A Hubterranean View of Syntax

A Hubterranean View of Syntax: An Analysis of Linguistic Form through Network Theory

By

Julie Louise Steele

CAMBRIDGE SCHOLARS

PUBLISHING

A Hubterranean View of Syntax:
An Analysis of Linguistic Form through Network Theory,
by Julie Louise Steele

This book first published 2012

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Copyright © 2012 by Julie Louise Steele

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-4438-3577-3, ISBN (13): 978-1-4438-3577-0

~This study is dedicated~

~to Robert and Becky~

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First and foremost, I'd like to express my deepest thanks to John Ingram my principal supervisor. Immensely patient and always supportive, his encouragement can be found in every aspect of my time at UQ. John constantly provided me with valuable feedback; always there to drive my understanding and offer insightful and thought provoking questions. Thank you John for everything but especially for the most important gift you could give: thank you for believing in me. Also, my warmest thanks to Mary Laughren, David Lee and especially Lynn Wales. Lynn was my first port of call at UQ and I am very thankful for her supportive role.

Some very patient and clever people saved my computer bacon by writing programs to assist my research and I am eternally grateful to them. Many thanks to Andrew Smith for tutoring me in the Leximancer software as well as tirelessly writing modifications especially for my research. I would also like to express my immense gratitude to Hans Jørgen Klarskov Mortensen for writing the program that calculated the frequencies and probabilities, for his patience and tenacity in the development of the software, for recommending literature to read and for enduring with such good humour my first tries and mistakes at writing the Danish language. Thank you so much Hans! Warmest thanks to Sean Fowler who wrote a very handy program that saved me days of tedious work and who never fails to make me laugh with his endearing humour. I'm also grateful to David Lee who shared his macro wizardry and made my computer life so much easier – thank you! Cheers also to Mike Proctor for his help with Perl.

I'd like to thank Mikael Bodén and Rob Pensalfini for reading through my candidature documents. Thanks also to Rob for his comments in the initial stages of my research. I'd also like to honour the contribution that John Ingram, Lynn Wales, Michael Harrington, Mary Laughren and Rob Pensalfini made in establishing a supportive environment for all the students in the Linguistic program at UQ by hosting parties and get-togethers. Their effort makes studying linguistics at UQ the pleasure it is. A big round of applause goes to those attendants of the UQ Linguistics Seminar Series, for their thought-provoking questions. Special thanks go

to John Ingram, Lynn Wales, Mary Laughren, Michael Harrington and Henry Brighton for their helpful comments. Many thanks to the EMSAH administration team, especially Annette Henderson, for their support.

I'd also like to give my fellow members of the Phonetics Lab a hearty pat on the back, especially Sacha DeVelle (an honorary member!), Thu Nguyen, Abdel el Hankari, Diana Guillemin, Tina Pentland and Sacha Rixon for their company. Thanks also to Nik Geard, Ben Skellet and Penny Drennan. To Lorraine and Jon Creedon, thank you for your friendship all these years. I also wish to express my appreciation for the support given by Misi Brody, Yves Le Clézio, Stefani Anyadi, Eleni Gregoromichelaki, Ann Warunwatcharin, Marsha Hill and Crispin Lane in those early days. My thanks to Ricard V. Solé for his permission to use the pretty graphic on the cover of this manuscript. And thank goodness for PhD comics!

I would especially like to convey my immense gratitude to the organisations that have helped me financially through my study. Firstly to the Victim support "Criminal Injuries Compensation" group in the UK whose compensation payment helped me finance my first year at UQ. I am immensely appreciative of the generosity of CRLPL at UQ who gave financial support for two years of my study and am indebted to the EMSAH department and the UQ Graduate School for the postgraduate research stipend which made studying at UQ financially possible. Also, I would like to express my appreciation and immense thanks for the support I received from Richard Fotheringham as the head of school.

