

A Hubterranean View of Syntax

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

By

Julie Louise Steele

CAMBRIDGE
SCHOLARS

P U B L I S H I N G

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

<i>Acknowledgements</i>	x
<i>Abstract</i>	xii
<i>List of Figures</i>	xv
<i>List of Tables</i>	xxi
 Chapter One.....	1
Introducing the Hubterranean World of Linguistic Form	
1.1 Network approach	
1.2 Hub structure	
1.3 Universal properties	
1.4 Content	
 Chapter Two	15
Syntactic Constructions	
2.1 Overview	
2.2 Constructions	
2.3 A network of constructions	
2.4 Principles of organisation	
2.5 Usage-based approach	
2.6 Groupings	
2.7 Frequency of co-occurrence	
2.8 The functional-lexical distinction	
2.9 Summary	
 Chapter Three	52
Perspectives on Language Learning	
3.1 Overview	
3.2 Usage-based approach	
3.3 Distributional analysis	
3.4 Segmentation and linear sequencing	
3.5. Extracting patterns	
3.6 Verb generalisation	
3.7. Optimal exemplars	
3.8 Summary	

Chapter Four	81
Network Perspectives	
4.1 Overview	
4.2 Small worlds	
4.3 Language networks	
4.4 Networks and language learning	
4.5 Hubs and small worlds	
4.6 Summary	
Chapter Five	105
The Adult Word Co-occurrence Network	
5.1 Overview	
5.2 Data	
5.3 Data preparation	
5.4 Small world properties	
5.5 Hierarchical modelling	
5.6 Assortative versus disassortative mixing	
5.7 Hubs and authorities	
5.8 Summary	
Chapter Six	135
The Child Word Co-occurrence Network	
6.1 Overview	
6.2 Data	
6.3 Data preparation	
6.4 Small world properties	
6.5 Hierarchical modelling	
6.6 Assortative versus disassortative mixing	
6.7 Hubs and authorities	
6.8 Summary	
Chapter Seven.....	157
Motifs: Significant Link Patterns	
7.1 Overview	
7.2 Automorphic subgraphs	
7.3 Subgraph extensions	
7.4 Subgraphs in the word co-occurrence network	
7.5 Detecting motifs	
7.6 The Z-score results for 3-node subgraphs	
7.7 Significance profiles for 4-node subgraphs	
7.8 Summary	

Chapter Eight.....	186
Investigating Motifs within the Adult and the Child Network	
8.1 Overview	
8.2 Automorphic nodes	
8.3 Non-automorphic motifs	
8.4 Summary	
Chapter Nine.....	225
The Hubterranean Structure and its Syntactic Implications	
9.1 Overview	
9.2 Segmentation	
9.3 Identifying (construction-specific) syntactic categories	
9.4 The five factors constraining early verb generalisation	
9.5 The verb-construction subpart network	
9.6 Optimisation factors	
9.7 Summary	
Chapter Ten	264
Concluding Discussion	
10.1 Networks	
10.2 The optimisation process	
10.3 Contrary factors	
10.4 Development of abstract syntax	
10.5 Limitations	
10.6 Second language learning	
10.7 Incorporating natural principles	
Appendix A.1	284
Word Pair Examples with Storage Cost Values from the Adult Word Co-occurrence Network	
Appendix A.2	308
Word Pair Examples with Storage Cost Values from the Child Word Co-occurrence Network	
Bibliography	323

