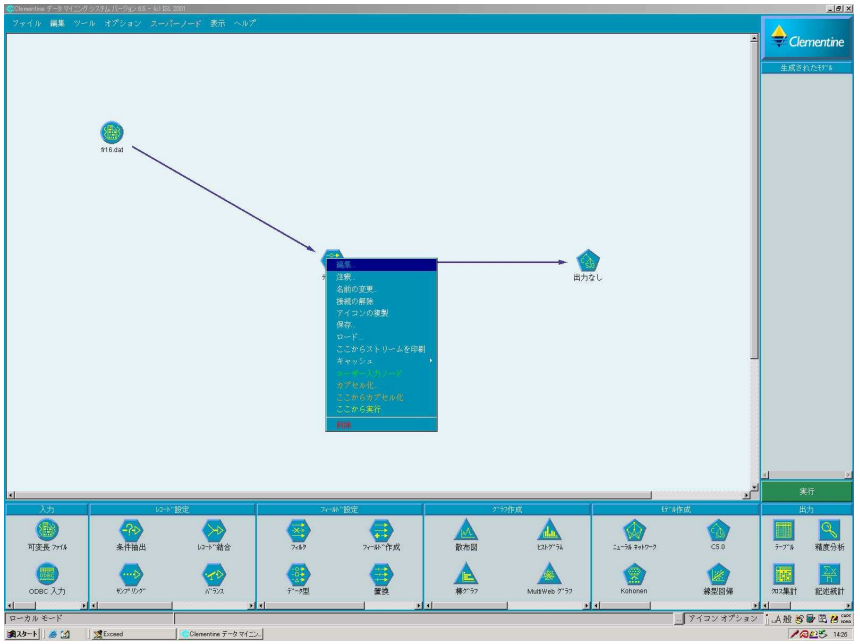
2) It is the persons of letters and the intelligentsia (and the writers of a quality paper ?) that prefer "si l'on" to "si on". The reason was also justified by a historical phenomenon that the word "si" had been in ancient times subject to an "elision“ (omission of a vowel) to make "s'on" like the word "que" to make "qu'on".

# C5.0

- As one of the AI algorithms, originally developed by J.R. Quinlan (1987), uses the Max-Gain method of selecting the best attribute, to build a learning Decision Tree with some rule sets.
- **With this diagram one can classify all examples as either positive (“l’on”: with “l”) or negative (“on”: without “l”) instances of the discursive process.**
- C5.0 algorithm automatically selects certain variables (attributes) and coordinates these, thus splitting the overall data into branches by generating if-then rules at each diverging point.

# Computation

- Classification of data by *Clementine*, data mining toolkit developed by SPSS
- We use the Build *C5.0* (of Quinlan)
  - to construct decision tree on the dataset obtained by using a windowing tool




-- to visualize the causal relationship between the independent parameters in the form of output nodes

# Decision Tree for different parameters

- **(Preceding word: si)** : [most frequent : ***l'on***]
  - **(Corpus: LeM)** : [frequency : 1643, 64.0%] → ***l'on***
  - **(Corpus: DNA)** : [frequency : 1096, 65.1%] → ***on***
- **(Preceding word: ou)** : [frequency : 48, 72.9%] → **on**
- **(Preceding word: où)** : [frequency : 1477, 74.1%] → **l'on**
- **(others)** : [frequency : 54808, 90.4%] → **on**

- **We have to consider the particularity of the words ("si","où") independently of their pronunciation.**
- **Phenomena that cannot be explained from the phonetic point of view.**

LeM (Le Monde) ; DNA (Dernières Nouvelles d'Alsace)

# Summary

- With C5.0 algorithm, we can put together into a stratification of nodes some heterogeneous factors--*phonetic, lexicological and socio-linguistic*– and then can communicate with each other in this causal network to choose “on” or “l’on”.
- The complicated interaction of these diverse parameters is discovered only by a computer-assisted calculation of data mining.
- We have come to conclude that our complex probabilistic analysis is fully effective for the discrimination of “on” and “l’on”.