In these closing words, I could not pass up the opportunity to tip my golden straw hat to my little friends in the land of grain. To my parents Mick and Carole, my brother David and my Nan Mona, much love and thanks for everything. In memory of absent family - especially beloved Nan Brenda and my cousin Linda - thinking of you all. And finally, to Robert Feyrer, whose supportive and wonderfully wise words continue to have such a profound impact on my life. Thank you with all my heart.

ABSTRACT

Language is part of nature, and as such, certain general principles that generate the form of natural systems, will also create the patterns found within linguistic form. Since network theory is one of the best theoretical frameworks for extracting general principles from diverse systems, this study examines how a network perspective can shed light on the characteristics and the learning of syntax.

It is demonstrated that two word co-occurrence networks constructed from adult and child speech (BNC World Edition 2001; Sachs 1983; MacWhinney 2000a) exhibit three non-atomic syntactic primitives namely, the truncated power law distributions of frequency, degree and the link length between two nodes (the link representing a precedence relation). Since a power law distribution of link lengths characterises a hubterranean structure (Kasturirangan 1999) i.e. a structure that has a few highly connected nodes and many poorly connected nodes, both the adult and the child word co-occurrence networks exhibit hubterranean structure.

This structure is formed by an optimisation process that minimises the link length whilst maximising connectivity (Mathias & Gopal 2001 a&b). The link length in a word co-occurrence network is the storage cost of representing two adjacently co-occurring words and is inversely proportion to the transitional probability (TP) of the word pair. Adjacent words that co-occur often together i.e. have a high TP, exhibit a high cohesion and tend to form chunks. These chunks are a cost effective method of storing representations. Thus, on this view, the (multi-) power law of link lengths represents the distribution of storage costs or cohesions within adjacent words. Such cohesions form groupings of linguistic form known as syntactic constituents. Thus, syntactic constituency is not specific to language and is a property derived from the optimisation of the network.

In keeping with other systems generated by a cost constraint on the link length, it is demonstrated that both the child and adult word co-occurrence networks are not hierarchically organised in terms of degree distribution (Ravasz and Barabási 2003:1). Furthermore, both networks are disassortative,

and in line with other disassortative networks, there is a correlation between degree and betweenness centrality (BC) values (Goh, Kahng and Kim 2003). In agreement with scale free networks (Goh, Oh, Jeong, Kahng and Kim 2002), the BC values in both networks follow a power law distribution.

In this study, a motif analysis of the two word co-occurrence networks is a richly detailed (non-functional) distributional analysis and reveals that the adult and child significance profiles for triad subgraphs correlate closely. Furthermore, the most significant 4-node motifs in the adult network are also the most significant in the child network. Utilising this non-functional distributional analysis in a word co-occurrence network, it is argued that the notion of a general syntactic category is not evidenced and as such is inadmissible. Thus, non- general or construction-specific categories are preferred (in line with Croft 2001).

Function words tend to be the hub words of the network (see Ferrer i Cancho and Solé 2001a), being defined and therefore identified by their high type and token frequency. These properties are useful for identifying syntactic categories since function words are traditionally associated with particular syntactic categories (see Cann 2000). Consequently, a function word and thus a syntactic category may be identified by the interception of the frequency and degree power laws with their truncated tails. As a given syntactic category captures the type of words that may co-occur with the function word, the category then encourages consistency within the functional patterns in the network and re-enforces the network's (near-) optimised state. Syntax then, on this view, is both a navigator, manoeuvring through the ever varying sea of linguistic form and a guide, forging an uncharted course through novel expression.

There is also evidence suggesting that the hubterranean structure is not only found in the word co-occurrence network, but within other theoretical syntactic levels. Factors affecting the choice of a verb that is generalised early relate to the formation and the characteristics of hubs. In that, the property of a high (token) frequency in combination with either a high degree (type frequency) or a low storage cost, point to certain verbs within the network and these highly 'visible' verbs tend to be generalised early (in line with Boyd and Goldberg forthcoming). Furthermore, the optimisation process that creates hubterranean structure is implicated in the verb-construction subpart network of the adult's linguistic knowledge, the mapping of the constructions' form-to-meaning pairings, the construction

inventory size as well as certain strategies aiding first language learning and adult artificial language learning.

