# The Genitive Ratio and its Applications

**Kevin John Glover**

A thesis submitted for the degree of Doctor of Philosophy

School of Computer Science and Electronic Engineering
University of Essex

January 2016

This thesis is dedicated to my partner, Anne, who has given years of love, support and encouragement to an ageing, serial student who probably should have known better.

Also in memory of Dr Derek Plumb, MA, PhD (Cantab.)

1947-2005

## Abstract

The genitive ratio (GR) is a novel method of classifying nouns as animate, concrete or abstract. English has two genitive (possessive) constructions: possessive-*s* (*the boy's head*) and possessive-*of* (*the head of the boy*). There is compelling evidence that preference for possessive-*s* is strongly influenced by the possessor's animacy. A corpus analysis that counts each genitive construction in three conditions (definite, indefinite and no article) confirms that occurrences of possessive-*s* decline as the animacy hierarchy progresses from animate through concrete to abstract.

A computer program (Animyser) is developed to obtain results-counts from phrase-searches of Wikipedia that provide multiple genitive ratios for any target noun. Key ratios are identified and algorithms developed, with specific applications achieving classification accuracies of over 80%. The algorithms, based on logistic regression, produce a score of relative animacy that can be applied to individual nouns or to texts. The genitive ratio is a tool with potential applications in any research domain where the relative animacy of language might be significant. Three such applications exemplify that.

Combining GR analysis with other factors might enhance established co-reference (anaphora) resolution algorithms. In sentences formed from pairings of animate with concrete or abstract nouns, the animate noun is usually salient, more likely to be the grammatical subject or thematic agent, and to co-refer with a succeeding pronoun or noun-phrase. Two experiments, online sentence production and corpus-based, demonstrate that the GR algorithm reliably predicts the salient noun. Replication of the online experiment in Italian suggests that the GR might be applied to other languages by using English as a 'bridge'.

In a mental health context, studies have indicated that Alzheimer's patients' language becomes progressively more concrete; depressed patients' language more abstract. Analysis of sample texts suggests that the GR might monitor the prognosis of both illnesses, facilitating timely clinical interventions.

## Acknowledgements

I have received a great deal of help, support and encouragement during the course of this work. A huge thank-you must go to my supervisor, Professor Massimo Poesio, for his guidance. It has been a real privilege and a pleasure to work with him.

My approach to this work was informed, particularly in the early stages, by the breadth of the courses - in psycholinguistics, pragmatics, conversation analysis, syntax and computational linguistics – that I was fortunate to undertake for my MA in Linguistic Studies at the University of Essex. I am particularly grateful to Dr Claudia Felser, Dr Rebecca Clift, Dr Doug Arnold and Professor Andrew Radford for their teaching and support. From the same department, Dr Vineeta Chand has been generous with her advice, particularly on the Alzheimer's chapter.

As a member of my academic board, Dr Sonja Eisenbeiss has been ever generous with advice and ideas, as have others whom I have never met but who have responded generously to my email requests. I think particularly of Dr Anette Rosenbach, Dr Alan Scott (University of Nottingham), Maarten van Casteren (MRC CBU Cambridge), Dr Alan Kawamoto (UC Santa Cruz), Professor Ken McRae (University of Western Ontario), Dr Richard Evans (University of Wolverhampton), Dr Martin Corley (University of Edinburgh), Professor David Crystal (Bangor University), Adam Feldman (Google), Eric Yeh (SRI International) and Tom de Smedt (University of Antwerp). Those who ignored my requests shall remain anonymous.

Although the World Wide Web has rendered almost redundant that

time-honoured phrase "to the best of my knowledge"… to the best of my knowledge, the term "genitive ratio" was original, never having been encountered in the course of an extensive literature review. It was only in June 2013 that a belated Google search for the precise phrase uncovered a Language Log (blog) posting by Professor Mark Liberman in September 2008, entitled "A correlate of animacy". During a quiet moment in the "wrap-up session" of a workshop, Professor Liberman used Google Web Search to calculate the results-count ratios of "X's Y vs. the Y of X" for 18 exemplars (of which 13 are proper nouns) in four categories: American politics (4), information technology companies (4), countries (5), and chemical elements (5). He comments that "the ratio held up in a semi-sensible way", and that, albeit on very sparse data, the categories did not overlap. Professor Liberman never developed his observations into a formal hypothesis, but tells me that he looks forward to reading my thesis.

Most of the work on this thesis was done in the university libraries of Leicester and Loughborough, where the wonderful ladies of the library cafe maintained my caffeine levels. Thank you, Alison, Sandra and Cristina.

## Contents

PART 1: FOUNDATIONS

**PART 2: WORDS**

**4 Computing the Relative Animacy of Text**

**PART 3: TEXTS**

**PART 4: CONCLUSIONS**

**LIST OF TABLES**

## LIST OF FIGURES

## SUPPLEMENTARY MATERIALS

Data files relevant to this thesis have been attached to the digital copy at the University of Essex Research Repository: http://repository.essex.ac.uk

CHAPTER 1

# Introduction and Overview

*Everything should be made as simple as possible,*
*but not simpler*
Albert Einstein

## 1.1 Animacy

The objective of the work discussed in this thesis is to develop models to compute the **animacy** of nouns and, by extension, texts. Animacy is a concept with both cognitive and linguistic dimensions. The linguistic definition of a noun's animacy (in a dictionary or an ontology) is binary, indicating whether the entity it represents either has life or it does not. The cognitive representation of an entity's animacy, on the other hand, is more fine-grained, not dichotomous but located within a **gradient** of animacy.

In this thesis we first of all show that there are a number of linguistic reasons to adopt this cognitively-motivated notion of an animacy gradient as the foundation for models to compute the animacy of nouns. Further, we show that there are good reasons to discretize this gradient into three main `sections', corresponding to nouns expressing **animate**, **concrete** or **abstract** concepts. Those three levels constitute a basic scale or **hierarchy** of **relative animacy**, progressing from animate to abstract. Finally, we present a new method of calculating the relative animacy of text, with textual analysis and discourse salience as potential applications.

Entities (usually human) that are more animate are typically more **salient** – more prominent or focused – than other referents within a **discourse** (a term that will encompass both spoken and written language). Why that matters is because salience is a factor in **co-reference** (or **anaphora**) resolution, which is a necessary process in many natural language processing (NLP) applications.

Co-reference is generally intuitive for people, but difficult for computers. Consider this text:

[1.1] Leading his horse, the outlaw strode along, followed by his dog. *He* was alert for danger from among the boulders.

There is ambiguity here. The gender-specific pronoun *he* could refer to any of the underlined animate nouns in the first sentence. Most people would spontaneously link the pronoun to *outlaw*, because it is the human referent that is salient – but why is it salient?

One view (Yamamoto, 1999: 16) is that the cognition of relative animacy is influenced by "biological distance", by our perception of the gap that separates us from other animate entities. So, we are closer to other humans, whether referenced by role (*teacher*) or particularly by name (*Anne*), than we are to a domestic pet (*cat*), than we are to an insect (*cockroach*). Fine, but natural language does not have the rationality of biology. Rosenbach (2008: 154) observes that "animacy as a linguistic factor crucially depends on whether and to what extent speakers treat referents linguistically *as if they were animate* [my emphasis]". Rosenbach's view is closer to Pinker (2012: 713), who posits a "human empathy gradient" that essentially substitutes psychological for biological distance.

Both Pinker's and Yamamoto's views are persuasive. There does appear to be a cognitive basis to our linguistic preference. From a computer science perspective, the problem is one of quantification: to be able to measure the 'distance' or 'gradient'. A computational algorithm of relative animacy requires a quantifiable measure, some kind of 'score' for each referent. How might that be achieved? By means of the **genitive ratio** (GR).

From a comprehensive review of the research evidence, Rosenbach (2014: 242) cautiously concludes that there is "some evidence for a general cognitive underpinning of the internal factors (mainly animacy) governing genitive choice". There are in English two basic genitive (possessive) constructions: possessive-'*s* (*the boy's head*) and possessive-*of* (*the head of the boy*). The empirical evidence (reviewed in chapter 3) is clear, that the animacy of the possessor (*the boy*) influences the selection of a possessive-*s* construction. By extension, an inanimate possessor (either concrete or abstract) influences the selection of a possessive-*of* construction. That observation prompts this hypothesis:

> **For any noun, the ratio of possessive-*s* constructions to possessive-*of* constructions, as quantified by a corpus analysis, should provide a proxy measure of that noun's relative animacy**.

The initial test of that hypothesis (in chapter 3) is facilitated by access to a database of genitive constructions, independently annotated for animacy. Analysis of those data reveals the significance of definiteness and number; the different GR characteristics of proper nouns and "measure nouns"; and provides verification that the genitive ratio differentiates, at a categorical level, animate, concrete and abstract nouns.

The necessary qualification is that the database analysis is a 'clean data' test. The only nouns analysed are those found in genitive constructions. To devise an algorithmic implementation of the genitive ratio concept involves identifying and resolving the exceptions and confounds that become apparent when the GR concept is tested on an unsupervised, un-annotated corpus (in chapter 4).

## 1.2 The Goal

**To devise a computational model, based on the genitive ratio, that will reliably classify nouns as animate, concrete or abstract.**

## 1.3 Structure of the Thesis

The thesis is structured in four parts:

| | | |
|---|---|---|
| Part 1 | Foundation | Chapters 1-3 |
| 2 | Words | 4-5 |
| 3 | Texts | 6-7 |
| 4 | Conclusion | 8 |

Each of the chapters 2-7 begins with an Overview and closes with "Caveats and Conclusions" – a heading borrowed from Jerry Fodor (*The Modularity of Mind*, 1983).

**Chapter 2** provides a broad foundation of theory and supportive research. Relevant literature is otherwise critically reviewed within the chapters to which it relates. A survey of empirical and theoretical work in four domains underlines the significance of the animate-concrete-abstract model. The chapter reviews how relative categories of animacy have been variously rated, codified and classified.

**Chapter 3** develops and tests the genitive ratio, from its theoretical rationale, through an analysis of a database of genitive constructions, to verification that the progression of the ratio at a categorical level matches the animate-concrete-abstract hierarchy of animacy.

**Chapter 4** presents a developmental process that tests different components of a workable genitive ratio. Two computational algorithms are defined, both reliant upon automated phrase-searches of Wikipedia: an animateness rating (AR) and a concreteness rating (CR).

**Chapter 5** tests the AR as a predictor of relative salience, in two experiments: online sentence production and a corpus-based analysis. Replication of the online experiment in Italian tests whether the genitive ratio method might be applied to other languages by using English as a 'bridge'.

**Chapter 6** applies the concreteness rating (CR) to tracking the language of patients diagnosed with Alzheimer's disease. Their language becomes progressively more concrete as the disease develops.

**Chapter 7** applies the CR to tracking the language of people with depression. Their language tends to become more abstract as their depression deepens.

**Chapter 8** reflects upon the main themes that emerge from the research findings, and evaluates this project's contribution to knowledge. Other potential applications and improvements are suggested.

## 1.4 Postscript

Three things will be repeatedly stressed throughout this thesis. First, the genitive ratio is a relative rather than a categorical method of noun classification. Second, although its level of success is significantly above chance, it would be most effectively applied in combination with other factors. Third, the three areas of application explored in chapters 5-7 are presented primarily as proofs of concept, as evidence that the genitive ratio is viable in the real world. Particularly with regard to possible applications in mental health, there are no pretensions to any clinical insights or expertise.

All examples and empirical materials are in British English, unless otherwise stated.

CHAPTER 2

# Animacy: From Animate to Abstract

*How do words ... catch hold of things?*

Michael Frayn: *The Human Touch* (2006)

## 2.0 Overview

This chapter develops a perspective that draws upon the literature concerning the categorisation of animacy. The chapter relates the evidence of prior studies to these questions:

- How should we define the categories *animate, concrete* and *abstract*?
- Why and how is the distinction of *animate* from *concrete,* and *concrete* from *abstract*, of benefit to researchers in a range of fields, from language development to neuroscience?
- Are there precedents for the development of computational models in the classification of animacy?
- What are the strengths and weaknesses of classification systems based on participant ratings?
- How have categories been deconstructed and coded in the annotation of corpora?
- What can be learnt from the hierarchies, or scales, of *animacy* (that is, animacy in its broadest sense, from animate to abstract) that have been proposed?

The chapter seeks to lay down a foundation for what follows. Terms are explained and defined. Alternative measures of animacy are critically examined. Research studies that rely upon classifications of animacy are reviewed. The scope of the topics covered testifies to the range of potential applications of the genitive ratio.

## 2.1 Animate, Concrete or Abstract?

Our perceptions of what is animate and inanimate are not based solely on objective, rational criteria. They are a product of "anthropocentric human cognition" (Yamamoto, 1999: 9) that introspectively holds humans to be more animate than all other living creatures and, possibly, oneself to be the most animate of all. Medieval naturalists constructed a *scala naturae* (Patton, 2008), based on the Aristotelian "great chain of being", in which every living thing has its proper place. This linear scale ranks humans as the highest of earthly beings, with "base creatures" such as worms and insects at the other extreme of the hierarchy.

