Aggregation and interpretation

Tom Ruette, Dirk Speelman, Dirk Geeraerts

Sociolectometry

Sociolectometry is interested in the structure and function of language varieties or *lects*. Lects are not unidimensional (e.g. dialects), but combinations of socio-demographic (age, gender) and stylistic (register, topic) dimensions. To describe a lect accurately, and to make sure that the description is representative, a large set of linguistic items (e.g. phonological variables) needs to be aggregated. Our study focuses on lexical items.

Research goal

A language variety is *a set of linguistic items with similar social distribution* (Hudson, 1980). Starting from this definition, the goal of this study is to test the usefulness of Weighted Multidimensional Scaling to analyze the relationship between multiple dimensions of social variation across numerous sets of lexical variables, measured as alternation variables.

Distance metric

To measure the distance between lects on the basis of alternation variables, we use the City-Block distance that was also applied in Speelman *et al.* (2003).

Imagine a linguistic function *L* with two possible realizations x_1 and x_2, which were counted in subcorpora representing lect$_1$ (V_1) and lect$_2$ (V_2). These counts were transformed into relative frequencies (*R*) by dividing the raw frequency of x_1 with the sum of x_1 and x_2. The City-Block distance between lect$_1$ (V_1) and lect$_2$ (V_2). for linguistic function *L* is then calculated as follows:

	lect$_1$	lect$_2$
x_1	0.8	0.1
x_2	0.2	0.9

$D_{CB,L}$ = 0.7

$$D_{CB,L}(V_1,V_2)=\frac{1}{2}\sum_{i=1}^{n}|R_{V_1,L}(x_i)-R_{V_2,L}(x_i)|$$

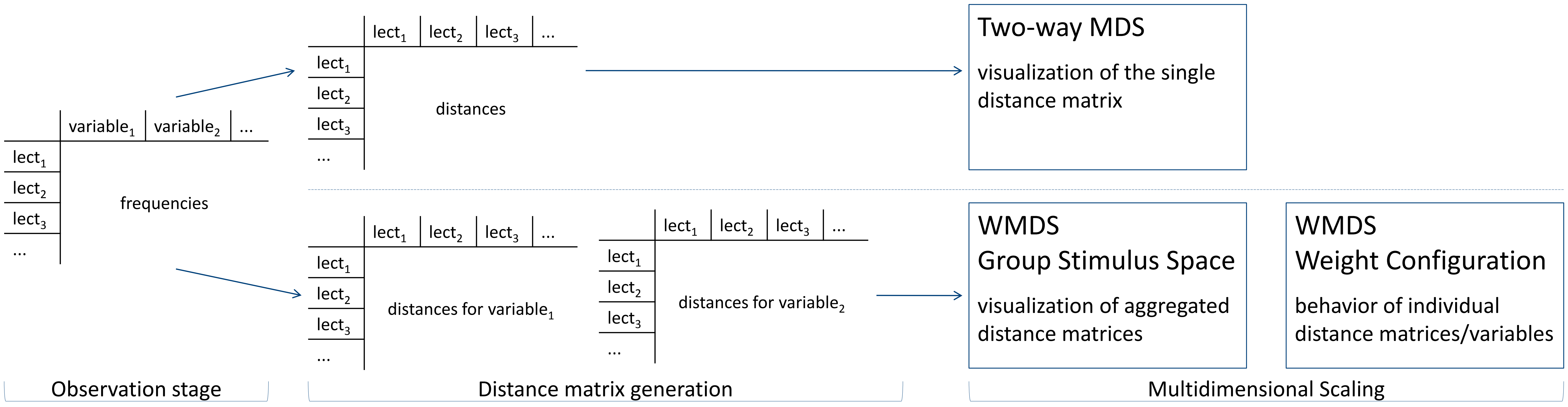
Weighted Multidimensional Scaling

Traditional approach

Aggregation of variables in one distance matrix will obscure the behavior of the input variables

Proposed approach

Postpone variable aggregation step to WMDS, which will grant access to variable behavior.