Keywords: syntax, network theory, child language, power laws, frequency

Australian and New Zealand Standard Research Classifications (ANZSRC): 200408 Linguistic Structures (incl. Grammar, Phonology, Lexicon, Semantics) 65%, 010299 Applied Mathematics not elsewhere classified 35%.

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCING THE HUBTERRANEAN WORLD OF LINGUISTIC FORM

This study explores the idea that patterns in nature may be realised in the linguistic form of our own conversations; that our words dance to the same tune that is played out in our world. It takes little effort to observe in nature how certain patterns reappear in quite diverse environments. For example, the branch configuration of a tree and its leaf structure echoed in the distributary arrangement in a river delta and the blood vessels of a kidney. Recall the spiral of a shell, its shape reflected in the wind currents of a tornado, the florets of a sunflower head and the curl of a ram's horn.

Interestingly, these pervasive patterns have quite simple factors underpinning their shape. Spiral structures may be captured by the mathematical equiangular spiral configuration, a structure which retains its shape as its size increases. The biological branching structure is captured by simple principles related to hierarchical branching, size invariance of 'terminal units', and the "minimization of the energy and time required to distribute resources" (Brown, West and Enquist 2005: 735; West, Brown and Enquist 1999). The enquiry of this study is whether language too reflects omnipresent patterns in nature and in the world at large; patterns that are created by very simple and general principles. If so, these principles may explain why language has certain properties once thought of as specific to language and/or to cognition.

Studies have shown that when one views linguistic form as a word network, language does share specific properties with other systems (Ferrer i Cancho and Solé 2001a), such as social networks (Granovetter 1973, 1983; Amaral, Scala, Barthélémy and Stanley 2000), genetic transcription networks in certain mammals (Potapov, Voss, Sasse and Wingender 2005), ecological food webs (Montoya and Solé 2002), to name but a few. These properties are known as the small world and the hubterranean structure. Before we start this discussion, it would be

pertinent to consider how such eclectic systems are compared.

1.1 Network approach

When comparing systems from diverse domains, the network approach is one of the most versatile ways of embodying such systems, as the theoretical machinery behind the network model is very simple. The entities being observed are represented by *nodes* or *vertices* (the circles in Figure 1.1) and the relation between nodes is denoted by a link. A link can be either undirected i.e. an edge or directed i.e. an *arc*. An arc is shown as the arrow in Figure 1.1. Any relation between entities can theoretically be modelled by a network.[1]

For example, in a food web, node A in Figure 1.1 may represent a predator and node B the prey; the link denoting that entity A feeds on entity B. In a language network, the nodes may represent words (Ferrer i Cancho and Solé 2001a; Steyvers and Tenenbaum 2005), phonological word-forms (Vitevitch 2008), sets of synonyms (Miller 1990) or word meanings (Steyvers and Tenenbaum 2005), etc. Links between words would then denote syntactic/collocation relations, association relations or the relation of phonological neighbour, respectively. Whereas sets of synonyms are linked by "conceptual-semantic and lexical relations" (Wordnet 2008), and word meanings by hypernymic and meronymic relations.

Figure 1.1. A network link (arc). Nodes represent entities whilst links represent relations

[1] Note that this thesis utilizes the network approach as a tool for representing relations between entities. On this view, the network is not a shorthand notation for a generative algorithm that generates grammatical sentences. If the network were such an algorithm, then the precedence relation represented by the network connections would seem overly generative. However, since the network is a means of representing distributional patterns of relations, the network facilitates the recognition of patterns within the domain of study and allows us to compare these patterns with those found in other systems in the world. (I'd like to thank Michael Harrington for discussing this issue).