This is not as far removed from the modern world as it might seem. Whilst most languages treat inanimates as an "undifferentiated class" (Comrie, 1981: 190), Navajo (for example) has a hierarchy of <u>in</u>animacy in which natural phenomena such as wind, rain and lightning are ranked above other inanimates. From the perspective of linguistic typology, Cruse (2006: 13) observes, in his definition of *animacy*, that "An examination of a wide range of languages suggests that there is a universal 'scale of animacy', and that different languages draw their distinction between animate and inanimate at different points on the scale. Underlying the scale is something like perceived potency, importance, or ability to act on other things, rather than a simple possession or non-possession of life".

Before proceeding to definitions, it is the scope of the three categories of nouns under examination – animate, concrete and abstract – that needs to be clarified:

- Unless stated otherwise, scales or hierarchies of *animacy* encompass both animate and inanimate, and sometimes also abstract, entities. Reference to a scale of *relative animacy* might be preferable, but has not been found in prior studies.
- *Concrete* has two scopes. In a binary distinction between *concrete* and *abstract*, *concrete* encompasses all entities, animate as well as inanimate, that are not *abstract*. When presented as an intermediate category between *animate* and *abstract*, *concrete* encompasses all entities that are neither sentient nor abstract.

In this chapter, and in the thesis as a whole, the general rule will be that **the adjective *animate* and the noun *animateness* refer to that specific category, whilst the noun *animacy* alludes to the full scale of reference from *animate* to *abstract***, unless there is a specific statement to the contrary.

The three categories – animate, concrete, abstract - are not discrete. They are rendered porous by the polysemy of the English language. A common example is the noun *chair*, which might be concrete (an item of furniture) or animate (the person responsible for the process of a meeting). Connell and Lynott (2012: 461) cite *substitute* and *kingdom* as examples of words that occupy a polysemous middle ground somewhere between concrete and abstract. Whilst not invalidating the utility of the genitive ratio, polysemy and other linguistic confounds ensure that it can never be absolute. The genitive ratio model is probabilistic, and probabilistic means uncertain.

In the Stanford Encyclopedia of Philosophy, Rosen (2014) concludes that "The abstract/concrete distinction ... is of fundamental importance. And yet there is no standard account of how it should be drawn". Hale (1988) makes an attempt, regarding concrete entities as those that occupy dimensions of "spacetime" that do not apply to abstract entities. On a more pragmatic level, "concreteness" is defined by  Brysbaert, Warriner and Kuperman (2014: 904) as "the degree to which the concept denoted by a word refers to a perceptible entity". This is reflected in the instructions given to their participants in a large-scale rating study *(ibid*: 906):

> "A concrete word … refers to something that exists in reality; you can have immediate experience of it through your senses (smelling, tasting, touching, hearing, seeing) and the actions you do …
>
> "An abstract word … refers to something you cannot experience directly through your senses or actions. Its meaning depends on language. The easiest way to explain it is by using other words".

These are descriptive definitions. An alternative approach is to classify words in terms of their linguistic (rather than just their semantic) characteristics. Gillie (1957: 216) identified abstract nouns by the presence of seven suffixes: *-ness*, *-ment*, *-ship*, *-dom*, *-nce*, *-ion*, and *–y*. This is clearly inadequate: what about *love*, *hate*, *risk*, *danger*, or *pence, lion*, *money*? Orăsan and Evans (2007: 80) would categorise a noun phrase (NP) as animate where "its [singular] referent can also be referred to using one of the pronouns *he, she, him, her, his, hers, himself, herself*, or a combination of such pronouns (e.g. *his/her*)".

The semantic diversity of a word correlates strongly with its frequency since, the greater the number of contexts in which a word might possibly feature, the more frequently that word will tend to be used. Hoffman, Lambon Ralph and Rogers (2013: 722-723) cite a consensus conclusion from prior studies that it is abstract words that have "inherently more variable and context-dependent meanings than do concrete or highly imageable words" (*ibid*: 722). Note the inference that abstract words are not "highly imageable", and the earlier inference from the instructions given by Brysbaert *et al* (2014: 906), that an abstract entity does not "exist in reality".

Reality is a psychological construct – what is real to you might not be to me: "Reality is constructed in the mind of the participants or researcher, or both" (Strang, 2015: 25). This becomes significant when researchers select their materials for experiments that explore the differences between concrete and abstract. Kousta *et al* (2011) applied especially stringent criteria to the preparation of materials for their lexical decision experiments. They constructed a set of 38 concrete-abstract word-pairs, matched on no less than 12 psycholinguistic and lexical criteria, including familiarity, age of acquisition, length (in phonemes, letters and syllables), range of meanings (based on WordNet synsets), measures of frequency, and context availability ratings obtained from participants. Six of these matched word-pairs are reproduced in Table 2.1.

I would argue that the supposedly abstract entities *demon*, *hell*, *angel*, *concert*, *paradise* and *ghost* are highly imageable, in the sense that it is difficult for us to read any of these words without mentally forming a visual image. Although we cannot 'see' an *angel*, we have a clear, stereotypical image in our minds of what an angel looks like, a picture formed by exposure to numerous paintings, sculptures and primary school nativity plays. Particularly when words carry a religious significance, they are "real" to significant numbers of people.

**Table 2.1**: Matched pairs of concrete and abstract nouns (from Kousta *et al*, 2011)

| **Concrete** | **Abstract** | **Concrete** | **Abstract** |
|---|---|---|---|
| creature | concert | asbestos | paradise |
| relic | demon | lamp | hell |
| cousin | angel | stick | ghost |

Yamamoto (1999: chapter 1) draws together the pragmatic and semantic evidence for animacy as a quality that has gradience, and that is based not on objective criteria but on more subjective perceptions. Our cognitive "mental models" (see section 5.3) of entities contain not just physical attributes such as motion, but also cognitive factors which bestow an "inferred animacy", factors such as a degree of empathy (with anthropomorphised animals such as horses, dogs and cats, for example) or of innate sentience (e.g. dogs vs. 'intelligent' computer software vs. amoebae). Rosenbach (2006: 106) argues that "animacy as a linguistic factor is dependent on how language users *conceptualize* referents as being more or less close to their own species".

We might define an animate entity, in simple terms for the purpose of this thesis, as one that is sentient, i.e. endowed with senses, feelings and awareness, but we need to bear in mind that, as Yamamoto (1999: 1) points out, 'animate-ness' is more complicated than merely the semantic feature [± alive]: there is a "cognitive domain" of animacy, that contains such concepts as empathy, locomotion and sentience.

The Harvard Mind Survey (Gray, Gray and Wegner, 2007) measured not animacy *per se* but "mind perception" – our perceptions of how "minded" various entities are. For "minded" read "sentient"; for "sentient" read "animate". Using the method of pairwise comparisons, the Harvard survey collected data on 18 "mental capacities" attributed by participants to 13 prototypical "characters". These included a chimpanzee, a dog, a foetus, a dead woman, God, a robot, and a man in a persistent vegetative state (PVS).

From 2,399 completed surveys, the researchers also collected data about their participants' spiritual and political beliefs, as well as standard demographics. Principal components factor analysis (varimax rotation) identified two factors that together accounted for 96% of the variance:

- Experience (i.e. the extent to which the characters are perceived to have feelings) at 88% included the mental capacities of hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment and joy.
- Agency (i.e. the extent to which the characters are perceived to have the capacity to act independently) at 8% included the capacities of self-control, morality, memory, emotion recognition, planning, communication and thought.

Figure 2.1 shows the distribution of the 13 "characters" on these two dimensions.

Two findings from the Harvard Mind Survey are relevant to the current thesis. First, individual differences, particularly in participants' spiritual beliefs, affected their judgments. This supports the argument that "reality" is a psychological construct. Second, the factor scores on the dimensions of experience and agency indicate a hierarchy that is similar to the animacy hierarchies reported in this chapter.

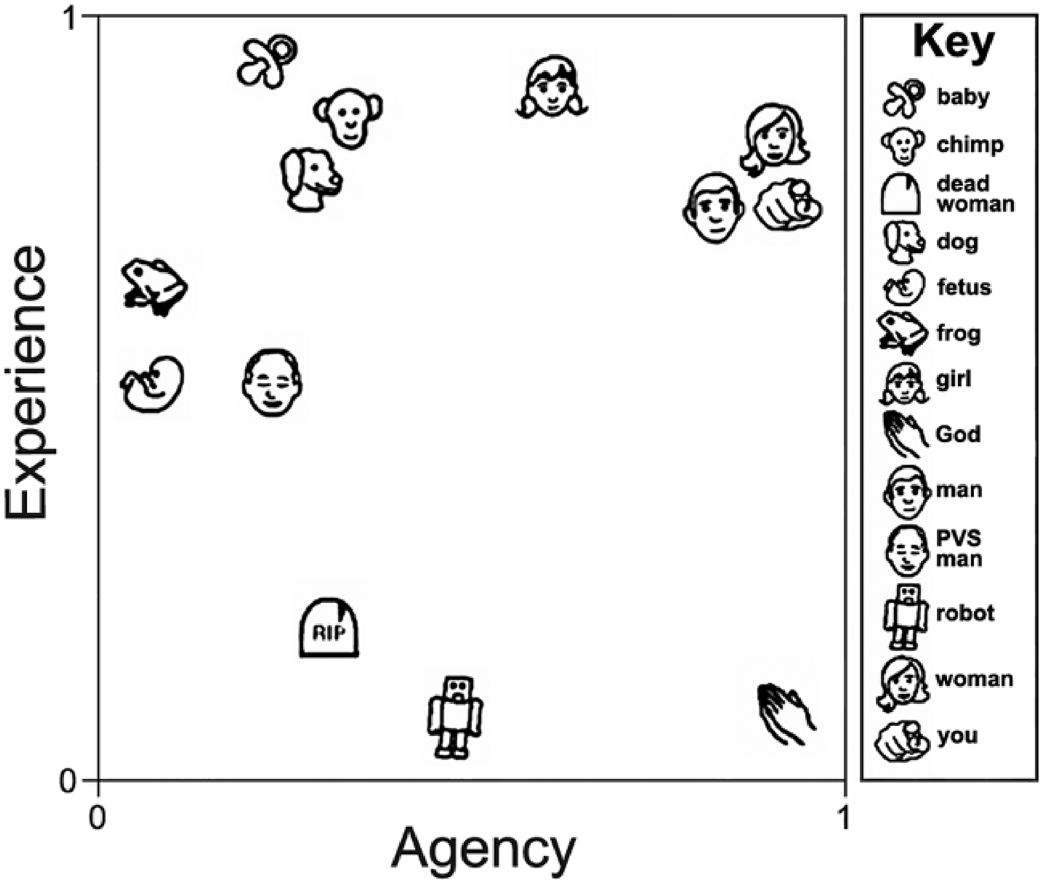


Figure 2.1: Adjusted character factor scores on the dimensions of mind perception. (Reproduced from *Science*, volume 315, 2 February 2007).

Researchers in the field of experimental philosophy (e.g. Knobe, 2008) have used the methods of cognitive science to probe people's intuitions about linguistically abstract entities that might yet be perceived as exhibiting some form of consciousness – such as an organisation, particularly a business such as Microsoft or Google whose existence is based on intellectual property; or a robot; or God. They have found that respondents are willing to assign certain cognitive attributes to these essentially abstract entities – certain beliefs and intentions in the case of a business, say – but they do not endow them with the more individually subjective attributes such as the experience of envy or happiness. In the case of

God, it is self-evident that people's intuitions will be influenced to a considerable degree by the extent of their personal religious beliefs.

### 2.2 The Many Uses of Relative Animacy

"The importance of concreteness for psycholinguistic and memory research is hard to overestimate" (Brysbaert, Warriner and Kuperman, 2014).
"Only if we can have a way to measure concreteness for words and senses consistent with human judgment in large scale will it be possible for us to pursue further studies on its role in natural language processing" (Kwong, 2013: 1150).

The significance of the abstract/concrete distinction, particularly in psycholinguistic and neurolinguistic research, is exemplified by the thirteen studies cited by Brysbaert, Warriner and Kuperman (2014: 904), all published in 2011-2013 and spanning multiple research domains that include clinical neuropsychology, long-term and working memory, and bilingual word processing (see Brysbaert *et al* for  references). The following sections will review four different domains of research where the differentiation of animate, concrete and abstract words has made a vital contribution.

The accounts of the psycholinguistic and neurolinguistic studies offer an explanation of why there is a progressive bias to concrete language in the development of Alzheimer's disease (see chapter 6). There is a common finding that, compared with abstract words, concrete words are stronger in sensory perception, particularly in imageability; richer (with more features) in their semantic representations; and hence activate a wider neural network that is more resistant to the depredation of dementia. The difference in feature-richness is clear from Table 2.2 (adapted from Clark and Begun, 1971).