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

- 1.1. A network link
- 1.2. Language as a network of word co-occurrence links (Solé, Corominas-Murta, Valverde and Steels 2005: 2)
- 2.1. “The symbolic structure of a construction” (Croft 2001: 18)
- 2.2. “The internal structure of a construction (exploded diagram)” (Croft 2001: 176)
- 2.3. “Radical Construction Grammar representation of construction taxonomy” (Croft 2001: 56)
- 2.4. An example of a simplified Generative tree structure representing hierarchical phrase structure
- 2.5. The relationship between transitional probability and emergent constituent structure in naturally occurring discourse. Utterances were taken from Carterette and Jones 1974/MacWhinney 1995 corpus (Bush 2001)
- 2.6. The four networks representing the causal grammars under investigation by Tenenbaum and Niyogi (2003). G₁ and G₂ were more easily learnt than G₃ and G₄ Griffiths and Tenenbaum (2007: 14)
- 3.1. Sourced from Sethuraman and Goodman (2004a: 86) depicting the different factors influencing the extension of verbs
- 3.2. Cumulative number of different verbs in VO and SVO word-combinations produced by a child learning English, as a function of age (Ninio 1999a: Figure 2)
- 4.1. Different types of networks are formed at various values of parameter p (Watts and Strogatz 1998: 441)
- 4.2. The child network at (a) 25 months (b) 26 months and (c) 28 months (Corominas-Murtra, Valverde and Solé 2007a: 4). Scale free degree distribution is observed at 28 months (c)
- 4.3. (i) link length distribution for an $n=100$, $k=4$ network optimised network for $0 < \lambda < 1$ range (Mathias and Gopal 2001b: 6, Figure 3). (ii) the evolution hubs for an $n = 100$, $k=4$ optimised network for $0 < \lambda < 1$ range (Mathias and Gopal 2001b: 9)
- 5.1. A log-log plot of the rank v frequency of the words in Nadult
- 5.2. The first regime of the cumulative degree (K) distribution - the truncated tail. Where $P(K)$ is the cumulative distribution
- 5.3. The cumulative degree (K) distribution $P(K)$ is captured by two power law regimes (a) the second regime with exponents -1 and (b) the third regime with exponent -0.7
- 5.4. The cumulative distribution $P(S)$ of bigrams with storage cost (S) in the first regime of the adult network

- 5.5. The second regime of the power law in the adult network. The cumulative distribution $P(S)$ of the storage cost (S)
- 5.6. The third regime in the adult network. The cumulative distribution $P(S)$ of the storage cost (S)
- 5.7. The fourth regime of the power law in the adult network. The cumulative distribution $P(S)$ of the storage cost (S)
- 5.8. The fifth regime of the power law in the adult network. The cumulative distribution $P(S)$ of the storage cost (S)
- 5.9. The networks on the left show the schematic version of (a) a scale-free network (b) a modular network and (c) a hierarchical network. The networks on the right correspond to a “typical configuration” respectively (each with 256 nodes). (Ravasz, Somera, Mongru, Oltvai and Barabási 2002, Ravasz and Barabási 2003: Figure 1)
- 5.10. Log-log plot of connectivity (degree(k)) against the clustering co-efficient for the adult network. The line dissecting the squares represents the slope of -1. A slope of +0.57 is indicated
- 5.11. Log-log plot of connectivity (degree (k)) against clustering co-efficient for a semantic network “connecting two English words if they are listed as synonyms in the Merriam Webster dictionary” (Ravasz and Barabási 2003: Figure 3b) The line indicates a slope of -1
- 5.12. The degree of a word plotted against the betweenness centrality (BC) value
- 5.13. The distribution of betweenness centrality for the adult network
- 5.14. For the adult network, the amount of words which are both hubs and authorities (H&A) was plotted against the number of hubs chosen to be represented (the amount of hubs chosen equalled the number of authorities also represented: $H=A$)
- 6.1. A log-log plot of the rank (r) versus frequency (f) of the words in child network
- 6.2. The degree distribution regimes. (a) Regime three has an exponent of -0.7 whilst in (b), regime two has an exponent of -1.1. (c) The truncated tail in regime one
- 6.3. A comparison of frequency and degree for the twenty-five highest ranked words in the child network
- 6.4. The first (a) and second (b) regimes in the Sachs (1983) child language network. Both plots show the cumulative distribution $P(S)$ and the storage cost (S)
- 6.5. The third (a) and fourth (b) regimes in the Sachs (1983) child language network. Both plots show the cumulative distribution $P(S)$ and the storage cost (S)
- 6.6. The fifth regime in the Sachs (1983) child language network. The graph shows the cumulative distribution $P(S)$ and the storage cost (S)
- 6.7. Log-log plot of connectivity (degree (K)) against the clustering co-efficient $C(K)$ for the child network. The line indicates a slope of -1
- 6.8. Log-Log plot of connectivity (degree (k)) against clustering co-efficient for the “Internet at router level”, a non-hierarchical network (Ravasz and Barabási 2003: figure 4a). The line indicates a slope of -1