In this study, the focus of attention is on how syntactic relations in adult and child speech ay reflect other general properties of systems found in our world. Basic syntax is concerned with word combinations and as such, basic syntactic relations will be identified by distributional analysis. Adopting a purely (non-functional) distributional analysis approach means that language may be represented as a network in which the nodes symbolise words and the links indicate the words that occur adjacently within a sentence.

An example of what a word co-occurrence network would look like is given below. The network is formed from converting word co-occurrences in the text (box a) to links (c). The direction of the link indicating the linear temporal order (precedency) of the words i.e. the preceding word points to the following word.

In representing linguistic form as a word co-occurrence network, this study suggests that (at least) two broad characteristics of language, namely basic constituency and the functional- lexical distinction may be explained by appealing to a certain property that is ubiquitous in the world of patterns: the small world effect. It was observed that many systems exhibit a small average shortest path and a high clustering co-efficient (Watts and Strogatz 1998; see below for definitions). Since these properties make it easier, that is, quicker to traverse from one node to any other in a given network, they were termed small world properties.

A path length between two nodes equates to the number of connections traversed across the network. For example, the path length between the word *about* and the word *and* in Figure 1.2 is two, since two connections are traversed between one node and the other. When the shortest path between all the nodes in the network is calculated and then averaged, in a small world network this value[2] is small compared to a random network[3] and "and offers a measure of a network's overall navigability" (Barabási and Oltvai 2004: 102).

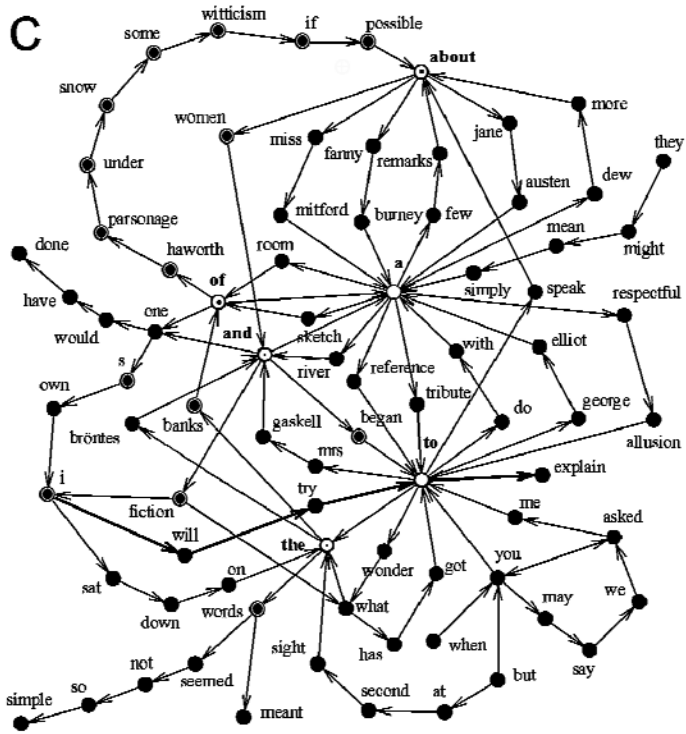
[2] This value is also known as the network diameter (Albert, Jeong and Barabási 1999).

[3] A network in which the nodes have roughly the same number of links is known as a random network (see Erdös and Rényi 1960).

Figure 1.2. Language as a network of word co-occurrence links. Sourced from Solé, Corominas-Murta, Valverde and Steels (2005: 2, Figure 2A&C).