**Table 2.2:** Features of six types of noun (Clark and Begun, 1971)

| | **Features** | | | |
|---|---|---|---|---|
| Noun Type | Human | Animate | Concrete | Count |
| (1) Human | + | + | + | + |
| (2) Animal | - | + | + | + |
| (3) Concrete-count | - | - | + | + |
| (4) Concrete-mass | - | - | + | - |
| (5) Abstract-count | - | - | - | + |
| (6) Abstract-mass | - | - | - | - |

### 2.3 Psycholinguistics and Animacy

There is extensive empirical evidence, from psychology, psycholinguistics and clinical neuropsychology (see Kroll and Merves, 1986, for references), of a differentiation between concrete and abstract words in the cognitive processes of comprehension and memory. Psycholinguistic experiments based on naming, word recognition and lexical decision, both written and spoken, have consistently demonstrated that recognition of concrete words is faster than recognition of abstract words.

Moss and Gaskell (1999: 73) explain this phenomenon in terms of a "richer semantic representation for concrete words". This is a common factor in three competing models of word-meaning – the dual code theory of Paivio (1971, 1986 and 2007); the context availability hypothesis of Schwanenflugel (1991);

and the connectionist theory of Plaut and Shallice (1993). All three models assume that concrete representations are richer than abstract representations.

The dual code hypothesis (see Paivio, 1971: 179-181) essentially puts imageability at the core of the difference. Both concrete and abstract words are coded verbally, but only concrete words are coded "imaginally", and that gives them their cognitive advantage.

Context availability (Schwanenflugel, Akin and Luh, 1992) is a measure of how easy it is to think of a context or circumstance that depicts a particular word. The hypothesis advanced by Schwanenflugel *et al* is that it should be easier (and therefore faster) to think of a context for a concrete word such as *book* (think of a library) than for an abstract word such as *knowledge* (think of … a library). Concrete entities thus have a processing advantage, because they typically access a more comprehensive network of semantic associations.

Plaut and Shallice (1993) developed a connectionist model that they claimed could account for dyslexia: surface, phonological, but also *deep* dyslexia. This latter type is a multiple-deficit dyslexia that exhibits semantic errors such as ‘screwpower’ for ‘ship’. Visual word recognition is intact, but semantic processing produces a "semantic neighbour" (the screw, or propeller, that powers the ship). Neuropsychological case-studies of deep dyslexics have reported significantly fewer errors in reading aloud concrete words than abstract words. According to Plaut and Shallice, the higher error rate exhibited by abstract words was caused by their having fewer semantic features. In their model’s simulations there were more visual than semantic errors on abstract words, but more semantic than visual errors on concrete words. The distinction between concrete and abstract words is therefore framed by the relative numbers of semantic features typical of each.

Kousta *et al* (2011) have cast doubt upon these models. They present evidence of contrary findings from lexical decision experiments that, if there is a rigorous matching of materials that eliminates potentially confounding linguistic variables, there is a processing advantage for abstract over concrete words. Their argument, based on a combination of experiments and regression analyses, is that neither the dual-coding theory (Paivio, 2007) nor the context availability hypothesis (Schwanenflugel, 1991) can "exhaustively" account for the empirical findings across a wide range of psycholinguistic and neuro-linguistic (EEG and fMRI) experiments. Kousta *et al* present an alternative hypothesis of the semantic representation of concrete and abstract concepts. They propose that, whilst both concepts "bind" linguistic information with sensory, motor and affective information, it is the weighting of the experiential information that determines the distinction. Concrete concepts have a higher weighting of sensory and motor information; abstract concepts are weighted towards affective information.

## 2.4 Neural Representations of Animacy

From fMRI studies of neural representations, Anderson, Murphy and Poesio (2014: 677) have provided evidence that there is a clear neural (and between-participants) distinction in the encoding of different concrete taxonomic categories (*Tools*, *Locations*, and to a lesser extent *Social Roles*) in the brain, although they found no such differentiation in the encoding of abstract categories.

In the field of cognitive neuropsychology, Warrington and Shallice (1984) provided, in a much-cited paper, evidence of a category-specific semantic deficit following damage to the brains of four patients who had contracted herpes simplex encephalitis. Supporting evidence for a dissociation between living (animate) and non-living (inanimate) entities has been put forward in many subsequent studies, although some studies have failed to control for 'nuisance variables', in that they did not match materials for factors such as frequency and familiarity (see Caramazza and Shelton, 1998: 1-4 for references and discussion).

Although the living/non-living distinction has been most apparent in clinical studies, it is important to stress that the impairments are relative, and that it is far from being an absolute distinction. Consider the patients originally studied by Warrington and Shallice (1984). Two patients were impaired in their comprehension of words for foods, as well as for plants and animals. One patient could name human body parts but not musical instruments, cloths or metals. Warrington and colleagues (see also Warrington and McCarthy, 1987) have interpreted these findings in terms of a "sensory/functional" theory of category-specific deficits. They contend that what defines the apparently categorical impairment is selective damage to the patient's sensory semantic sub-systems, and that this restricts the patient's ability to respond to the very different sensory attributes of animate entities (primarily visual) and inanimate objects (primarily functional). The reported findings that living entities are more often impaired are thus explained by objects being supposedly easier to process, since they have fewer sensory attributes.

Caramazza and Shelton (1998) argue that the diversity and dissociation of categories that are impaired in individual patients is problematic for the sensory/functional account of Warrington and colleagues. Their alternative theory, based on evolutionary principles, is the *domain-specific knowledge hypothesis* (*ibid*: 9), which proposes that conceptual knowledge is organised categorically in the brain. In support of this hypothesis, they cite findings from developmental studies of children who, from as early as three months of age, can distinguish animate from inanimate entities (Bertentahl, 1993), and also findings from their own case-study of patient EW. She (EW) presented with a disproportionate impairment in both naming and recognising (either visually or audibly) the specific semantic category of *animals*, compared both to artefacts and to other living things (there were no impairments for *body parts*, *fruits* or *vegetables*, for example). This disproportionate impairment extended to poor comprehension of statements about the attributes of animate entities – evidence that the deficit is semantic and conceptual, rather than based on visual or lexical processing problems.

Caramazza and Shelton (1998: 19-21) speculate that there is an evolutionary basis for positing the existence of "specialized processes" and "dedicated neural circuits" in respect of plants and animals, given their importance for survival (food and medicine, fight or flight): "The evolutionary adaptations for recognizing animals and plant life would provide the skeletal neural structures around which to organize the rich perceptual, conceptual, and linguistic knowledge modern humans have of these categories" (*ibid*: 20).

### 2.5 Animacy and Language Development

In an empirical investigation involving two tasks (sentence acceptability and causal explanations), Tunmer (1985) has demonstrated that four- and five-year-old children acquired the cognitive distinction between animate and inanimate entities before they acquired the more specific distinction between sentient and non-sentient entities.

Cognitive development studies by Gelman and colleagues have demonstrated that children as young as three years old are able to categorise unfamiliar objects as animate or inanimate (Gelman, 1990), and they argue that the defining feature of animacy at that age is that animate entities have a natural capacity for self-generated motion.

In an earlier review of developmental research into children's acquisition of the animate/inanimate distinction, Gelman and Spelke (1982: 44-47) provided a more fine-grained analysis of that distinction, under four headings as summarised here:

**Collections of properties**. Animate entities are characterised by a capacity for action and by an ability to grow and to reproduce. They derive sustenance for themselves and for their offspring. They change over time, and that change is generated internally rather than by external forces. They are capable of perception, emotion, learning, thought and knowledge, and they develop the structures – limbs, brains, neurosystems – that support those capabilities.

**Objects of perception**. Animate entities are perceived not simply in terms of their physical characteristics, but also in psychological terms of actions, feelings, motives and intentions. There is also the possibility of communication between the perceiver and the perceived animate entity.

**Recipients of action**. Animate entities respond to actions independently, with reactions, and not always in predictable ways. Communication is a means of co-ordinating the actions of one animate entity with another. Reciprocal interactions, role-reversals (for example, turn-taking in a conversation), co-operation and competition are all factors of animacy.

**Domains of systematic knowledge**. Inanimate objects may be understood in terms of physical laws, but an understanding of animate objects relies equally upon psychological and social organisation.

Note that this analysis equates animacy specifically with people and (at most) 'higher order' animals. It does not embrace a *collective human* category, and Gelman and Spelke (1982: 44) specifically exclude the "classification of ambiguous cases ... such as viruses and chess-playing computers".

## 2.6 Animacy and Agency

Dahl (2008: 142) has drawn attention to the correspondence between animacy and syntactic roles, with statistical evidence from corpus analysis. Transitive subjects are predominantly animate, direct objects are predominantly inanimate, and this distinction is even more evident in spoken than in written English. Noting the interactive tendencies of animacy, "yielding bundles of syntactic, semantic and pragmatic properties that tend to occur together", Dahl tentatively proposes the two "incomplete" bundles of contrasting linguistic characteristics in Table 2.3.

**Table 2.3**: Linguistic characteristics associated with animacy (from Dahl, 2008: 142)

| **Animate** | **Inanimate** |
|---|---|
| Definite | Indefinite |
| Pronominal | Lexical |
| Subject | Non-subject |
| Count | Mass |
| Proper | Common |
| Rigid designation | Non-rigid designation |
| Independent reference | Dependent reference |
| Proximate (salient third person) | Obviative (non-salient third person) |
| Agent | Non-agent |

Although animacy is closely associated with agency, there is a clear distinction between them. Animacy is defined by the semantic features and ontology of an entity. Agency is defined by its action role as the argument of a particular verb, by its propensity for intentional action, for 'doing'. Yamamoto (2006: 41) draws this distinction between animacy and agency: "Whereas 'animacy' is concerned with the intrinsic features and ontological status of animate and inanimate entities themselves, the notion of 'agency' characterises the entities (at least partially) according to what they are 'doing' ... in a nutshell, the former is largely a matter of noun phrases, whereas the latter is concerned with verb phrases". Animacy is grounded in the noun phrase (NP); agency is grounded in the verb phrase (VP).

Where a verb has both animate and inanimate arguments, it is most probable that the subject NP will be animate and the object NP will be inanimate. However, the ultimate determinant resides in the semantics of the verb. Thus, *please* might take an animate object, whereas *like* will always take an animate subject:

[2.1] The dinner pleased the man.
The man liked the dinner.

Dixon (1979: 85) identifies the concept of "agent propensity", whereby specific verbs typically select their agents from differing spans of the animacy hierarchy. Verbs such as *calculate* and *lend* will have human agents. *Listen*, *choose*, and *decide* might feasibly apply to higher animals as well as to humans. *Eat* and *die* could apply to any animate being.

Dowty (1991: 578) has proposed a hierarchy of "proto-agentivity", in which thematic roles are ranked in relation to their agentive potential:

Agent < Instrument/Experiencer < Patient < Source/Goal

## 2.7 Computational Models (Orăsan and Evans)

This section reviews the work on the computational identification of animate entities by Richard Evans and Constantin Orăsan of the University of Wolverhampton. This review is of interest on three levels. First, their approach provides a comparative model for the computation of animacy. Second, it demonstrates the incremental nature of systems development, progressing through three stages (2000, 2001 and 2007, though none since). Third, it exemplifies the dependence of prior systems on WordNet.

Several prior studies (see Poesio *et al*, 2004, for references) have commented on the limitations of WordNet, limitations inherent in both its scope and its design, but WordNet has undoubtedly been a beneficial resource for natural language processing, particularly in word sense disambiguation (see Navigli, 2009, for a survey of that field). The WordNet database is a lexical hierarchy (Fellbaum, 1998). Within each lexical category (nouns, verbs, adjectives and adverbs) there are a number of top-level "semantic domains", each with a top-level root word or "unique beginner" (there are 25 root words for nouns). Each semantic domain is itself a hierarchy of hypernyms and hyponyms, organised around sets of synonyms ("synsets"). For example, the nouns *salmon* and *cod* are hyponyms (subordinates) of the hypernym (superordinate) noun *fish*. WordNet locates the unique beginner {*animal*, *fauna*} in a separate lexical domain from {*person*, *human being*}, on the basis that they adopt reasonably distinct "possible adjective-noun combinations" (Miller, 1998: 29-30). Both, though, are located within the hypernyms {*entity*} ~⟶ {*organism*}, along with {*plant*}.

Evans and Orăsan (2000) put forward a computational method to assess the animacy of an entity that is not gender-marked: role-descriptions such as *lawyer* and *machinist* are gender-neutral. Sociolinguistic trends are tending to reduce the gender-marking of roles, so that *fireman* has become *fire-fighter,* and *chairman* has become *chairperson* or simply *chair* (Simpson and Mayr, 2010: 16-17).

Their algorithm for "animate entity recognition" is based on WordNet synsets. Specifically, they identify three noun and four verb hierarchies as strongly indicative of animacy. The three animate noun hierarchies, with their

WordNet file numbers, are *animal* (05), *person* (18), and *relation* (24). The four animate verb hierarchies are *cognition* (31), *communication* (32), *emotion* (37), and *social* (41). The algorithm first deploys a parser, to identify the lemma (i.e. the canonical form) of the head-noun of every NP in a text, and the lemma of every subject-NP's verb. For every lemma, the algorithm then extracts from WordNet a count of the number of senses within the animate noun and verb hierarchies, and the number of senses in other (inanimate) hierarchies.