- 6.9. The plot of word degree against betweenness centrality (BC) for the child's network (a) compared with the adult network (b)
- 6.10. The distribution of betweenness centrality (BC) for the child network (a) and the adult network (b)
- 6.11. For the child network (a), the amount of words which are both hubs and authorities (H&A) were plotted against the number of hubs chosen to be represented (the amount of hubs chosen equalled the number of authorities also represented: $H=A$). Figure b shows the equivalent adult network plot
- 7.1. (a) An example of a subgraph 6. (b) A network in which two examples of the subgraph 6 represented by blue links
- 7.2. An illustration of subgraph 6, repeated for convenience
- 7.3. (a) An illustration of subgraph 6 and (b) its extension in which the anchor X fixes the automorphic nodes Y1, Y2, and Y3
- 7.4. A word co-occurrence network.(A) The text is converted into a word co-occurrence network (C) whose links represent precedence relations. The lighter the node's colour, the higher its degree i.e. connectivity (Solé, Corominas Murta, Valverde and Steels 2005: 2, Figure 2A&C)
- 7.5. An artificial word co-occurrence network drawn by Pajek (de Nooy, Mrvar and Batagelj 2005) from sentences 7.2 (a&b)
- 7.6. Pictured in the inset is the bi-fan (204) subgraph. An extension of subgraph 204 is shown, with anchors *a* and *the* fixing the instances of the subgraph. The input nodes (nodes that the arrows point to) are the automorphic nodes: {*berlei*, *kettle*, *pig*, *ration*, *saucepan*, *sort*}
- 7.7. (a) This small (undirected) network represents the minimum disjoint instances of subgraph 110 in the adult network (the arrows have been omitted to simplify the scheme, aiding the identification of disjoint groups). (b) The grouping below, represented by the coloured area, is one possible representation of three disjoint node groups; there are other alternatives but each one still produces at least three disjoint groups of nodes
- 7.8. Subgraphs within separate networks composed of English, French, Spanish and Japanese sentences as well as an artificial bipartite network. The X- axis numbers 1-13 correspond, respectively, to the following binary notated motifs: 6, 36, 12, 74, 14, 78, 38, 98, 108, 46, 102, 110, 238 (Milo, Itzkovitz, Kashtan, Levitt, Shen-Orr, Ayzenshtat, Sheffer and Alon 2004: 1539)
- 7.9. 3-node z-scores for the adult and child network. The X-axis numbers 1-13 correspond, respectively, to the following binary notated motifs: 6, 36, 12, 74, 14, 78, 38, 98, 108, 46, 102, 110, 238. Pictures are from the motif dictionary (Kashtan, Itzkovitz, Milo and Alon 2005c)
- 7.10. 3-node uniqueness scores normalised for the adult and child network. Equation 7.2 gives the normalisation procedure
- 7.11a The significance profile of the first forty-eight 4-node subgraphs (14 to 862. See motif dictionary Kashtan, Itzkovitz, Milo and Alon 2005)
- 7.11b The significance profile of the subgraphs 904 to 4564
- 7.11c The significance profile of the subgraphs 4566 to 5084
- 7.11d The significance profile of the subgraphs 5086 to 15326 (there were no 31710 subgraphs in either the studied or the random networks generated)

