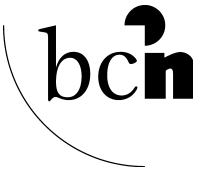
Measuring Dialect Pronunciation Differences using Levenshtein Distance

Wilbert Heeringa

The research has been carried out at the University of Groningen, Faculty of Arts, the Humanities Computing department.



The work in this thesis has been carried out under the auspices of the Behavioral and Cognitive Neurosciences (BCN) research school, Groningen.

Groningen Dissertations in Linguistics 46
ISSN 0928-0030

Rijksuniversiteit Groningen

Measuring Dialect Pronunciation Differences using Levenshtein Distance

Proefschrift

ter verkrijging van het doctoraat in de
Letteren
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
donderdag 8 januari 2004
om 14.15 uur

door

Wilbert Jan Heeringa

geboren op 2 augustus 1970
te Groningen

Promotor	:	Prof. dr. ir. J. Nerbonne
Beoordelingscommissie	:	Prof. dr. H. Goebl Prof. dr. H. Niebaum Prof. dr. G. De Schutter

Acknowledgements

This thesis is attributed to exactly one author as can be seen on both the cover and the title pages. The author is the one who is responsible for the content. But it should be emphasized that many people contributed to the coming about of this thesis. I would like to mention them.

First of all, I thank my promotor, John Nerbonne. More than five years ago, he encouraged me to work on the dialectometry project, and during the project he gave invaluable support. Without his support this thesis would never have been published.

Probably just the pictures in this thesis will catch the eye of the reader. While I implemented programs for calculating distances between language varieties and for clustering them, Peter Kleiweg developed software for creating dendrograms, multidimensional scaling plots and different types of (color) maps. I am grateful to Peter for developing and making available this excellent software, and for his extensive help when creating the figures.

During a visit on a cloudy afternoon, one of my best friends, Martin de Vries suggested that one seek speech segment distances on the basis of an acoustic representation. Becoming the inventor of a new type of voice-producing prosthesis, this approach seemed obvious to him. I thank him for this valuable suggestion.

In cooperation with Roberto Bolognesi I worked on the comparison of Sardinian dialects. In this small project, the use of acoustic segment distances was developed and the use of the Levenshtein distance was improved. I thank Roberto for his help and his friendly cooperation.

In the field of phonetics and phonology I got the help of many persons. I thank David Weenink, Dicky Gilbers, Vincent van Heuven, Wouter Jansen, Angelika Braun, Tjeerd de Graaf and Christine Siedle for explanation and advices. I thank Paul Boersma and David Weenink for making available their excellent PRAAT program.

Norwegian dialects play an important role in this thesis. Jørn Almberg made Norwegian recordings and transcriptions of the fable 'The North Wind and the Sun'. I am grateful to him for this his permission to use this material and for his help during the whole investigation. I thank Charlotte Gooskens for the good cooperation in our Norwegian research, and for her permission to use the results of her perception experiment in Norway. Thanks are due to Charlotte Gooskens

and Sabine Rosenhart for cutting the Norwegian word samples. I thank Arnold Dalen for his help in finding a reliable dialect map and for classifying each of the Norwegian varieties in the right dialect group in accordance with this traditional dialect map.

When exploring the Dutch dialect area, it is important to do this on the basis of a well-chosen network of sites. I thank Jo Daan, Arjen Versloot (Friesland), Sybren Dyk (Friesland), G. H. Kocks (Drenthe), Jacques van Keymeulen (French, West, East and Zealand Flanders) and Joep Kruijsen (Brabant and Limburg) for their useful advice when selecting the varieties. Once the dialects were chosen, the transcriptions had to be digitized. I owe a lot to Saakje van Dellen, Rogier Nieuweboer, Marcus Bergman and Johan Dijkhuis who did the greater part of this donkey work.

Now and then I asked Cor Hoppenbrouwers some questions about the interesting book 'De indeling van de Nederlandse streektalen' which he wrote in cooperation with his brother. I thank him for patiently answering all my questions. I thank Henk Kiers for his help with the Mantel test. I thank Frits Steenhuisen for some help in geography.

Many people help me prevent my English becoming too much like Dutch. I thank John Nerbonne, Jennifer Spenader, James Hammerton and Menno van Zaanen for making corrections. Remaining errors are for my own responsibility! Furthermore, Menno van Zaanen gave also some useful remarks as regards content. Thanks!

Special thanks are due to the members of the reading committee: Hans Goebl, Hermann Niebaum and G. de Schutter. During the hot summer they read the manuscript. Especially I thank Hermann Niebaum for his valuable comments, which I used profitably in this thesis.

I thank all who contributed to this thesis anyway. In the department of Alfa-Informatica, sometimes I felt like a dwarf among giants metaphorically speaking. I am grateful to all colleagues for their collegiality. Furthermore, I thank my two paranimfs for rendering assistance.

Last but not least I thank my parents for their empathy during the more than five years of this project.

In the center of the coat of arms of the university of Groningen, an open Bible is shown. On the left page the words *Verbum Dni* are found, and on the right page the word *lucerna* is printed. This text is a short reproduction for *Verbum Domini lucerna pedibus nostris*: 'the Word of the Lord is a lamp unto my feet' (Psalms 119:105). In the light of this lamp I confess that language variation is a gift of the Creator. It is my wish that this thesis will help scholars in further research of language variation.

Contents