Their basic method is ratio analysis. By dividing the animate sense-count by the total (animate plus inanimate) sense-count, the algorithm calculates ratios for noun animacy (NA) and verb subject animacy (VSA). By dividing the inanimate sense-count by the total sense-count, the algorithm calculates ratios for noun non-animacy (NN) and verb subject non-animacy (VSN). The computation of an animacy judgment is then obtained from these ratios, combined with a set of heuristics and three animacy thresholds derived from a "relatively small number of discrete experiments" which are not specified. The authors themselves suggest that these threshold values might better be obtained by combining a large corpus with neural network or genetic algorithm techniques. They report precision of 77% for their method.

Orăsan and Evans (2001) next presented a method of animacy identification, still using WordNet and machine learning, that improved on their earlier efforts (Evans and Orăsan, 2000). The new method is based on an animacy classification of synsets in WordNet. They use the SEMCOR corpus (Landes *et al*, 1998), which has been annotated with WordNet senses, in conjunction with WordNet itself.

Their system's utilisation of WordNet is bottom-up, starting with the terminal-node hyponyms in each hierarchy. These hyponyms are classified as animate or inanimate, based on their frequencies in the annotated corpus. The more general sense-nodes within the WordNet hierarchy are then classified as animate or inanimate, either on the basis that all the hyponyms are thus classified, or on the basis of a chi-square test (expected vs. observed values) of all-animate or all-inanimate hypotheses for that node – whichever passes the test at a significance level of 0.05. If the tests are inconclusive, so is the node. An additional level of processing using the TiMBL machine-learning program classifies these nouns for animacy, based on other data: WordNet hypernyms, verb-biases, and co-referent singular pronouns. The combined methods achieved a creditable accuracy (ratio of correct to total classifications) of 97% in a corpus test.

Orăsan and Evans (2007) obtained similar results from two methods of animacy identification: one rule-based, and one based on machine learning with access to a WordNet database that was enhanced by assigning animacy information to synsets. Based on intrinsic and extrinsic evaluations, Orăsan and Evans judged that it was the machine learning method that gave the best results when applied to anaphora resolution.

## 2.8 Ratings of Animacy

The dominant paradigm for differentiating abstract and concrete words has been the rating experiment. Participants are given a set of instructions and a list of words. They then make their judgments as to differences of degree, usually along a Likert scale from 0 or 1 to between 5 and 8 points.

Stöber and Borkovec (2002: 92) provide a case-study of this approach. Their experiment relies on ratings of concreteness by two "trained" graduate students, using a scale from 1 (*abstract*) to 5 (*concrete*), with *abstract* defined as "indistinct, cross-situational, equivocal, unclear, aggregated", and *concrete* as "distinct, situationally specific, unequivocal, clear, singular". Their employment of just two raters points to the costs of a large-scale rating exercise, and makes no provision for resolving a significant difference of opinion. The need for training acknowledges the problems of subjectivity and consistency, whilst the given definitions of *abstract* and *concrete* illustrate the difficulty of framing clear instructions for participants. The graduate status of their raters points to the fact that raters are generally drawn from a narrow population of student volunteers (see Foot and Sanford, 2004, for a discussion of bias in student participant populations).

Kwong (2013: 1150) has identified two further problems with the rater methodology. Realistically, it can only ever address a subset of our total vocabulary. The rating of 40,000 words by Brysbaert, Warriner and Kuperman (2014) is by far the most ambitious to date, but still falls well short of the average person's vocabulary. Furthermore, rating has no objective solution for the problem of polysemy. For example, many location words occupy a semantic position on the border between concrete and abstract – words such as *home*, which might refer to a building:

[2.2] It is a <u>home</u> for the elderly [concrete]

or to a concept:

[2.3] <u>Home</u> is where the heart is [abstract]

Or consider the word *border* itself:

[2.4] The border is guarded and mined [concrete]
The border between interest and obsession [abstract]

As an alternative to rating by human participants, Kwong (2013) investigated the surface analysis of dictionary definitions as a means of estimating degrees of concreteness and abstractness, but found only a "mild" correlation with prior classifications.

An account of the preparation of Italian-language lexical materials for an experiment by Anderson, Murphy and Poesio (2014: 661) further illustrates the problems involved in a process of norming that is based on participants' judgments. Their stimulus words encompassed two domains (*Music* and *Law*) across seven taxonomic categories, rated on a seven-point scale, from highly abstract (1) to highly concrete (7). Most consistently rated as concrete was the category *Tool*, followed by *Location*. Within the other five taxonomic categories (*Social Role*, *Event*, *Communication*, *Attribute* and *Urabstract*), 20 (out of 50) words had participant ratings within a range from highly concrete to highly abstract. Anderson *et al* (2014: 679) conclude that their findings "raise doubts about the value of rating concepts on a concrete-abstract continuum".

For psychologists and psycholinguists, the principal source of concreteness and imageability ratings, and the benchmark for alternative rating systems, has been the MRC Psycholinguistic Database (Coltheart, 1981). The MRC database now contains 4,292 words rated for concreteness, and imageability ratings for 8,900 words (Brysbaert *et al*, 2014). As a source of experimental materials, the MRC database has well over a thousand citations, but it is not without critics.

The adequacy of the ratings of concreteness and imageability in the MRC Database has been challenged by Connell and Lynott (2012). Those ratings consist of a list of 925 nouns compiled by Paivio, Yuille and Madigan (1968), with normative values for concreteness, imagery [*sic*] and meaningfulness. Connell and Lynott's (2012) retrospective analysis of the instructions that were given to the original raters goes some way towards explaining the high correlation between the ratings of concreteness and imageability, a factor that has been particularly influential on the dual-coding theory (Paivio, e.g. 2007). Those original raters were directed to give a *high imagery* rating to "any word which … arouses a mental image (i.e. a mental picture, or sound, or other sensory experience)". Although this instruction was trying to be multi-modal, the very word "image" strongly biases the visual modality. It was, in any event, asking a lot from the raters, to integrate a multi-sensory experience into a single composite rating.

Connell and Lynott (2012) experimented with an enhancement to the standard rating method. For each word presented, they asked their participants to rate all five perceptual modalities in turn: auditory (hearing), gustatory (taste), haptic (touch), olfactory (smell), and visual (sight). They found that it was the visual modality that was most highly correlated with prior concreteness and imageability ratings for both concrete and abstract words. Every word in their data set scored above zero for visual perception. Even *atom* scored 1.38 on a 0-5 scale. They conclude that "so-called concreteness effects in lexical decision and naming are better predicted by perceptual strength ratings than by concreteness or imageability ratings" (*ibid*: 460).

In the context of studying the "cognitive cost" of language switching in German-English bilingual speakers, von Studnitz and Green (2002) tested their participants on a semantic categorisation task: when presented with a stimulus word, categorise it as either animate ("living") or inanimate ("non-living"). This necessitated a prior animacy rating study, in which a different set of participants rated each stimulus word on an eight-point scale of animacy. The average ratings were 7.06 for animate and 1.63 for inanimate items.

Von Studnitz and Green (2002) derived their materials, in eight animate and eight inanimate categories, from Battig and Montague (1969). These materials and their ratings (provided by Professor David Green, personal communication) demonstrate the limitations of such a rating exercise. The high ratings given to the categories *fruits*, *vegetables*, *flowers* and *trees* imply a broad definition of animacy, and one that is not constrained by sentience; there are no human referents in the animate categories; and the ratings themselves indicate little differentiation across the animate categories, as is clear from Table 2.4.

**Table 2.4**: Animacy ratings of stimuli in von Studnitz and Green (2002), based on categories and exemplars derived from Battig and Montague (1969)

| Categories | | Mean | Categories | Mean |
|---|---|---|---|---|
| (Animate) | | Rating | (Inanimate) | Rating |
| Fruits | | 6.25 | Clothing | 1.48 |
| Vegetables | | 6.16 | Toys | 1.81 |
| Flowers | | 6.71 | Kitchen utensils | 1.47 |
| Trees | | 6.55 | Furniture | 1.50 |
| | FLORA | 6.42 | Carpenters' tools | 1.49 |
| Four-footed animals | | 7.98 | Stores | 1.74 |
| Fish | | 7.65 | Weapons | 1.69 |
| Birds | | 7.76 | Musical instruments | 1.89 |
| Insects | | 7.44 | | |
| | FAUNA | 7.71 | | |
| | ALL | 7.06 | ALL | 1.63 |

In the course of their experiments on the sensitivity of lexical decision to concrete and abstract stimuli, Kroll and Merves (1986: Appendix) created a dataset of 212 concrete and abstract nouns, matched on word length (number of letters) and on word frequency (from Kučera and Francis, 1967). Their definition of a concrete noun encompasses both animate and inanimate entities. The 212 nouns were rated by 101 undergraduate participants who were "encouraged to use imageability and the availability of sensory experience as criteria in making their judgments" (Kroll and Merves, 1986: 106). Participants rated the nouns on a seven-point scale from (1) *highly abstract* to (7) *highly concrete*. The resultant mean ratings were 2.7 for abstract nouns and 6.2 for concrete nouns, with no overlap between the two categories. Post-hoc analysis tested whether frequency was a determinant of concreteness, but the effect was only marginally significant. The results of the rating exercise were compared with the concreteness ratings derived by Paivio, Yuille and Madigan (1968). Of the 212 nouns, 130 were included in the Paivio *et al* norms. The correlation of these common items was a very strong 0.96.

Amazon Mechanical Turk, an internet-based crowdsourcing facility, has been used by Brysbaert, Warriner and Kuperman (2014) to collect concreteness ratings from over 4,000 participants who were all current US residents. Brysbaert *et al* (*ibid*: 907) excluded from their published concreteness ratings any words that were rated as "not known" by more than 15% of their raters. This reduced their original list of 63,039 items (including 2,940 two-word expressions) to 39,954 (including 2,896 two-word expressions), a reduction of 37%. They compared their concreteness ratings (*ibid*: 908) with matching items on the MRC Psycholinguistic Database (Coltheart, 1981). They report a high correlation ($r$ = 0.92).

Altarriba, Bauer and Benvenuto (1999) have argued that emotion words are significantly different from other abstract words, and should be regarded as a separate category. In their experiment, 326 words (155 abstract, 100 concrete and 71 emotion words), taken from previous studies of concreteness, were matched by frequency and word-length, then rated by 78 participants on three scales: concreteness, imageability and context availability. Altarriba and colleagues found that "the three word types [concrete, abstract and emotion] are reliably different from each other" (*ibid*: 579).

Their essential point is that emotion and abstract words are sufficiently differentiated to raise the possibility that previous research findings have been biased by their amalgamation. They therefore recommend their ratings (itemised in an extensive Appendix) for use by future researchers. However, they have not compared like with like. All 100 concrete words are nouns, but only 11 of the 71 emotion words are nouns, the other 60 are adjectives. Of the 151 abstract words, 14 are verbs, 16 are adjectives, one is an adverb (*now*) , and the remainder (120) are nouns.

Rating is an expensive exercise (though crowdsourcing offers a much cheaper alternative) and is heavily dependent on presenting participants with instructions that are clear and unequivocal without any bias – a difficult task. On the other hand, the rating scores obtained (after aligning the different scales) from the wide range of studies reviewed generally correlate highly. Subsequent chapters will test the genitive ratio as an alternative method of differentiating *concrete* from *abstract*, whilst also acknowledging the limitations of the genitive ratio and its own reliance on ratings for noun categories that defy genitive ratio analysis.

## 2.9 Corpus Coding of Animacy

Bresnan and Hay (2008: 249) cite two typical schemes for coding animacy. Garretson *et al* (2004) coded for seven categories of animacy:

Human > animal > organisation > concrete inanimate > non-concrete inanimate > place > time.

Bresnan *et al* (2007) opted for just four categories:

Human > organisation > animal/intelligent machine > inanimate.

Their *human* category contained individual humans, "humanoids" (gods, ghosts, robots), and groupings of humans that do not qualify as organisations, e.g. students or customers. Categories such as organisations, communities, nations, metaphorical and metonymical references, are all examples of what Dahl (2000: 100) calls "borderline cases of personhood".

Denison, Scott and Börjars (2008) identified eight categories of animacy in the spoken component of the British National Corpus (BNC):

animal
body part
collective human
human
inanimate abstract
inanimate concrete
place
time

A subsequent re-analysis of the same data by Börjars, Denison and Krajewski (2011) eliminated *collective human* and combined the two *inanimate* categories into one. When the research objectives changed, so did the categorisation.

A corpus study by O'Connor, Maling and Skarabela (2013: 97-98) defined animacy by three superordinate variables with subordinate levels:

| | |
|---|---|
| ANIMATE | a. human(oids) |
| | b. animals |
| ORGANISATION | a. human organisation |
| INANIMATE | a. concrete objects |
| | b. locations |
| | c. temporal entities |
| | d. other non-concrete entities |

Zaenen, Carletta, Garretson *et al* (2004) devised a corpus annotation scheme that assigned three main categories of animacy (human, other animate, inanimate) to noun phrases in a corpus, but tagged with sub-categories for the *other animate* and *inanimate* categories. Table 2.5 summarises their coding system.

