Calculating the sum of language: An interview with John Nerbonne

On January 27, the department of Alfa-Informatica celebrated its 30th anniversary. Alfa-Informatica Professor John Nerbonne has played a key role in the development and success of the department. After many years of devotion to the study of computational language, John Nerbonne gave his farewell lecture on the day that his department turned 30. Time to learn more about his work!

You have a great interest in research on dialects, or Dialectometry. Could you explain more specifically what Dialectometry is?
Dialectology studies how languages vary geographically, and sociolinguistics how they vary socially. Modern studies typically try to gauge both. Dialectometry adds exact measurement to dialectology, for example by checking what fraction of a list of concepts are realized as the same words. In Groningen we have especially championed the measurement of pronunciation differences using the edit distance, also known as Levenshtein distance. The Levenshtein distance indexes the number of changes that are needed to change one word to another word. For example, the Levenshtein distance between *melk* 'milk' (pronounced mElEk in Haarlem) and *molke* (same word in Grou, Friesland) is 3:

- mElEk -> molEk (substitution of 'o' for 'e')
- molEk -> molEke (insertion of 'e' at the end)
- molEke -> molke (deletion of 'E' between 'l' and 'k')

We normally work on phonetic transcriptions rather than orthographics (spellings), and the operations are often associated with different costs, but the example gives you an idea. The focus of the work on dialectometry (six dissertations, including two that won prizes) was methodological, but we have also contributed to theoretical discussions, such as on the diffusion of changes, on the relation between views of dialect landscapes as continua versus partitioned (into areas), and on the relative importance of social and geographic factors.

Which dialects were most interesting to study and why?
This is a tough one. Because of the methodological focus, we have studied dialects of around twenty languages, including Bantu and Turkic languages, and as a linguist, I'm chuffed at that (to borrow a Briticism). However, Dutch remains a big favorite for the incredible density of its variety, and for the opportunity to hear it personally!

> *> Dutch remains a big favorite for the incredible density of its variety, and for the opportunity to hear it personally! <*

Does studying dialectology make you more sensitive to hearing people's accent /dialect?
Yes, I am definitely more sensitive to others' accents. I watched the BBC 'Earthflight', and found myself paying attention to David Tennant's Scottish accent almost as much as the incredible videos of birds in flight. It was also fun to listen for Balkenende's Zeeuws features or the late Cruiff's Amsterdam's.

Studying dialectology also makes that one notices one's own peculiarities a bit more. I come from the East coast of New England, where I might have said "Others notice the differences, and so don't I", which lead most people, once they have heard it, to suspect a home between here and Jupiter. If they have listened carefully to that last example, they sometimes try to clarify whether or not I notice the differences.

In addition to dialectology, you have a passion for statistics. What drives this strong interest in statistics?
A lot of the work in dialectology progressed as we discovered and applied more powerful statistical analyses, so there has been no tension in the two, on the contrary. Martijn Wieling's (2012) dissertation is

>> CONTINUATION OF CALCULATING THE SUM OF LANGUAGE: AN INTERVIEW WITH JOHN NERBONNE

> If nothing else, the statistics give us a chance to examine problems from a fresh perspective, and that is usually interesting scientifically. In fact, statistics offers much broader and deeper possibilities. <

the best example of that. Intellectually I grew up at a time when linguistics, logic, cognitive science and even computer science were all dominated by ideas best modeled categorically, i.e. in discrete mathematics. Just consider the Chomsky hierarchy (Linguistics), modal and intensional logics and Montague Grammar (Logic), Miller, Galanter and Pribram's *Plans and the Structure of Behavior* (CogSci), or all the work on computational complexity (Garey & Johnson, 1979) in computer science. It was no accident that statistics rose in popularity in the nineties as large amounts of data became available, allowing much more sensitive analyses. If nothing else, the statistics give us a chance to examine problems from a fresh perspective, and that is usually interesting scientifically. In fact, statistics offers much broader and deeper possibilities.

Personally, after working from the categorical perspective for fifteen years, I experienced the "statistical revolution" in Linguistics and Computational Linguistics as an exciting development, and I have really enjoyed exploring that perspective. My involvement with BCN also began in the mid-nineties, and the interest in statistics dovetailed nicely with that. Whatever the brain is, it is not a discrete processor, but rather one where multiple, imperfect information sources are combined. It was great to experiment with neural nets at that time – with Erik Tjong Kim Sang and Ivo Stoianov (Ph.D. students), and also with the BCN bio-physicists Hans Stavenga and Diek Duifhuis, from whom I also learned a lot.

You have made important contributions to the success of Alfa-Informatica and you have worked at the department for a long time. What is the biggest development of the department that you witnessed?

Lots has happened over the past twenty-five years, but the most important thing with respect to everyday life in the department has been the enormous growth in interest among students, industry and colleagues in our work. The work has not changed much but now about 50 students per year enter the bachelor programme, a large range of companies clamor for graduates, and, perhaps most gratifying, lots of colleagues are interested in collaboration!

So I suppose this will sound arrogant, but I think the big change is that more people now see the opportunities afforded by computational language processing. We worked hard not to "hide our light under a bushel basket", but lots of others should share the credit for showing the potential of computational linguistics.

Where do you see the department of Alfa-Informatica in the future?

This is also not easy, as I do not think the department covers a discipline in the way that Dutch Language and Literature does, or maybe Neuro-Anatomy. However, I am confident that the work with computational methods will continue to grow and flourish. I am cautious only about the organizational form that it is likely to take.

What can BCN learn from Alfa-Informatica (or the humanities in general)?

The humanities is a fantastic source of questions about the human mind. How do people produce and understand language, how do they learn it? How do we recognize allusions, "In the beginning Zwarts created the office and the organization"? How much is needed to recognize other allusions to *Genesis 1:1*? Why do some words sound unpleasant, like *runt*, *moist*, *scum*, *fester* or *phlegm*? And there are many more questions!

Which question that you have never solved do you hope to see answered in the future?

Formulating a question is tough, but, as I see it, we have made enormous progress with respect to the understanding of languages from the perspective of cognitive psychology, a programme due to Chomsky. At the same time, language is a social mechanism and is shaped by the functions it serves socially – allowing people to exchange information, or inviting inferences about speakers based on how they speak, a topic studied in dialectology. I would like to see us make progress toward understanding how the cognitive and social perspectives interact. They interact in simple ways that we have explored in dialectology.
For example, we can induce phonetic differences (cognitive) from the distributions of dialect pronunciations (social), and some of the constraints that play a role in spoken word recognition (cognitive) also play a role in which pronunciation differences are socially interpreted (social). Currently, there is no overarching research programme trying to link the two, and I would hope to see that in the future.

■ BY AMÉLIE LA ROI
■ PHOTOS BY MIGUEL SANTÍN

References

Garey, M. R., & Johnson, D. S. (1979). Computers and intractability: a guide to the theory of NP-completeness. 1979. San Francisco: Freeman.

Miller, G. A., Galanter, E., & Pribram, K. H. (1986). Plans and the structure of behavior. New York: Adams Bannister Cox.

Wieling, M. (2012). A quantitative approach to social and geographical dialect variation. PhD dissertation, University of Groningen.