LSE Department of Methodology,
MY428/528 - LT 2014

Qualitative Text Analysis

Course Convenor: Aude Bicquelet
(a.j.bicquelet@lse.ac.uk)
Office Hours:
Thursday 11:30-13:30

The Alceste Software

(seminar 8)

The Alceste Software

General Background

- Alceste stands for: *Analyse des Lexemes co-occurents dans les énnoncés simples d'un texte*
- Analysis of the co-occuring lexemes within the simple statements of a text
- Lexeme: Fundamental Unit of the lexicon of a language.
  Some lexemes such as *put up with* consist of more than one word.

The Alceste Software

- Alceste aglorithm is based on Benzecri's contributions to textual statistics
- The software was designed by Max Reinert at the CNRS
- It is distributed by the company Image: http://www.image-zafar.com/en/alceste-software

The Alceste Software

- Alceste combines textual and statistical analyses.
- It relies upon co-occurrence analysis.
  = The statistical analysis of frequent word pairs in a text or *corpus.*
- The technical procedure lead to selecting classes, each determined by a pool of words mathematically linked together and having the highest significant frequency of occurrence.
- These classes with their content and function words subscribe to different types of discourse with their specific vocabulary and syntax.

The Alceste Software

Suitable data types

- Parliamentary Debates (Schonhardt-Bailey, 2008; Bara, Weale, Bicquelet, 2007)
- Open-ended questionnaires (Brugidou, 2003)
- Campaign speeches (Schonhardt-Bailey, 2005)
- Interview transcripts (Lahlou, 1996;1998)
- Consultations documents (Bicquelet and Weale, 2011)
- Political Manifestos (Bicquelet 2007)
- Newspaper articles (...)

Any type of text in digitised format above 10.000 words (maximum around 2 million)

The Alceste Software

ICUs

- Initial Context Units (*ICUs*) = Sampling units corresponding to the divisions of the text specified by the user, to which one or several variables can be assigned.

  For instance, in the analyses of parliamentary debates, each utterance (speech act) constitutes an *ICU*.

ECUs

- The corpus is then fragmented into Elemental Context Units (*ECUs*).

- These are gauged sentences that the program automatically constructs based on word length and punctuation in the text.

Alceste crosses the ECUs and the presence/ absence of words (*forms*) in the following way:

	Form 1	**Form 2**	**Form 3**	**…**	**Forms p**
ECU 1	0	1	0	…	…
ECU2	1	0	0	…	…
ECU3	1	1	0	…	…
…	…	…	…	…	…
CUn	…	…	…	…	…

	Doctor	**Army**	**Hospital**	**weapons**	**Nurse**
ECU 1	0	1	0	1	0
ECU2	1	0	1	0	1
ECU3	0	1	0	1	0
ECU4	1	0	1	0	1
ECU5	0	1	0	1	0

	Doctor	Nurse	Hospital	weapons	Army
ECU 2	1	1	1	0	0
ECU4	1	1	1	0	0
ECU1	0	0	0	1	1
ECU3	0	0	0	1	1
ECU5	0	0	0	1	1

This classification proceeds by successively splitting the ECUs into classes on the basis of vocabulary oppositions.

What is obtained is a number of classes of words which should be representative of the main themes of the analysed text.

The Alceste Software

Medical Issues	**Military Issues**
Doctor	Weapons
Nurse	Army
Hospital	ECU1
ECU2	ECU3
ECU4	ECU5

The Alceste Software

- The profile of each class is established in relation to their specific vocabulary.
- Key words and sentences are ranked in terms of their Khi2.
- This indicates the degree of association of the word to the class.

The Alceste Software

- Complementary operations are then performed during the last stage of the analysis.

- Supplementary classifications within different classes (Listing of ECUs)

- *Ascending Classification* (graphic representation of the association between different groups of words)

The Alceste Software

- Factorial analysis of the Correspondences (spatial representation of the relations between the classes)

- The spatial representation provides a way of transforming numerical information about classes, speakers and *forms* (words) into a visual format.

- The distance between points (representing the variables, forms and classes) accounts for their degree of association and is measured by the “chi-squared distance”.

Exercise

Exercise

- Look at the output produced by Alceste for the analysis of newspaper articles in the aftermath of Lance Armstrong's interview by Oprah Winfrey.

- You wish to investigate whether the event was reported in similar/different ways in the **US** and in the **UK**.

- You are also looking at whether the **type** of newspaper (**Tabloid/Broadsheet**) had an influence on the way the story was reported.

- The **corpus** is made up of **8 Articles**:

✓ 2 UK Newspaper articles
✓ 2 US Newspaper articles
✓ 2 UK Tabloid articles
✓ 2 US Tabloid articles