- 7.12a The normalised uniqueness profile of the first forty eight 4-node subgraphs (14 to 862) for the adult and child network
- 7.12b The normalised uniqueness profile of the subgraphs 904 to 4564 for the adult and child network
- 7.12c The normalised uniqueness profile of the subgraphs 4566 to 5084 for the adult and child network
- 7.12d The normalised uniqueness profile of the subgraphs 5086 to 31710 for the adult and child network
- 7.13 The abundance value of the most highly significant motifs represented in table 7.3 above for child (diamond) and adult (square) data. Pictures are from the motif dictionary (Kashtan, Itzkovitz, Milo and Alon 2005c)
- 7.14 The normalised uniqueness scores for the most significant 4-node motifs in the adult and child network
- 8.1. (a) Motif 6 from Mfinder motif dictionary (Kashtan, Itzkovitz, Milo and Alon 2005c). (b) In the adult network, the word *the* anchors motif 6. The motif was drawn by PAJEK (de Nooy, Mrvar and Batagelj 2005).
- 8.2. The 204 motif. The words *the* and *a*, which are traditionally classed as determiners or articles, anchor the set of words $\{N1, N2, N3, N4, N5, N6\}$, all of which are traditional nouns.
- 8.3. The 204 motif captures a distribution environment associated with nominals. The function words $\{the, a, and, of, in, for\}$ are represented by the out-degree nodes and the nominal set $\{N1, N2, N3, N4, N5, N6\}$ by the in-degree nodes. For two general categories to be posited, both sets of automorphic nodes e.g. $\{the, a, and, of, in, for\}$ and $\{N1, N2, N3, N4, N5, N6\}$, must reduce. Thus, the two syntactic categories are mapped one-to- one.
- 8.4. Motif 36 with automorphic nodes and the corresponding anchor shown (Kashtan, Itzkovitz, Milo and Alon. 2002-2005c)
- 8.5. Motif 78 pictured with the anchor node and two automorphically equivalent nodes N_1 and N_2
- 8.6. In motif 78, the nodes N_1 and N_2 are automorphically equivalent. (b) The word *and* anchors the automorphic nodes *clothed* and *fruit*. (c) The word *of* anchors the automorphic nodes (N) *one* and *some*
- 8.7. Motif 12 is pictured with two non-automorphic nodes N_a and N_b (from the motif dictionary, see Kashtan, Itzkovitz, Milo and Alon 2002-2005c). In (b) the word *and* anchors the different automorphic nodes, as does the word *of* in (c)
- 8.8. Motif 14 is pictured with two non-automorphic nodes N_a and N_b (the motif is from the mfinder dictionary, Kashtan, Itzkovitz, Milo and Alon (2002-2005c). In (b) the word *and* anchors the different automorphic nodes, as does the word *of* in (c)
- 8.9. (a) Motif 12 is pictured with two non-automorphic nodes N_a and N_b (the motif is from the Mfinder dictionary, Kashtan, Itzkovitz, Milo and Alon (2002-2005c). In (b) the word *and* anchors the non-automorphic nodes N_a and N_b which represent the words given in the table
- 8.10. (a) Motif 14 (b) The independent variable of the Covariational Conditional, the Xer the Y is represented by motif 14 where X denotes *bigger* and Y, *household*

- 8.11. (a) Subgraph 46. Word 1 and word 2 are automorphically equivalent and so may act as anchors or automorphic nodes. In (b) the anchors are *I* and *can*, with the automorphic nodes representing {*do*, *draw*, *drink*, *eat*, *get*, *go*, *jump*
- 8.12. The subgraph 4548 representing the pronominal schema 'I + Verbal₂ + it' (from the Mfinder dictionary, Kashtan, Itzkovitz, Milo and Alon 2002-2005c)
- 9.1. The distribution of function and lexical words in the list ranked by connectivity (in the adult network)
- 9.2. The distribution of function and lexical words in the list ranked by frequency (in the adult network)
- 9.3. The connectivity and frequency ranking for the fifty most highly connected words in the adult network
- 9.4. Motif 6 (left) is reduced to an isomorphic mapping (right) of the given function word and the syntactic category which represents the automorphic nodes {N₁, N₂, N₃, N₄}
- 9.5. (A) The isomorphic relation between the function word (f-word) and the syntactic category (s-cat) is represented by the pink line. (B) As the system moves towards a more automorphic state, the relation between the f-word and the s-cat becomes automorphic (pink dashed line). (C&D) The links form a 204 subgraph that straddles two levels of linguistic representation
- 9.6. Escher's 'up and down' (1947) lithograph (© Cordon Art - Baarn - the Netherlands), illustrating the representation of the same object (the buildings and the people) but from two different perspectives. In the same way that the ceiling of the eaves and the floor intercept, both the specific function word and the abstract syntactic category form part of the 204 motif.
- 9.7. The comparison of degree and frequency rankings of the most frequent verbs in the child word co-occurrence network.
- 9.8. The connectivity of the word *draw* in the child's word co-occurrence network. The picture is generated by PAJEK (de Nooy, Mrvar and Batagelj 2005).
- 9.9. The 904 motif anchored by *don't* and *it* from the child network. The dotted links represents the potential formation of a new 904.
- 9.10. An example of part of a subpart network derived from source sentences such as *Sam gave Mary a cake* (ditransitive); *Sam gave the piece of land to his son* (transfer-caused motion). Examples from Goldberg (1995: 111, Examples 20-21).
- 9.11. Sethuraman and Goodman (2004a: 86, Figure 3) showing the five factors of early verb generalisation.
- 9.12. The five factors of early verb generalisation in respect to the hub structure of a word co-occurrence and the verb-construction network with corresponding factors postulated by Sethuraman and Goodman (2004a: 86)
- 9.13. The power law distribution of the type frequency of a construction. The distribution represents the number of verbs which appear in a given construction
- 9.14. The power law distribution of the storage cost of a verb within a construction. The distribution represents the ratio of appearances of the verb within a given construction
- 10.1. The isomorphic forms (left) and the automorphic form i.e. the universal hub