**Table 2.5** Summary of animacy coding in Zaenen *et al* (2004)

| | |
|---|---|
| HUMAN | Look and act like humans |
| OTHER ANIMATE | |
| Organisation | Group of humans with collective identity, voice or purpose |
| Animals | All non-human animates, including cellular |
| Intelligent machines | E.g. computers or robots (very rare) |
| Vehicles | Because sometimes referred to as if animate (very rare) |
| INANIMATE | |
| Concrete | Tangible inanimates |
| Non-Concrete | Events, non-tangibles (e.g. *air*, *voice*, *wind*) |
| Place | E.g. *at work* |
| Time | Expressions describing periods of time |

## 2.10 Hierarchies of Animacy

Animacy "can be regarded as an assumed cognitive scale of some measure, extending from human through animate to inanimate" (Yamamoto, 2006: 29). A basic animacy hierarchy might resemble the one proposed by Rosenbach (2006: 105):

| [2.5] | human | > | animal | > | collective | > | inanimate |
|---|---|---|---|---|---|---|---|
| | (*girl*) | | (*dog*) | | (*family, church*) | | (*chair*) |

The examples are Rosenbach's. It is interesting to note that *church* might be construed primarily as an inanimate place (along with, say, *school* or *hospital*) rather than as an animate collective, whilst *chair* might be construed as a person (a gender-neutral equivalent of *chairman*).

Many animacy hierarchies are in fact derived from models of salience or semantic dominance. Clark and Begun (1971) postulated a semantic 'dominance hierarchy' that distinguished between count-nouns and mass-nouns:

Human nouns (*teacher*)

Animal nouns (*dog*)

Concrete count-nouns (*tree*)

Concrete mass-nouns (*snow*)

Abstract count-nouns (*fact*)

Abstract mass-nouns (*harm*)

Siewierska (2004: 149) situates animacy as the third of five "familiar topicality hierarchies" that determine the "inherent and discourse saliency" of linguistic factors:

a. The person hierarchy:

1st > 2nd > 3rd

b. The nominal hierarchy:

pronoun > noun

c. The animacy hierarchy:

human > animate > inanimate > abstract

d. The referential hierarchy:

definite > indefinite specific > non-specific

e. The focus hierarchy:

not in focus > in focus

Foley and Van Valin (1985: 288) proposed an anthropocentric "individuation scale", a hierarchy of persons that is based on how animacy is encoded linguistically. With minor changes in wording, their scale is almost identical to that of Lyons (1999: 213-215), whose scale is itself a distillation of the various animacy hierarchies proposed in the literature up to that point (see also Croft, 1990). The terminology differs slightly, the trajectory not at all:

| **Foley and Van Valin (1985)** | **Lyons (1999)** |
|---|---|
| Speaker/addressee | First- and second-person pronouns |
| Third-person pronouns | Third person pronouns |
| Human proper nouns | Proper names |
| Human common nouns | Common nouns with human reference |
| Other animate nouns | Non-human animate nouns |
| Inanimate nouns | Inanimate nouns |

This is not so far removed from the *scala naturae* (see section 2.1), or from Yamamoto's (1999: 16) suggestion that we might discriminate different levels of animacy in terms of the "biological distance" between ourselves and (say) an amoeba or a water flea.

Dahl and Fraurud (1996: 62-63) propose extending the basic animacy hierarchy to include:

[2.6] Metaphorical extensions

Time heals all wounds.

[2.7] Metonymical extensions

London has secured the 2012 Olympics.

[2.8] Collective nouns

The team won the Cup.

[2.9] Non-personal agents

Microsoft increased its annual profit.

[2.10] Mythological beings

The gods are smiling down on us today.

[2.11] Animals

Pigs are really very clean.

Our cat, Daisy, is a real character [anthropomorphism]

Dixon (1979: 85) classifies animals as "higher" (e.g. dogs) and "lower animal forms". In a discussion of animacy and gender, Dahl (2000: 100) observes that "gender distinctions often cut through the animal kingdom, with at least some higher animals being treated as persons and at least some lower animals being seen as inanimate". Yamamoto (1999: 22) has presented a "simplified" radial model of a General Animacy Scale. This has been reproduced in Figure 2.2

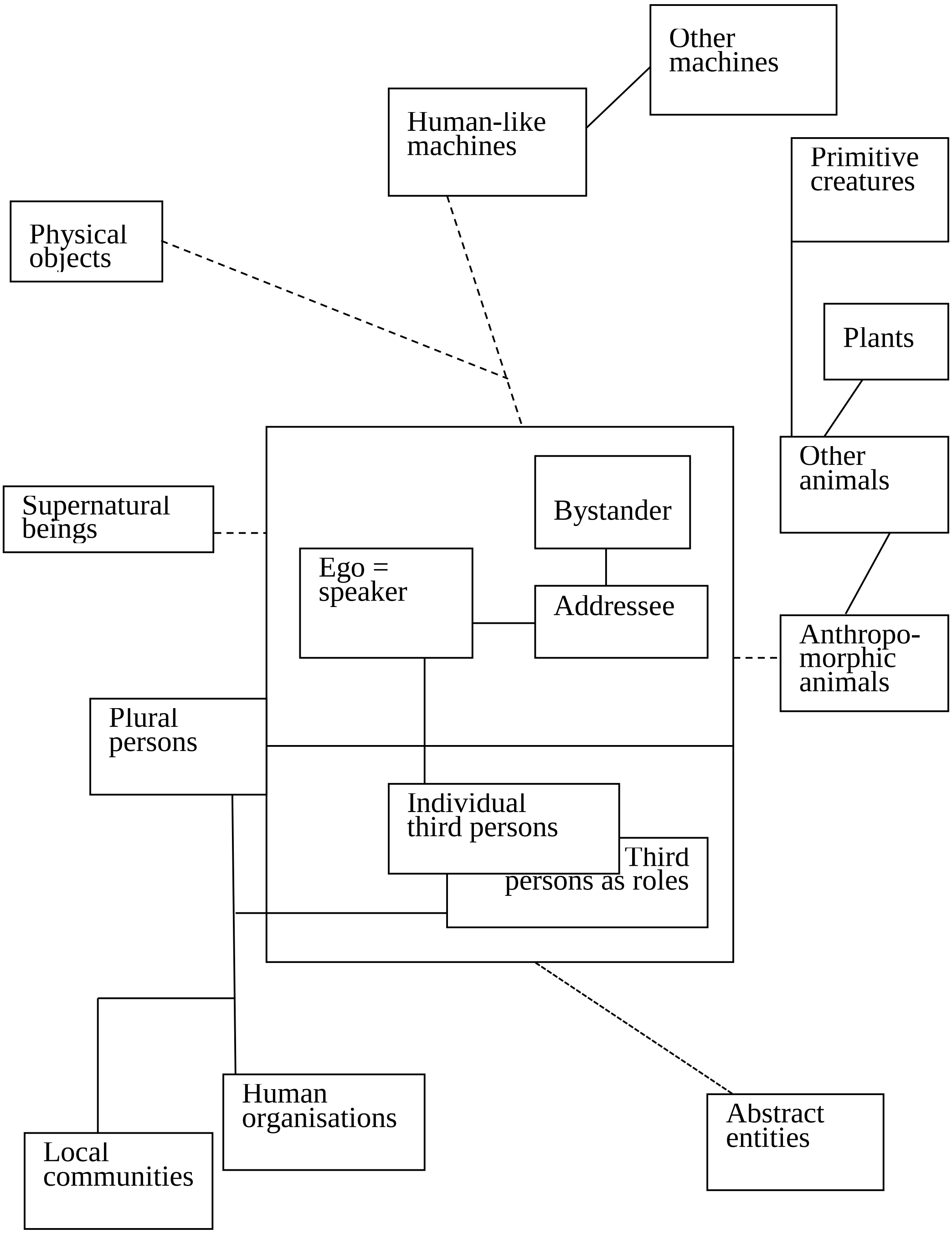


Figure 2.2: Animacy: Radial gradience with human sub-categorisation

Reproduced from Yamamoto (1999: 38, Figure 3)

With individual human beings at its centre, the model extends to categories of physical objects, machines, plants, primitive creatures, supernatural beings, abstract entities, metonymic organisations and communities at its perimeter. The human category in the central box of Yamamoto's model represents an interaction with two parameters: a Hierarchy of Persons and an Individuation Scale.

The Hierarchy of Persons (from Langacker, 1991: 306-307) differentiates the participant roles in a discourse into a salience hierarchy of first person (speaker) > second person (addressee) > third person (others, bystanders). The Individuation Scale (from Foley and Van Valin, 1985: 288) is a measure of how uniquely identifiable an entity is: singular or plural, pronoun or common noun, role-name or proper name; all affect the degree of definition of an entity and the "psychological distance" between a speaker and the referent.

Figure 2.3 presents a tabular comparison of four animacy hierarchies. Yamamoto's (1999) seventeen-category 'radial gradience' model has been reduced by the omission of three 'minor' categories (*supernatural beings, human-like machines* and *primitive creatures*), for the sake of clarity and simplicity. The other animacy hierarchies are from Clark and Begun (1971), Rosenbach (2006: 105) and Siewierska (2004: 149).

| **Clark & Begun (1971)** | **Rosenbach (2006)** | **Siewierska (2004)** | **Yamamoto (1999)** |
|---|---|---|---|
| human | human | human | ego/speaker<br>addressee<br>bystander<br>individual 3rd persons<br>3rd persons as roles<br>plural persons |
| animal | animal | animal | human organisations<br>local communities |
|  | collective | collective | anthropomorphised animals<br>other animals |
| concrete – count<br>concrete – mass | inanimate | inanimate | plants<br>physical objects<br>other machines |
| abstract - count<br>abstract - mass |  | abstract | abstract entities |

**Figure2.3**: Categorical comparison of animacy hierarchies

## 2.11 Caveats and Conclusions

The animate – concrete – abstract categories might be defined objectively, but their "reality" is subjective and constructed, moderated by individual differences that are psychological, social and cultural. Natural language is ambiguous, 'fuzzy'. English nouns often have a number of different senses that might cross categorical boundaries. Consider the noun *set*: animate (as in a group of people), concrete (the scenery or backdrop of a film), abstract (a series or sequence). The three categories introduced in 2.1 are not discrete, and the genitive ratio is not an absolute measure. This is an important caveat that will be re-emphasised in each of the applications considered in subsequent chapters.

The number and range of studies that rely on the distinction of *living* from *non-living*, or of *concrete* from *abstract*, prompts the thought that the genitive ratio might find a role in the selection of experimental materials, perhaps deployed as a filter of potential words, alongside other matching criteria.

The review of computational applications developed by Orăsan and Evans (2000, 2001 and 2007) exemplifies an incremental and progressive development to which the genitive ratio has not yet been exposed. Their incremental progress brought in additional ideas and processes that were integrated for the enhancement of the earlier work. Chapters 5-7 will advocate a future for the genitive ratio as a significant component of co-reference resolution and of risk assessment or diagnostic models.

The question of objectivity is critical to an assessment of participant-based rating schemes, that still represent (in the absence of viable alternatives) the gold standard of noun categorisation. The most interesting development is the use of Web-enabled crowdsourcing by Brysbaert *et al* (2014), utilising the "wisdom of crowds" (Surowiecki, 2005) at relatively low cost. Chapter 4 will suggest that genitive ratio analysis is a viable method for differentiating the majority of concrete and abstract nouns, but must itself fall back on ratings for certain sub-categories such as temporal nouns.

The evident diversity of the corpus coding schemes and hierarchies of relative animacy conveys the point that there is no 'canonical' categorisation of animacy. Categorisation is a product of the data and the research objectives of the analysis.

# CHAPTER 3

# The Genitive Ratio

*We as humans resonate more with creatures like ourselves ... charismatic mega-vertebrates.*
Professor Lord (Robert) May

## 3.0 Overview

The English possessive is formed by selecting either an *s*-genitive construction (*my mother's*) or an *of*-phrase genitive construction (*of my mother*). There is a consensus in the literature that the relative animateness of a possessor noun is the principal factor that biases an *s*-genitive construction. Concrete and abstract nouns are much more likely to take an *of*-genitive construction.

If a noun's degree (or absence) of animateness is the key factor that affects the probability of selecting a particular genitive construction, then the reverse hypothesis is that a noun's ratio of *s*-genitives to *of*-genitives, as measured in a corpus, will be a reliable proxy for that noun's animateness, concreteness or abstractness, relative to other nouns within a text or discourse. That ratio is the noun's **genitive ratio** (GR).

At a categorical level, the genitive ratio hypothesis (see chapter 1) is tested on a database of 41,000 genitive constructions that have been independently annotated with categories of animacy (from *human* to *inanimate abstract*). Analysis of those data provides supporting evidence of a *prima facie* correlation between a noun's genitive ratio and its categorical animacy.