Two-way MDS

visualization of the single distance matrix

WMDS
Group Stimulus Space

visualization of aggregated distance matrices

WMDS
Weight Configuration

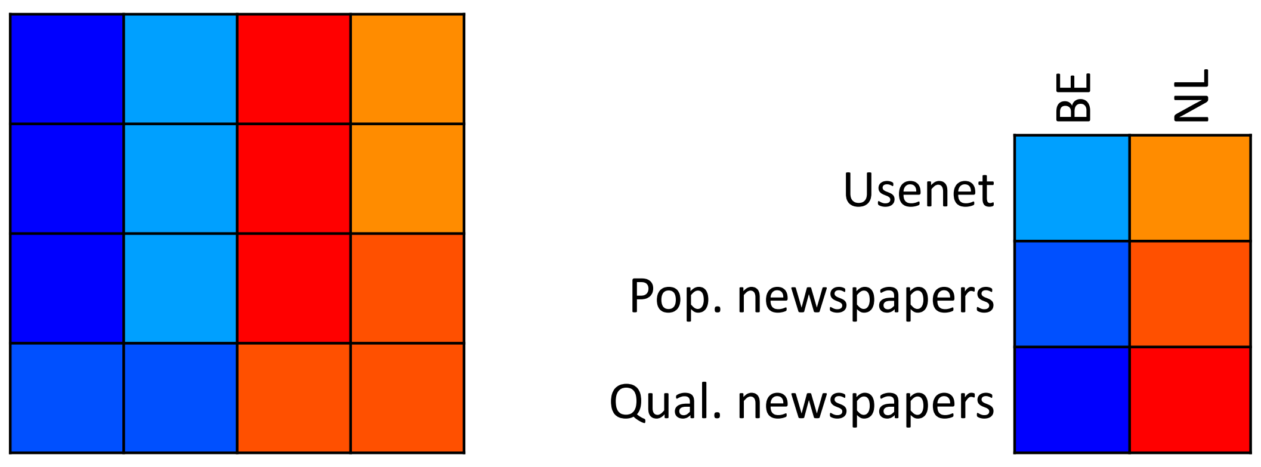
behavior of individual distance matrices/variables

Case study: Register variation in two national varieties of Dutch

Objective

Dutch is a pluricentric language, used in Belgium and The Netherlands. Previous research has already thoroughly investigated variation along this single dimension. In sociolectometry, we extend this unidimensional approach and look into both national and registral variation in Dutch at the same time by aggregating categories of lexical choices.

