Representation, information theory and basic word order

Luke Maurits

Thesis submitted for the degree of
Doctor of Philosophy
in
Psychology
at
The University of Adelaide

School of Psychology

THE UNIVERSITY
of ADELAIDE

September 2011

Contents

Abstract

Many of the world's languages display a preferred ordering of subject, object and verb, known as that language's *basic word order*. There are six logically possible basic word orders, and while each occurs in at least one known language, not all are found equally frequently. Some are extremely rare, while others are used by almost half the world's languages. This highly non-uniform cross-linguistic distribution of basic orders is a fundamental explanatory target for linguistics.

This thesis tackles this problem from a psychological perspective. It constitutes an advance over previously proposed explanations in that it is compatible not only with the distributions observed today, but with what is known of broad trends in the word order change which happen over hundreds of years. There are two largely independent components of the explanation given in this thesis, which is necessary to be compatible with both synchronic and diachronic evidence.

The first component is focused on the structures which the human mind uses to represent the meanings of sentences. While mental representations of meaning are not inherently serial (hence ordered) like spoken language, we can think of the different components in these representations as being ordered in a different sense, based on some components being more accessible to cognitive processing than others. This thesis develops the idea that the word order used most often in the earliest human languages, which are taken to rely on a direct interface between mental representations and motor control systems, were determined by a "word order of the language of thought".

The second component is focused on the functional adequacy of different word orders for high speed, reliable communication. The driving idea here is that human language represents a rational solution to the problem of communication. The mathematical formalism of information theory is used to determine the gold standard for solutions to this problem, and this is used to derive a ranking of word orders by functionality. This thesis develops a novel perspective on word order functionality in which cross-linguistic preferences are ultimately a reflection of statistical properties of the events which languages describe.

Signed Statement

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

SIGNED: ..... DATE: .....

Acknowledgements

First and foremost, I owe a tremendous debt of gratitude to my supervisors, Dan and Amy. Before beginning this PhD I knew very close to nothing about cognitive science or linguistics, and at the end it of a huge proportion of what I know is either something one of them taught me or something I taught myself after one of them pointed me in the appropriate general direction. Shortly after meeting Dan and having my academic worldview (which I developed as an undergraduate exclusively in maths and physics) violently shifted by learning that there were people who called themselves psychologists *and* used maths (at the same time!), I was offered the opportunity to start a PhD under him, and I am thankful for the leap of faith on his behalf in doing this despite my total lack of relevant background. Looking back at the research proposal I wrote six months into the endeavour, it's embarassing how naively overconfident I was about breaking amazing new ground in psycholinguistics (using Markov chains, no less), and it is a small wonder that when Amy turned up at around this time she took me as seriously as she did. I am very lucky to have had two approachable and helpful supervisors with strong mathematical and computational backgrounds to guide me while I found my feet in a new field.

Natalie May and Tim Ck, my fellow members of Dan and Amy's labs at the University of Adelaide, assisted with recruiting and scheduling participants for the experiments presented in this thesis, which was a tremendous time-saver for me. I am grateful to both of them.

While working on this thesis, I was lucky enough to be able to travel to Vancouver for the Neural Information Processing Systems conference to talk on an embryonic form of the material which appears in Chapter 9 of this thesis. I was able to travel to Canada to attend NIPS thanks to a number of sources of funding beyond that provided by the University of Adelaide, including a grant from the Walter & Dorothy Duncan Trust and a NIPS travel grant which was sponsored by Google.

As part of the trip to NIPS, I was able to visit a number of universities in the northern hemisphere to discuss my work, in many cases with scholars upon whose ideas I was directly buiding. I am very grateful to everyone who allowed me to visit their lab: Nick Chater at the University College, London's Cognitive, Perceptual and Brain Sciences Research Department and University of Warwick's Warwick Business School, Tom Griffiths at the University of California, Berkeley's Computational Cognitive Science Lab, Simon Kirby at the

University of Edinburgh's Language Evolution and Computation research unit and Roger Levy at the University of California, San Diego's Computational Psycholinguistics Lab. Not only did they allow me to speak at their labs, but they and many of their colleagues took particular interest in my ideas on word order and Uniform Information Density and discussed their thoughts on this matter with me at length. Having my ideas taken seriously by "real linguists" was a tremendous confidence boost for someone who very much felt like they had been seriously theorising by the seat of their pants.

I am thankful to the many graduate students at the institutions above who offered their companship during my visits, and particular to those who provided me with transportation or accommodation, who were: Sean Roberts, Marton Soskuthy and Rachael Bailes in Edinburgh, Joseph Austerweil and Karen Schloss in Berkeley and Klinton Bicknell in San Diego. I am also indebted to Matt Hall at UCSD for a very enthusiastic hour of discussing and swapping references on improvised gestural communication.

Merrit Ruhlen of Stanford University, Matthew Dryer from the University at Buffalo and Albert Bickford at the Summer Institute of Linguistics all provided me with copies of papers which I was unable to locate otherwise.

Richard Sproat, from Oregon Health & Science University, whom I met at the HCSNet WinterFest event in Sydney in 2010, introduced me to the World Atlas of Language Structure, which proved to be incredibly useful for speculating about word order typology.

Kayako Enomoto, at the University of Adelaide's Centre for Asian Studies, bent rules against strong Faculty resistance in allowing me to audit the Japanese 1A class, giving me my first practical experience with a language typologically dissimilar to English, which was tremendously helpful in thinking about language, and particularly syntax, in a more general way than I could have otherwise.

Finally, I am grateful to my parents for their encouragement and support throughout the last three and a half years, and to my wife Kirsty for the same, as well as for proofreading and for putting up with a lot of neglect in the final weeks of preparing this thesis!