The findings from a sub-categorical analysis of proper nouns within the *human, collective human* and *place* categories indicate that proper nouns and common nouns follow different conventions of genitive selection. With supporting evidence from prior studies that there is a specific 'naming effect', this argues for the exclusion of names from the GR method. A sub-categorical analysis of two conditions, number and definiteness, suggests that the effect of relative animacy upon the selection of a possessive construction is most evident in the singular condition, regardless of the definiteness condition.

### 3.1 The English Genitive

The genitive, or possessive, construction in English is formed by selecting either an *s*-genitive (poss-*s*) or an *of*-phrase genitive (poss-*of*). Each genitive construction typically contains a possessor and a possessum. For example:

[3.1] My mother'*s* [possessor] thoughts [possessum]

[3.2] The thoughts [possessum] *of* my mother [possessor]

"In any period of the English language and in any variety of English the *s*-genitive has always been more frequent with animate possessors than with inanimate ones". (Rosenbach, 2014: 241).

The *s*-genitive, with its origins in Old English, is sometimes called the Saxon or Germanic genitive. A full account of genitive constructions is beyond the scope of this thesis, but detailed analyses exist. Altenberg's (1982: 296) study of 17th century English identified more than forty factors potentially affecting the choice of genitive. No less than 13 variants of English genitive constructions have been identified by Denison, Scott and Börjars (2008), in their study of Germanic possessives in English and Swedish. Rosenbach (2014) provides a comprehensive survey of the empirical research carried out on the "genitive variation" in English, and suggests that it is "arguably the best researched of all syntactic alternations in English" (*ibid*: 215).

### 3.2 Animacy and the Genitive

There is a strong and growing body of evidence that animacy influences the genitive form in English (e.g. Dahl and Fraurud, 1996: 49; Yamamoto, 1999: 28; Rosenbach, 2008: 152). In a review of English genitive variation, Grafmiller (2014: 471) has found that "No single factor has been shown to influence this choice [of genitive construction] more than possessor animacy".

The *s*-genitive is becoming more frequent in informal, particularly in journalistic, language and in American than in British English (for empirical evidence, see Rosenbach, 2003). Hinrichs and Szmrecsanyi (2007) have reported, based on diachronic  (1960s and 1990s) corpus analyses of journalistic texts, a steady increase in the relative use of the *s*-genitive. Wolk *et al* (2013) estimate that the overall proportion of *s*-genitives in Present Day English (1950-1999) is 38%. Grafmiller (2014: 472) has reviewed cumulative evidence of the increasing frequency of poss-*s*, in both spoken and written language, over the last 50 years.

Nevertheless, based on a corpus analysis of 40 million words, Biber *et al* (1999: 301) conclude that "*s*-genitives are outnumbered by *of*-genitives in all registers", although "nouns with human/personal reference, especially proper nouns, tend to occur with the *s*-genitive rather than an *of*-phrase" (*ibid*: 302). Their generalised findings indicate that the *of*-phrase construction is particularly favoured by inanimate nouns (both concrete and abstract), by collective nouns for groups of people, and by plural nouns.

Hinrichs and Szmrecsanyi (2007) conclude that "a human possessor is ... the single most powerful categorical predictor" of an *s*-genitive construction (*ibid*: 462). A corpus study by O'Connor, Maling and Skarabela (2013: 103) supported the findings of Rosenbach (2002) and of Hinrichs and Szmrecsanyi (*ibid*) in determining (through a logistic regression) that animacy is the strongest factor in determining the genitive form.

In a multivariate analysis using logistic regression, Denison *et al* (2008) found that the category with the strongest bias to possessive-*s* was the "non-collective human referent", i.e. *human*. Yamamoto (1999: 50-52) cites the findings of a corpus study by Leech, Francis and Xu (1994) of the animacy of possessors in English *s*-genitive constructions. With a near-zero occurrence of inanimate *s*-genitives reported by Leech *et al*, there would seem to be a clear correlation between the *s*-genitive and human or quasi-human characteristics of the possessor.

However, such an unequivocal conclusion has inevitably been challenged. Dabrowska (1998) tested the Leech, Francis and Xu (1994) model against a very specific text type: computer (software) manuals. She found that 39% of inanimate 'computer nouns' did take an *s*-genitive. A possible explanation is that computer terminology is conceptualised within a framework of predominantly animate metaphors, e.g. *memory*, artificial *intelligence*, *mouse*, *virus*, *bug*, *malicious* software. From a comparison of the software manuals with journalistic texts, Dabrowska found that, when such nouns were used in their non-computer sense, as in the journalistic texts, they were significantly less likely to take an *s*-genitive.

### 3.3 Other Factors Affecting Genitive Constructions

The selection of a genitive construction is the cumulative product of a number of factors that combine to affect the relative frequency of the selected construction.

Although animacy "exerts a strong statistical bias" (Rosenbach, 2008: 152) on the selection of a genitive construction, there is prior research evidence of a complex interaction of additional factors with animacy.

Keizer (2007: 314) finds that the preference for a particular genitive construction is the outcome of an interaction involving no less than six potential factors. Whilst animacy is again the primary factor, Keizer proposes (*ibid*: 353) that there is a "degree of interdependence" between all of these six factors:

[3.3] 1. Gender/animacy of the possessor.

2. Number of the possessor.

3. Structural complexity of the possessor.

4. Presence of certain types of pre- or post-modifier of the head noun.

5. Centrality or prominence or topicality of the possessor/possessee.

6. Stylistic considerations.

Rosenbach (2002) has studied the effects of animacy, topicality and prototypicality (of possessor relations) on genitive constructions – factors which, as Rosenbach acknowledges, are strongly correlated. In order to separate their effects, Rosenbach tested all possible combinations of the three factors, by presenting her participants with passages extracted from a detective story, in which one noun phrase (NP) is a forced choice of either the *s*-genitive or the *of*-genitive construction. Table 3.1 (from Jäger and Rosenbach, 2006: 949) shows the conditions tested in Rosenbach's (2002: 137) experimental study, with examples of the items tested.

**Table 3.1**: Factors affecting choice of genitive construction: animacy, topicality and prototypicality (Jäger and Rosenbach, 2006: 949)

| [+animate] | | | | [-animate] | | | |
|---|---|---|---|---|---|---|---|
| [+topical] | | [-topical] | | [+topical] | | [-topical] | |
| [+proto] | [-proto] | [+proto] | [-proto] | [+proto] | [-proto] | [+proto] | [-proto] |
| *The boy's eyes/*<br>*The eyes of the boy* | *The mother's future/*<br>*The future of the mother* | *A girl's face/*<br>*The face of a girl* | *A woman's shadow/*<br>*The shadow of a woman* | *The chair's frame/*<br>*The frame of the chair* | *The bag's contents/*<br>*The contents of the bag* | *A lorry's wheels/*<br>*The wheels of a lorry* | *A car's fumes/*<br>*The fumes of a car* |

Statistical analysis of the results shows that the three factors are clearly differentiated, in spite of their correlation, and form the following hierarchy:

[3.4] animacy > topicality > prototypicality

However, the dominance of animacy is not absolute. Refer to the eight conditions and items shown in Table 3.1. In participants' choices, the first three conditions showed a significant preference for the *s*-genitive construction, but the fourth condition, negative for both topicality and prototypicality, favoured the *of*-genitive (*the shadow of a woman*). Jäger and Rosenbach (2006: 950) cite this as "clear evidence" for what they term "ganging-up cumulativity". In other words, the combined effect of the 'weaker' factors is potentially sufficient to overcome the strength of the animacy factor.

Eisenbeiss, Matsuo and Sonnenstuhl ( 2009: 158) have argued that the principal factors that bias the selection of poss-*s* or poss-*of* are, together with animacy, the "topicality and syntactic weight" of the possessor. Their typical possessor in an *s*-genitive construction is topical (or, more specifically, definite), syntactically unmodified (with a low word-count, e.g. a non-adjectival noun phrase), and animate.  Topicality carries two possible senses (see Seoane Posse, 1999, for a review). One sense of 'topic' is that of a referent that occupies initial position in a clause, its first constituent. The other sense is that the 'topic' is 'what the clause is about'. This "pragmatic aboutness" is the definition preferred by Reinhart (1981: 78). Börjars, Denison and Krajewski (2011) propose that definiteness might serve as a "proxy for topicality", whilst conceding that this is "clearly an oversimplification". Börjars, Denison, Krajewski and Scott (2013) do actually use definiteness as a proxy for topicality, though again acknowledging that it is "imperfect".

In their regression analysis of an "unconventional" dataset of English from different eras (1650-1999) and registers, Szmrecsanyi, Ehret and Wolk (2014) found that lexical weight (based on word-length) was second only to animacy in affecting the choice of genitive construction. "Syntactic weight" is generally considered to be a product of phrasal complexity, quantified by the number of nodes in a particular syntactic construction or simply by word-count, as proxy measures of the processing cost associated with comprehension. Grafmiller and Shih (2011) argue that syntactic weight is a "highly reliable" factor in the selection of a genitive construction. So, as the word-count of the possessor NP increases, relative to the word-count of the possessum NP, an *of*-genitive construction becomes more likely. For example:

[3.5]  The businessman's house

?The house of the businessman

[3.6] ?The thrice-married senior businessman's house

The house of the thrice-married senior businessman

## 3.4 Categorical Tests of the Genitive Ratio

Although the preceding review sets out a strong case for the *s*-genitive as a marker of animate nouns, application of the *of*-genitive to inanimate (concrete and abstract) nouns might have seemed not much more than the default option. A categorical analysis of genitive constructions will now test whether the animate-concrete-abstract progression, which is common to the hierarchical models of animacy reviewed in chapter 2.10, is emulated by the genitive ratio.

A database of possessive constructions (*s*-genitives and *of*-genitives) has been extracted from the spoken component of the British National Corpus (BNC) by Denison, Scott and Börjars (2008). This database was accessed at http://www.llc.manchester.ac.uk/research/projects/germanic-possessive-s/data/. The 'Denison database' will populate a model of animacy that will test the genitive ratio hypothesis at a categorical level.

These are the eight 'categories of animacy' with which Denison *et al* (2008) annotated their data:

[3.7] animal
body part
collective human
human
inanimate abstract
inanimate concrete
place
time

The analysis of these categories will explore both the feasibility and the limitations of the genitive ratio, identifying names and measurement nouns as 'special cases', and the significance of the factors of number and definiteness in defining (in chapter 4) operational algorithms based on the genitive ratio.

It is important to stress that Denison, Scott and Börjars (2008) are not advancing the hierarchy of animacy that this author has inferred from their data (Alan Scott, personal communication, 10 May 2011). Indeed, Börjars, Denison and Krajewski (2011) have subsequently revisited the original (Denison *et al*, 2008) dataset and reported on a new analysis that features a reduced set of six animacy categories: *collective human* has been eliminated, and the two inanimate categories (*concrete* and *abstract*) have been combined.

Denison *et al* (2008) hypothesise that relative animacy is one of a number of factors that influence the use of possessive-*s* in English. Their method is inductive and data-driven. They identify every genitive construction within their source corpus and then infer from those data eight categories of animacy that will differentiate those constructions for the purposes of their own research design. There is no category for "others".

A first consequence of this method is that the Denison *et al* taxonomy contains three categories (*time*, *body part*, and *place*) that do not feature in the Yamamoto model or in most other hierarchies of animacy (see chapter 2.10, figure 2.4). A second consequence is that Denison *et al* make no categorical distinctions between common and proper nouns. Hence their *human*, *collective human* and *place* categories contain (respectively) human names as well as roles, named organisations as well as collective nouns, and place-names as well as locations.

The independent annotation of the data provided by Denison, Scott and Börjars (2008) facilitates a restructure of the data in line with my own, very different research objectives. The original database provided by Denison *et al* (2008) contains 41,798 English BNC records and 2,361 Swedish records from the Gothenburg Spoken Language Corpus. Each record is annotated with 45 attributes. Most of these data (e.g. relating to the possessum or to phonology) were irrelevant to the current investigation and were therefore deleted. In addition to deleting the Swedish records, the database was 'cleaned' of items that did not lend themselves to the current analysis, by deleting 212 'double possessives' (e.g. *of Russell's*) and 169 items with a rating of 'none' or 'unclear' for animacy, number, or topicality. These latter deletions were also made by Börjars, Denison and Krajewski (2011) when they revisited the original dataset.

The reduced database, constructed for the purposes of this research, contained 41,417 records with the following attributes:

- Genitive type (*of*-possessive or possessive-*s*)
- Text (the context of the genitive phrase from the BNC)
- Animacy (one of the eight categories in [3.7] above)
- Number (singular or plural)
- Topicality (definite or indefinite)
- Line number (to provide a reference back to the original database)

The database was then partitioned into the eight categories in [3.7], in order to calculate a categorical genitive ratio (GR) for each category. This analysis yielded the results shown in Table 3.2. The two *inanimate* categories (*concrete* and *abstract*) are differentiated from the other categories, as well as from each other.

In Table 3.2 and subsequent analyses in this chapter, a chi-square (χ2) test for independence has been applied to the frequency data that have been extracted from the Denison database and classified into mutually exclusive categories within a series of contingency tables. Chi-square tests whether an observed pattern of distribution has occurred by chance.