(right)

- 10.2. An illustration of isomorphism (left) between a word (W) and a morpheme (M) and a 'universal hub' (right). Isomorphism is typical in isolating languages such as Chinese. The 'universal hub' word on the right is typical of polysynthetic languages such as Navajo, in which a sentence may correspond to one word
- 10.3. The two contrary sets of properties that are operating in language. Namely, contrast~cohesion (see Lindblom 1989) and connectivity ~isomorphism (see Ninio 2008)

LIST OF TABLES

- 2.1. “Examples of constructions, varying in size and complexity” Goldberg (2006: 5, Table 1.1)
- 2.2. Summary of different kinds of frequencies and their associated linguistic effect
- 3.1. “Frames that were frequent frames in at least two corpora, organized by number of corpora in which each occurred” Sourced from Mintz (2003: 102, Table 5)
- 3.2. Adapted from Goldberg, Casenhiser and Sethuraman (2004): “Table 4. 15 Mothers’ most frequent verb and number of verb types for 3 constructions” in Bates, Bretherton and Snyder (1988) and Carlson-Luden (1979) corpus (Goldberg, Casenhiser and Sethuraman 2004)
- 5.1. Small world properties of the adult language network N_{adult} compared to the equivalent random network N_{random} with 5365 nodes and 31602 arcs. Multiple lines were removed for the calculation in PAJEK (de Nooy, Mrvar and Batagelj 2005)
- 5.2. The first words of the bigrams are shown with their respective regime. The words which occur in the previous regime are bolded. The fifth regime is not present as the number of words in this regime numbered approximately 4,500 – too many to include. Power law regimes are indicated by (PL)
- 5.3. The twenty hub words are compared with words with the highest betweenness centrality and the twenty most frequent words. Bolded font identifies the words with the largest BC values and hubs (ranked twenty or above). Italicised words denote the twenty most frequent words
- 5.4. Bolded text indicates the words that are also in the ten most frequently used words in the adult text
- 5.5. The breakdown of hubs and authorities in the adult co-occurrence network. As the number of hubs chosen to be represented in the network increases (first column), so to does the number of words given as both hubs and authorities (fourth column)
- 6.1. Small world properties of the adult and child language network. N_{adult} and N_{child} compared to the equivalent random networks ($N_{random\ adult} = 5365$ nodes and 31602 arcs whilst $N_{random\ child} = 1346$ nodes and 3061 arcs). Multiple lines were removed for the calculation in PAJEK (de Nooy, Mrvar and Batagelj 2005)
- 6.2. The first words of the bigrams are shown with their respective regime. The words which occur in the previous regime are bolded. Power law regimes are indicated by the abbreviation (PL)
- 6.3. For the child network, the twenty largest hub words are compared with words

- with the highest BC value and the twenty most frequent words. A bolded word is both in the twenty most highly connected hubs and the BC list. An italicised word is in the twenty most frequent word list. Words which occur in the frequency list but not in the twenty largest hubs and BC list are underlined
- 6.4. The breakdown of hubs and authorities in the child co-occurrence network. As the number of husband authorities chosen to be represented in the network increases (first column), so too does the number of words given as both hubs and authorities (fourth column)
 - 6.5. For both the adult and the child co-occurrence network, the ten largest hubs and authorities are compared with the ten most frequent words. Bolded text indicates words are found in the ten most frequent words
 - 7.1. Z-scores of the triad subgraphs for the adult and child network. Z- scores over 2 identifying network motifs are italicised. The Z-score of subgraph 46 for the child network, which is borderline significant, is bolded
 - 7.2. Uniqueness values of the triad subgraphs for the adult and child network. Z-scores over 2 identifying network motifs are italicised - all motifs have a uniqueness value of four and over. The bolded numeral of subgraph 46 indicates a borderline 1.86 Z-score value in the child network. Pictures are from the motif dictionary (Kashtan, Itzkovitz, Milo and Alon 2005c)
 - 7.3. A comparison of the Z-score and the uniqueness value for each of the most significant motifs in the adult and child network. The bracketed Z-score indicates the motif is not found in significant numbers in the network. Pictures are from the motif dictionary (Kashtan, Itzkovitz, Milo and Alon 2005c)
 - 8.1. For the ten largest extensions, the anchor word is given as well as the category assigned to the majority of the automorphic nodes. The nominal category is bolded. The anchors *the*, *and*, *of* fix the most automorphic nodes
 - 8.2. The extensions of motif 6 that have more than ten automorphic nodes with their corresponding anchor. Degree (connectivity) and frequency rankings are given respectively. The ten largest extensions are bolded
 - 8.3. For the ten largest extensions of motif 6, the anchor word is given as well as the traditional syntactic category assigned to the majority of the automorphic nodes.
 - 8.4. For the ten largest extensions, the anchor word is given as well as the category assigned to the majority of the automorphic nodes. The nominal category is bolded
 - 8.5. The ten largest extensions of motif 6 and 36 from the adult network. The most predominant traditional syntactic category of automorphic nodes is given
 - 8.6. The largest extension of motif 36 with respective anchors with the corresponding number of automorphic nodes. The degree and frequency ranking of the anchor is also given respectively. The ten largest extensions of motif 36 are bolded
 - 8.7. For the ten largest extensions of motif 36, the anchor word is given as well as the category representing the majority of the automorphic nodes
 - 8.8. The largest extensions of motif 78. The constructions represented by the motif are given as are the respective anchors and the number and details of the automorphic nodes