Corpus



Period: 1999-2004

□ : 8 million words

Full size: 128 million words

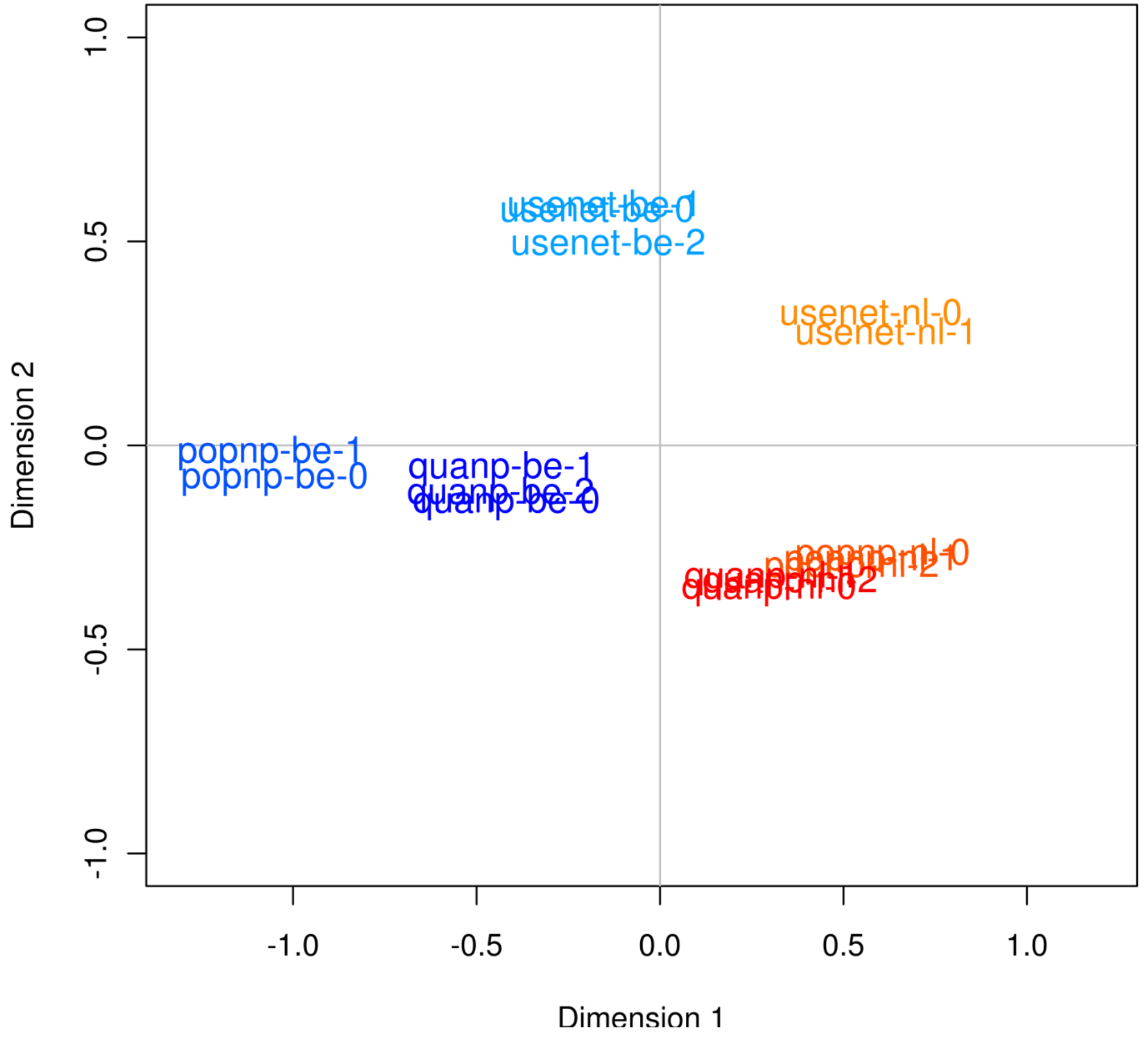
Input Features

From the Reference List of Belgian Dutch (RBBN), described in Martin (2005), we generated 1310 lexical alternation variables. These variables are grouped in six categories (below). We aggregated the variables from each category by using Speelman *et al.* (2003) so that we had six distance matrices. It is the behavior of these categories that we study.

Type	Description	Example	Expected pattern
"vrij"	Competing variants in BE, only one variant in NL	*dollekoeienziekte / gekkekoeienziekte*	Country
"uniek"	Mutually exclusive variants	*confituur / jam*	Country
"cultuur"	Belgian phenomena	*schepen / wethouder*	Country
"omgang"	Colloquial words	*gilet / vest*	Country + register
"restrictie"	Restricted usage, e.g. jargon	*unief / universiteit*	Country + register
"substandaard"	Inappropriate in standard	*bareel / slagboom*	Country + register