Chi-square is well-established in corpus analysis as a test for independence (Oakes, 1998: 25). It is essentially "a measure of how much expected counts *E* and observed counts *N* deviate from each other. A high value of χ2 indicates that the hypothesis of independence, which implies that expected and observed counts are similar, is incorrect" (Manning, Raghavan and Schütze, 2008: 255).

Chi-square is a non-parametric test, i.e. it does not assume a normal distribution of the source data. "The only inference drawn from a significant result is that the different samples do not represent the same population distribution" (Hays, 1973: 734). The chi-square tests for independence in this chapter have *p*-values (two-tailed) set at the p < .05 level of significance. The Yates correction that is applied when sample sizes are small was not considered necessary. The effect size Φ is Cramer's phi, "a measure of the efficacy of prediction" (Ferguson 1976: 411):

$\sqrt{}( \chi 2 / (N\,(k - 1))$

where *N* is the total number of items and *k* is the smaller of the number of rows or columns. The guidelines for the significance of effect size are 0.1 small, 0.3 medium and 0.5 large.

**Table 3.2**: Categorical analysis of genitive constructions extracted from the British National Corpus by Denison *et al* (2008)

| **Category** | **Poss-*of*** | **Poss-*s*** | **GR** | **Rank** |
|---|---:|---:|---:|---:|
| Human | 3599 | 5491 | 0.66 | 1 |
| Time | 1991 | 1273 | 1.56 | 2 |
| Animal | 200 | 110 | 1.82 | 3 |
| Collective Human | 4333 | 1724 | 2.51 | 4 |
| Place | 4269 | 890 | 4.80 | 5 |
| Body Part | 363 | 20 | 18.15 | 6 |
| Inanimate Concrete | 3570 | 148 | 24.12 | 7 |
| Inanimate Abstract | 13346 | 90 | 148.29 | 8 |
| $\chi$2 | 12286.65 | | | |
| Degrees of freedom | 7 | | | |
| $p < .001$ $\Phi = 0.54$ | | | | |

The focus of this thesis is on the trichotomy of animate, concrete and abstract, and it is relatively simple to re-structure the Denison data into those three categories. The *animate* category combines *human*, *animal* and *collective human*, whilst the *concrete* category combines *place*, *body part* and *inanimate concrete*. The *inanimate abstract* category is unchanged. Table 3.3 presents the genitive ratios of the three categories. The omission of the *time* category will be discussed in the next section.

Hundt and Szmrecsanyi (2012: 246) specifically classify body parts as inanimate nouns, and a body part is typically referenced by the inanimate pronoun *it*. No other model or hierarchy of animacy has been found that includes *body part* as a separate category. Whilst it might be argued that body parts have a conferred animacy, an equally strong case might be made for regarding them as no more than the dependent components of an animate being, not as animate in their own right (a severed hand 'lives' only in horror films).

**Table 3.3**: Three-category model derived from the database constructed by Denison *et al* (2008)

| **Category** | **Poss-*of*** | **Poss-*s*** | **GR** |
|---|---|---|---|
| Animate | 8132 | 7325 | 1.11 |
| Concrete | 8202 | 1058 | 7.75 |
| Abstract | 13346 | 90 | 148.29 |
| $\chi 2$ | 9904.41 | | |
| Degrees of freedom | 2 | | |
| *p* < .001 Φ = 0.51 | | | |

The closest comparison to the categorisation by Denison *et al* is the previously cited study by Leech, Francis and Xu (1994), who analysed genitive constructions in nine sections of the Lancaster-Oslo-Bergen (LOB) Corpus. Their aim was to construct a logistic regression model that would predict the choice of a genitive construction, poss-*of* or poss-*s*, given just three factors: semantic class (equivalent to animacy category); style or text type (journalistic, academic or fictional); and the semantic relation of the possessor to the possessum. They analysed seven 'semantic classes' of animacy, with the same classification as in Denison *et al*, except for *body part*. Leech, Francis and Xu (1994: 71) conclude that it is the animacy of a referent's semantic class that is the most significant factor in determining the choice of genitive construction. Their data are presented in Table 3.4, with my addition of genitive ratios (poss-*of* divided by poss-*s*) and a consequent ranking.

**Table 3.4**: Genitive ratios and consequent ranking derived from a corpus study reported by Leech, Francis and Xu (1994: 62)

| **Category** | **Poss-*of*** | **Poss-*s**** | **GR** | **Rank** |
|---|---:|---:|---:|---:|
| Human | 186 | 224 | 0.83 | 1 |
| Place | 61 | 29 | 2.10 | 2 |
| Time | 41 | 14 | 2.93 | 3 |
| Collective Human | 76 | 25 | 3.04 | 4 |
| Animal | 26 | 4 | 6.50 | 5 |
| Inanimate Abstract | 399 | 2 | 199.50 | 6 |
| Inanimate Concrete | 467 | 1 | 467.00 | 7 |

**Note**: *****Because there were reportedly no instances of poss-*s* in the *inanimate concrete* category, the data in that column have been 'smoothed' by adding 1 to every category.

The data provided by Leech *et al* (1994) are useful in providing a comparison to the analysis of the Denison data, but there are reservations in that comparison. First, their dataset was only 3.7% of the size of the Denison database. Second, Leech *et al* analysed only "those occurrences of [X's Y] which could, in principle, be replaced by [the Y of X]" (*ibid*: 62) and vice versa. Third, two of the three text types analysed, journalistic and academic, have "by far" (according to Biber *et al*, 1999: 301) the highest frequency of poss-*s* and poss-*of* respectively. With those reservations in place, the analysis presented by Leech *et al* is broadly similar to the one presented below: the same differentiation of *human* from the other animate categories, and the clear differentiation of the two *inanimate* categories, *abstract* and *concrete*.

### 3.5 Time-Genitives

Leech, Francis and Xu (1994: 71) concluded that their own category of *time* was not only anomalous but might "suggest the existence of an independent class of genitives". That independent class has been defined by Payne and Huddleston (2002: 470) as "measure genitives" that most frequently measure either a length of time or a value. Rosenbach (2007: 182) refers to the "fuzzy categorical status" of measure genitives.

*Time* has a vocabulary that is well-defined and limited in its range. Less than fifty nouns constitute almost all the *time* data in the Denison database – the days, months and seasons plus the twenty words in Table 3.5. In fact, just four phrases ("of the day/ week/ month/ year") constitute 30% of the *time of*-genitives in the database.

**Table 3.5**: Words that measure time

| | | | | |
|---|---|---|---|---|
| *afternoon* | *century* | *day* | *decade* | *second* |
| *evening* | *fortnight* | *hour* | *instant* | *term* |
| *millennium* | *minute* | *moment* | *month* | *week* |
| *morning* | *night* | *quarter* | *season* | *year* |

A key property of such phrases is that they do not merely bias a particular genitive construction; they generally exclude the alternative construction:

[3.8] time of the day

*the day's time

[3.9] a week's holiday

*the holiday of a week

There are some exceptions to this exclusion. As observed by Börjars, Denison, Krajewski and Scott (2013), the following alternative constructions would be equally acceptable to native speakers:

[3.10] There was a delay of twenty minutes

There was twenty minutes' delay

Other exceptions are arcane and 'literary' phrases. Grafmiller (2014: 474) explains these unconventional genitive constructions as "artistic playfulness":

[3.11] the end of August

?August's end

[3.12] the first months of 1941

?1941's first months

When Börjars *et al* (2011) revisited the 2008 Denison database, they cited Payne and Huddleston's analysis as grounds for their decision to exclude such "measure possessives", particularly but not only those for *time*, precisely because "they often lack a poss-*of* alternative". This calls into question the *time* data reported by Leech, Francis and Xu (1994) since, as has already been noted, their data included only "those occurrences of [X's Y] which could, in principle, be replaced by [the Y of X]" (*ibid*: 62) and vice versa. Leech *et al* must therefore have reported only a subset of the *time* phrases in their corpus.

In conclusion, genitive constructions of *time* follow pragmatic conventions unrelated to animacy. They therefore do not follow the criteria for genitive selection observed in other categories. Because the vocabulary of *time* is limited and well-defined, it should be possible to treat *time* phrases as a special case in any computational analysis, by reference to a look-up table of *time* words.

## 3.6 Names

Human names are similar to time-words in that they follow different conventions of genitive selection. We would say *Kevin's hat*, but not *the hat of Kevin*, and definitely not *the hat of the/a Kevin*. The resulting ratios would be so absolute that they might usefully feature as markers in a model of named entity recognition and classification (NERC), but that is beyond the scope of this thesis. Collective names and place names are also subject to constraints that set them apart from their common noun counterparts.

Three of the original eight Denison categories contain a high proportion of names. In the Denison database, names and generic labels – proper nouns and common nouns -  reside together within the same categories (*human*, *collective human*, and *place*). Almost 20,000 items of data were manually coded to facilitate a sub-categorical analysis of the three categories.

### *Category: Human*

The *human* data were hand-coded into:

***Names***, any individual identified by a proper noun.

***Roles***, a sub-category that includes familial relationships, e.g. 'brother', 'parent', 'widow', 'kinsman', 'twin'; generic or demographic classifications, e.g. 'child', 'woman', 'adult'; occupations, e.g. 'con-man', 'student', 'councillor', 'prince'; and words that define relationships either to other people or to institutions, e.g.

'member', 'friend', 'patient', 'witness', 'communist', 'expert', 'employer'. The *role* sub-category also includes supernatural entities that are generally personified in human form, e.g. 'angel', 'saint', 'God', 'Allah', 'Satan'. 'Jesus' is a name, whereas 'Messiah' or 'Christ' are defined as supernatural/roles.

***Category: Collective Human***

The items in this category were coded into:

***Names***, representing 33% of the *collective human* items, and defined as a noun phrase in which by convention all nouns would be initially capitalised. This definition therefore excludes NPs in which the collective name functions as an adjectival modifier of the head noun. These were classed as *association* (e.g. 'Conservative group', 'Royal Navy contingent'). Thus, 'Labour government' and 'local authority' were coded as associations, whereas 'Labour Party' and 'Local Education Authority' were coded as names. Initials (e.g. NHS, IBM, IRA, ITN) and full names have very similar genitive ratios (1.46 for initials and 1.52 for names). The use of initials reflects a high degree of shared familiarity with the referent. The four examples given above are all more frequently used than their full names.

***Associations***, defined as entities with some formal structure and/or official constitution (e.g. 'guild', 'team', 'department', 'jury', 'charity'); and as collectives that are constituted informally and/or temporarily, for example by a shared environment (e.g. 'audience'), or kinship (e.g. 'family'), or age (e.g. 'generation'), or shared interest (e.g. 'membership', 'readership'). This sub-category also accommodates generic classifications such as 'class', 'public', 'gentry', 'group'.

### *Category: Place*

The *place* data were coded into ***place names*** and ***locations***. This distinction was not always as straightforward as simply defining a place name as somewhere that features in an atlas or street guide. That definition of place name was broadened, to include the 'label' of any geographical – or, indeed, imaginary or mythological – location that would have an individual identity within a discourse. Thus 'earth', 'heaven' and 'hell' were all coded as place names, as were 'East Oxford' and 'Eastern Europe', whereas 'east' on its own was coded as a location word. The most common location words were 'city/ies' (81), 'house/s' (115), 'county/ies' (140), 'world/s' (188) and 'country/ies' (208). These five together account for 27% of the location items.

### *Analysis*

Table 3.6 summarises the sub-categorical analyses of the *human*, *collective human* and *place* categories. The significant degree of independence ($p<.001$) of proper nouns from common nouns, within all three categories, indicates that the genitive ratios of names are significantly different from those of common nouns. There is equally a significant degree of independence between the three sub-categories of names (Table 3.7), indicating that 'names' are not a single category. The evidence argues for treating names – of people, organisations and places – as distinct categories that are individuated from their common noun referents, as *Philby* is from *spy*, *Google* is from *company*, or *Big Ben* is from *clock*.