- 8.9. The counts of the anchors for the largest extensions of 204 and 904 motifs as shown in Table 8.10 below
- 8.10. The largest extensions for motif 204 and 904 with respective anchors and automorphic nodes (N)
- 8.11. The counts of anchors in the largest extensions of 204 and 904 as shown in Table 8.12
- 8.12. The largest extensions for motif 204 and 904 with respective anchors and automorphic nodes (N)
- 8.13. Within a motif 14 extension, pictured in Figure 8.8, the words represented by the automorphic nodes N_a and N_c anchored by the words $\{and, of\}$ respectively
- 9.1. Pathbreaking Verbs (Ninio 1996: table 6)
- 9.2. Verbs with the highest connectivity are shown. All these verbs are anchors to the largest extensions of the two most highly significant motifs (motifs 6 and 36). Columns show how many automorphic nodes (N) each verb anchors in the respective motifs
- 9.3. The rankings for frequency (token) and degree (type frequency) for pronouns in the large BNC adult network (chapter five)

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 hyponymic 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 may 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 value² is small compared to a random network³ and “and offers a measure of a network’s overall navigability” (Barabási and Oltvai 2004: 102).

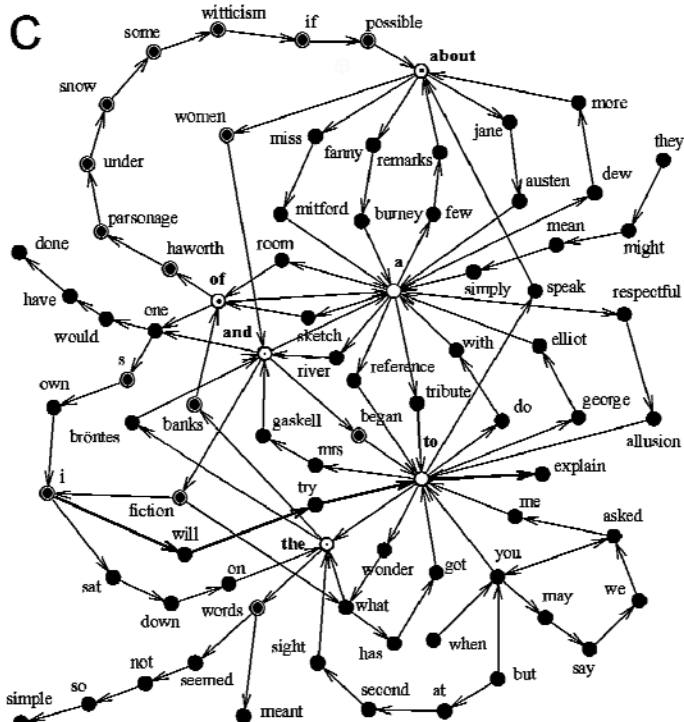
² This value is also known as the network diameter (Albert, Jeong and Barabási 1999).

³ A network in which the nodes have roughly the same number of links is known as a random network (see Erdős 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 Bröntes 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, $\langle C \rangle$, 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.