A *"But, you may say, we asked you to speak about women and fiction -what has got to do with a room of one's own? **I will try to explain**. When you asked me to speak about women and fiction I sat down on the banks of a river and began to wonder what the words meant. They might mean simply a few remarks about Fanny Burney; a dew more about Jane Austen; a Tribute to the Brönte and a Sketch of Haworth Parsonage under snow; some witticism if possible about Miss Mitford; a respectful allusion to George Elliot; a reference to Mrs Gaskell and one would have done. But at second sight the words seemed not so simple."*

- Virginia Wolf, *A Room of One's Own*



> "The average clustering coefficient,<C>, characterizes the overall tendency of nodes to form clusters or groups" (Barabási and Oltvai 2004: 102) in the sense that if node A connects to node B, and node B connects to node C, then how likely is it that node A is connected to node C. A small world has a high propensity to form such groups, in that "the average clustering coefficient of all nodes with *k* links" is high (Barabási and Oltvai 2004:102).

1.2 Hub structure

One way of generating a small average shortest path length - as well as a high clustering co- efficient - is to introduce hubs into the network. Hubs are highly connected nodes, relative to the connectivity of all the nodes (Jeong, Tombor, Albert, Oltvai and Barabási 2000). Hubs increase the overall connectivity of the network by providing short-cuts between two nodes. This idea can be easily visualised by a transportation network. In this network, the nodes are population bases e.g. villages, towns, cities whilst the connections represent the road system servicing and connecting these places. Major cities are the hubs of the network since many highways and roads converge in these cities. If one wanted to travel from a village in one state to a village in another state, the quickest way is to aim for largest city closest to the target village, perhaps passing through other major cities on the way i.e. hubs in the network, rather than travelling on the back roads. Thus, it is easy to see how hubs help navigation through a network and so tend to create a small world.

You may not be surprised that language exhibits these short-cuts, that is, hubs, when language is represented as a word co-occurrence network. It has been shown that the connectivity i.e. degree distribution of nodes in a linguistic network may be represented by a power law (scale- free) function (Ferrer i Cancho and Solé 2001a). Formally, the function represents "the probability [P] that a selected node has exactly *k* links" which is inversely proportional to the power *y* of *k* i.e. $P(k) \sim k^{-y}$ (Barabási and Oltvai 2004:102). This distribution equates to a network with many sparsely connected nodes with a few nodes that are highly connected i.e. hubs. These hub nodes tend to represent the 'function' words (Ferrer i Cancho and Solé 2001a).

If the hubs tend to denote words that are construed as function words, then perhaps the hub nodes within a word co-occurrence network are implicated in the definition and the explanation of the theoretical concept. This

proposition hints at the possibility that the hubterranean structure may account for very basic syntactic characteristics of language. The functional-lexical (f-l) distinction is the obvious place to start investigating but also since the network encodes precedence relations between words, it would be worthwhile exploring if constituency (i.e. how words combine) might also be related to this hubterranean structure.

Thus, this study examines the possibility of the f-l distinction and constituency being correlated with the properties of the hubterranean structure in a small world network. Furthermore, when this correlation stands (and it will be argued that it does), it would be interesting and productive to know the mechanisms and processes behind hub formation since this will inform us of why language exhibits these particular properties.

Mathias and Gopal (2001 a&b) propose that a hub structure is formed by optimising two factors related to the links in the network. Firstly, the number of links in the network is maximised and secondly, the cost of wiring i.e. the length of a new link, is minimised. When these two factors are optimised, a hubterranean network arises. In the context of the word co- occurrence network, the connectivity factor relates to the shortcuts within the network. That is, shortcuts are words that form many connections and as such maximise the number of word combinations that occur adjacently within a sentence.

An ideal candidate for the cost factor in the linguistic network is the storage cost of linguistic form. It is well known that when two elements occur frequently together, a form of chunking occurs which is very often accompanied (in a speech context) by the reduction of phonetic form as well as a loss of internal structure (Bybee 2002a). The chunk also gains autonomy and each element within the chunk primes the other (Bybee 2002a; see section 2.7.1). Thus in storage terms, the cost of storing two elements is reduced when it is stored in a chunk. Since chunking occurs when the two elements occur frequently together, then it is proposed that frequency of co-occurrence or transitional probability relates to the network cost factor.