Results

Group Stimulus Space



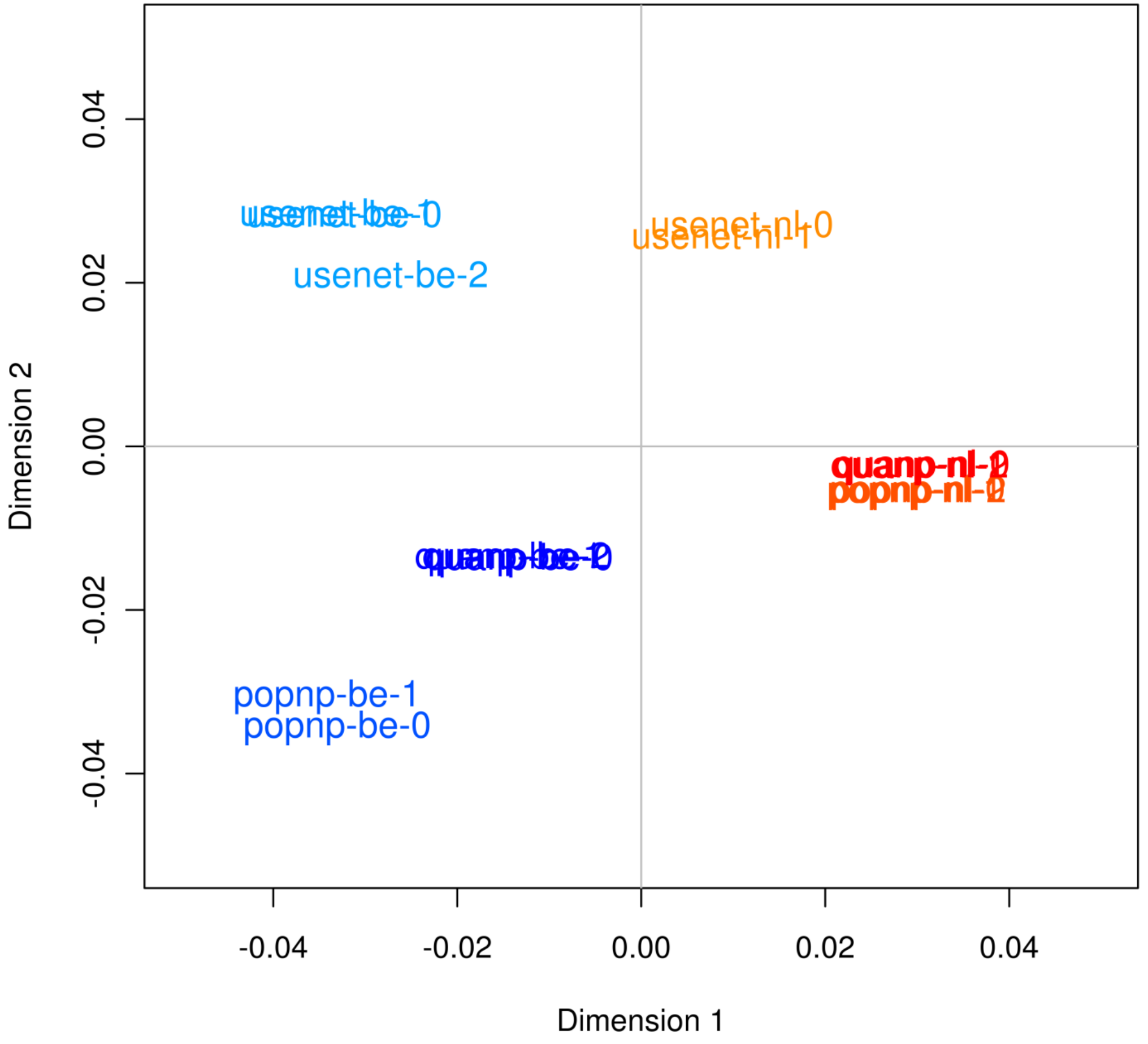
Group Stimulus Spaces visualize the position of the objects on the basis of aggregated distance matrices.

Dimension 1 clearly separates Netherlandic Dutch subcorpora from Belgian Dutch subcorpora.

Dimension 2 separates Usenet subcorpora from newspapers.

Within the Belgian newspapers, there is a clear split between the popular and quality newspapers.

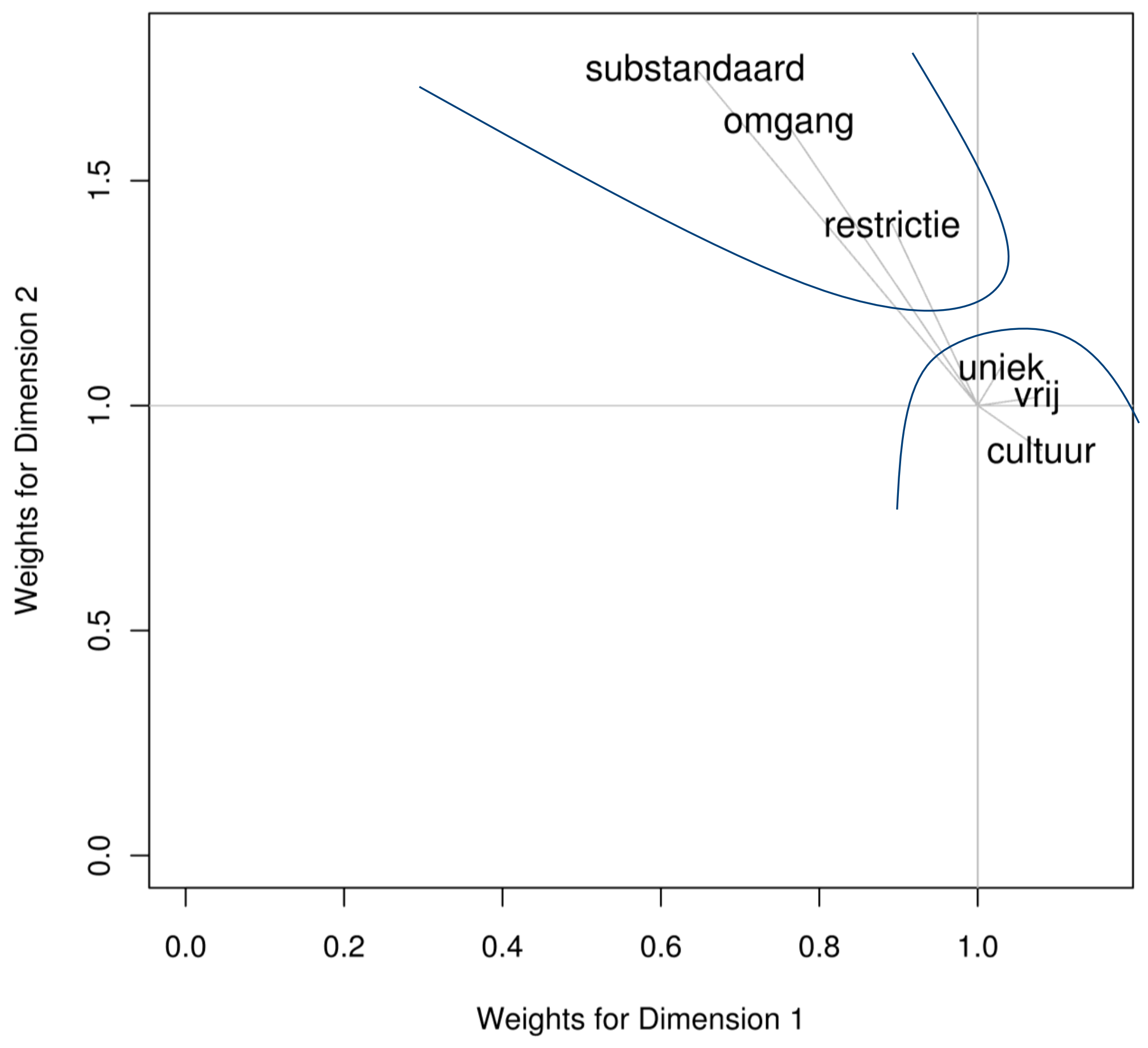
Two-way MDS solution



A two-way MDS approach returns more or less the same positioning of lects as the Group Stimulus Space of WDMS.

However, a two-way MDS does not allow for further research on the individual RBBN categories.

Configuration Weights



The Configuration Weights give an insight in the behavior of the individual RBBN categories.

"substandaard", "omgang" and "restrictie" all stretch Dimension 2 (usenet versus newspapers), because they have a weight > 1 on Dimension 2. This indicates the expected register sensitivity.

"uniek", "vrij" and "cultuur" do not stretch the dimensions since their weights are more or less 1.

Summary

a. The six categories of the Reference List of Belgian Dutch summarize 1310 lexical alternation variables that have been picked to show a national distribution.

b. The Configuration Weights of the WMDS confirm that the national distribution is present in every RBBN category. The aggregation of these categories also causes the primary split of the subcorpora on dimension 1 of the Group Stimulus Space.

c. Three of the RBBN categories were expected to have a registral distribution. The Configuration Weights of the WMDS link up with the expectations, and the Group Stimulus Space shows a register split on dimension 2.

d. These results validate the accuracy of WMDS for a sociolectometric approach that aggregates lexical alternation variables.

Tom Ruette - tom.ruette@arts.kuleuven.be - perswww.kuleuven.be/tom_ruette
University of Leuven/Faculty of Arts/Quantitative Lexicology and Variational Linguistics
Blijde Inkomstraat 21 (PO Box 3308) - 3000 Leuven, Belgium

KATHOLIEKE UNIVERSITEIT LEUVEN