**Table 3.6**: Summary sub-categorical analysis of named and generic genitive constructions extracted from BNC (Denison *et al*, 2008)

| **Category** | **Poss-*of*** | **Poss-*s*** | **GR** | | | |
|---|---|---|---|---|---|---|
| **HUMAN** | | | | | | |
| Names | 397 | 1916 | 0.21 | χ2 | *p* < | Φ |
| Roles | 2900 | 3192 | 0.88 | 651.6 | .001 | 0.28 |
| **COLLECTIVE** | | | | | | |
| Names/Initials | 1395 | 923 | 1.51 | χ2 | *p* < | Φ |
| Associations | 2891 | 703 | 4.11 | 290.1 | .001 | 0.22 |
| **PLACE** | | | | | | |
| Names | 1785 | 612 | 2.92 | χ2 | *p* < | Φ |
| Locations | 2434 | 277 | 8.79 | 207.6 | .001 | 0.20 |

**Table 3.7**: Sub-categorical analysis of names extracted from BNC (Denison *et al*, 2008)

| **Sub-category** | **Poss-*of*** | **Poss-*s*** | **GR** |
|---|---|---|---|
| Human | 397 | 1916 | 0.21 |
| Collective | 1395 | 923 | 1.51 |
| Place | 1785 | 612 | 2.92 |
| χ2 | 1665.97 | | |
| Degrees of freedom | 2 | | |
| *p* < .001 Φ = 0.49 | | | |

### *A Naming Effect*

Additional evidence for the individuation of names comes from studies of discourse salience and anaphoric reference that have identified a 'naming effect'. Sanford, Moar and Garrod (1988) manipulated a sentence continuation task to measure three antecedent factors: primacy of mention, centrality of role, and introduction by name versus by role-description. Only the latter factor was significant. They found that the introduction of a character by a proper name (e.g. *Harold*) rather than by role (e.g. *the publican*) was twice as likely to result in pronominal reference to that character. Fraurud (1996: 82) has suggested that labelling an entity with a proper name rather than a description somehow increases the entity's "discourse status". Yamamoto (1999: 28-29) observes that a name, encoded as a proper noun phrase, is much more individuating (and by implication therefore more salient) than a generic role that is encoded as a common noun phrase. Because the factors of animacy and topicality are highly correlated, and because names are consistently highly topical, Rosenbach (2003: 387) excluded proper nouns from an analysis of genitive choice.

### *Conclusion*

**The cumulative evidence supports Rosenbach's exclusion of proper nouns. In subsequent chapters, the genitive ratio analysis will be applied to common nouns only.**

### 3.7 Three Categories: Animate, Concrete and Abstract

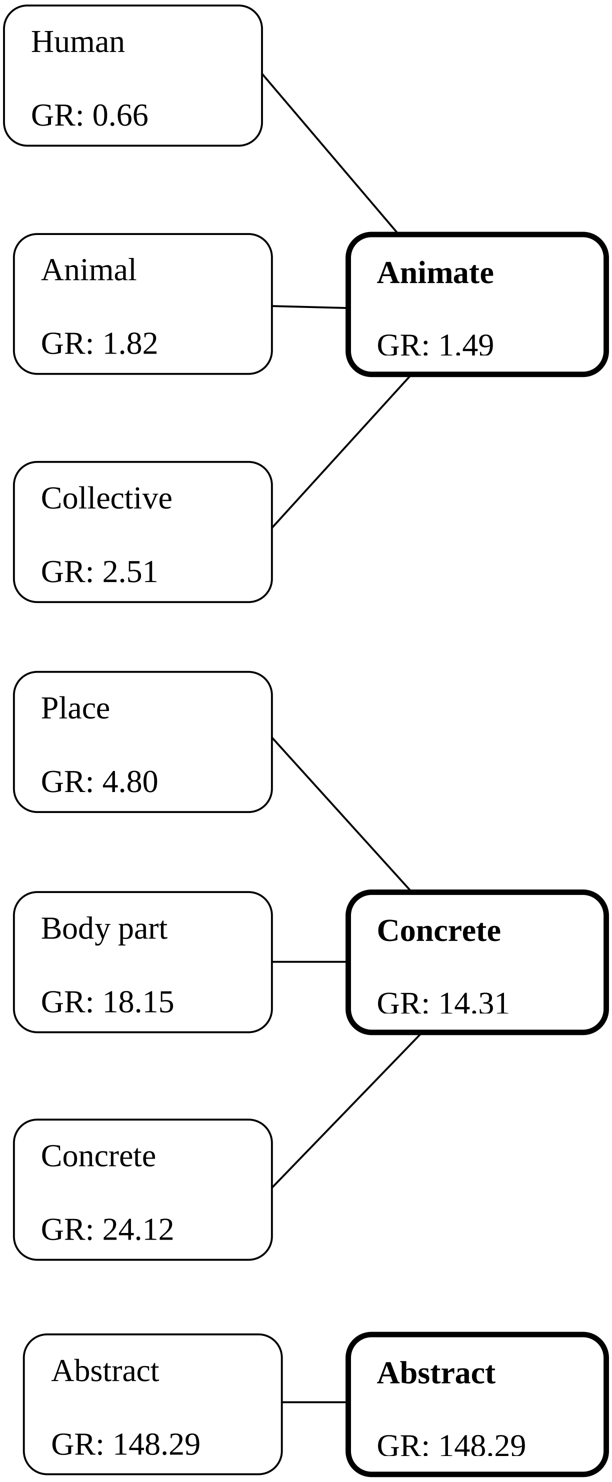


**Figure 3.1**: Denison categories and genitive ratios (from Table 3.3) mapped on to a three-category model

Figure 3.1 depicts the mapping of the Denison categories on to the three-category model. The exclusion of proper nouns yields the three-category analysis of common nouns in Table 3.8, with the resultant ranking.

**Table 3.8**: Consolidated categories derived from a re-analysis of data extracted from the British National Corpus by Denison *et al* (2008)

| **Category** | **Poss-*of*** | **Poss-*s*** | **GR** |
|---|---|---|---|
| Animate | 5973 | 4000 | 1.49 |
| Concrete | 6367 | 445 | 14.31 |
| Abstract | 13346 | 90 | 148.29 |
| χ2 | 7475.74 | | |
| Degrees of freedom | 2 | | |
| *p* < .001 Φ = 0.50 | | | |

### 3.8 Number and Definiteness

It is clear from the review of prior research evidence (see section 3.3) that, whilst animacy might be the primary factor affecting the choice of a genitive construction, it is not the only factor. Here we look for evidence of an interaction between animacy and the factors of number and definiteness ('topicality') that are coded in the Denison database.

Table 3.9 provides a number comparison of singular and plural genitive constructions. Table 3.10 similarly provides a comparison of definite and indefinite genitive constructions. In both comparisons, the data show the anticipated progression from *animate* to *abstract*.

**Table 3.9**: Number analysis of genitive constructions extracted from BNC by Denison *et al* (2008)

(**Sg** = Singular **Pl** = Plural)

| **Category** | **Sg-*of*** | **Sg-*s*** | **Sg-GR** | **Pl-*of*** | **Pl-*s*** | **Pl-GR** |
|---|---|---|---|---|---|---|
| Animate | 5263 | 5547 | 0.95 | 2502 | 1292 | 1.94 |
| Concrete | 6623 | 942 | 7.03 | 1529 | 115 | 13.3 |
| Abstract | 10952 | 86 | 127.35 | 2394 | 4 | 598.75 |
| χ2 | | | 8610.66 | χ2 | | 1313.66 |
| Degrees of freedom | | | 2 | Degrees of freedom | | 2 |
| *p* < .001 Φ = 0.54 | | | | *p* < .001 Φ = 0.41 | | |

The **number** data show that, in all categories, the *of*-genitive is more dominant in the plural than in the singular constructions. Both Keizer (2007: 314) and Biber *et al* (1999) have reported an inherent bias to the *of*-genitive construction for plural nouns, whilst the relatively low incidence of the *s*-genitive in the plural form is most likely to have a morphological or phonological explanation (Börjars *et al*, 2013).

Close analysis of the plural-*s* constructions shows that, in the majority of cases, the possessor is a unit of measurement, usually monetary (76 cases) and most usually pounds sterling (73) or else synonyms for pounds (2). Examples (with line numbers in brackets referring to the original data) are:

[3.13] three hundred thousand <u>pounds</u>' worth (37019)

fifty-seven <u>quids</u>' worth (40981)

Non-monetary units of measurement in this category include yards and metres, e.g.

[3.14] eighty thousand square <u>metres</u>' worth (20207)

These are further examples of what Payne and Huddleston (2002: 470) have called "measure genitives". As with the *time* examples discussed above, most do not permit the alternative genitive construction:

[3.15] *the worth of three hundred thousand pounds

*the worth of eighty thousand square metres

If these "measure genitives" were excluded, the plural-GR of the *concrete* category would be significantly different: 39.2. There were no plurals in the originally *body part* data that have been consolidated into the *inanimate concrete* category.

There is sound evidence here for basing the genitive ratio calculation on singular data. Singular constructions are more frequent (by almost 4:1); plural nouns are inherently biased to the *of*-genitive construction; and the plural *s*-genitive construction is skewed by measure genitives with no *of*-genitive alternatives. This finding, that singular constructions are the more reliable indicator of relative animacy is of particular significance, since the method to be deployed in the subsequent analysis (see chapter 4) will not accommodate plural constructions.

The analysis of definiteness data in Table 3.10 demonstrates that the ratios for the definite and indefinite conditions are similar, and might both be utilised in an operational implementation of the genitive ratio.

**Table 3.10**: Definiteness analysis of genitive constructions extracted from BNC by Denison *et al* (2008)

***Df** = Definite **In** = Indefinite

| **Category** | **Df-*of*** | **Df-*s*** | **Df-GR** | **In-*of*** | **In-*s*** | **In-GR** |
|---|---|---|---|---|---|---|
| Animate | 5623 | 5611 | 1.00 | 2147 | 1234 | 1.74 |
| Concrete | 6449 | 945 | 6.82 | 1706 | 115 | 14.83 |
| Abstract | 7439 | 49 | 151.82 | 5908 | 42 | 140.67 |

| | | | |
|---|---|---|---|
| χ2 | 6632.19 | χ2 | 2605.44 |
| Degrees of freedom | 2 | Degrees of freedom | 2 |
| *p* < .001 Φ = 0.50 | | *p* < .001 Φ = 0.48 | |

### 3.9 Languages Other than English

Could the genitive ratio method be applied to other Germanic languages, such as German itself? The answer is a qualified 'yes', though it would be more dependent on a tagged corpus. For example, English has just one definite article (*the*) both singular and plural, whereas in German the singular definite article is distinguished by gender (masculine, feminine or neuter) and by case (nominative, accusative, genitive and dative).

German does have an *–s* or *–es* genitive suffix, but it applies only to masculine and neuter nouns and it is not marked by an apostrophe. In written German particularly, possession is conveyed by the genitive case of the article (singular *des* or *der*), but *der* is also the article in the masculine nominative and feminine dative declensions:

[3.16] Das Dach des Hauses [neuter]

The roof of the house

[3.17] Das Dach der Kirche [feminine]

The roof of the church

In conversational German, the genitive is more often conveyed by the preposition *von* (*of*) which takes the dative article:

[3.18] Das Dach von der Kirche

An alternative and perhaps more viable approach is to use the English genitive as a 'bridge', by translating the other language into English and then carrying out the genitive ratio analysis. This approach is tested empirically in a later chapter (see section 5.16).

### 3.10 Caveats and Conclusions

This is the genitive ratio (GR) hypothesis as stated in chapter 1:

> **For any noun, the ratio of possessive-*s* constructions to possessive-*of* constructions, as quantified by a corpus analysis, should provide a proxy measure of that noun's relative animacy.**

Animacy is relative because it exists on a continuum, from animateness to abstractness via concreteness. Animacy is therefore not the same as animateness. Animacy is the continuum; animateness is a semantic band within that continuum.

This chapter has provided both theoretical and empirical support for that hypothesis, but only within a test environment of supervised data. The database constructed by Denison, Scott and Börjars (2008) provides 'clean data': the only nouns analysed are those encountered in genitive constructions. Their analysis consequently sidesteps some of the confounds of natural language. Whilst the genitive ratio has passed its first test, with verification at the categorical level, the

GR method must still be 'proved' in the more demanding context of unsupervised natural language.

The sub-categorical analysis of the Denison data has concluded that proper nouns (names of people, organisations and places) are significantly differentiated from common nouns. Analysis also identified the singular form of a genitive construction as the preferred basis for the ratio calculation.

The genitive ratio is primarily a gradient measure of relativity rather than of classification. It will not classify a noun as human or abstract in a sentence such as:

[3.18] The writer seethed with anger

but it should differentiate the two nouns in terms of their relative animacy. Just how it will do so is the subject of the next chapter.

CHAPTER 4

# Computing the Relative Animacy of Text

*Everything is vague to a degree you do not realise till you have tried to make it precise.*

Bertrand Russell

## 4.0 Overview

In the previous chapter, a categorical analysis of data supported the feasibility of the genitive ratio, based on a three-category model of animate vs. concrete vs. abstract nouns. There were two other significant findings:

- Calculation of the ratio is most accurately based on counts of singular nouns.
- There are exceptional cases in which the genitive ratio is not feasible, particularly proper nouns and temporal nouns.

This chapter will build on the foundation of that analysis, with the ultimate objective of defining and constructing  computational models that will assign values of relative animacy (on a gradient from animate to abstract) to individual nouns or to a text.

Looking ahead to the different applications discussed in subsequent chapters, two models are defined: an animateness rating that will differentiate animate from inanimate (concrete and abstract) nouns; and a concreteness rating that will differentiate concrete (both animate and inanimate) from abstract nouns.

The chapter introduces Animyser, a program that utilises Wikipedia as a corpus to obtain the phrase-search results-counts required to calculate a genitive ratio. The feasibility of an empirical analysis that utilises phrase-search and Wikipedia is supported by a review of prior studies.

Specifically, the chapter addresses three questions:

- Is the basic three-category model robust when tested on a new dataset?
- Which of the possible ratios, or combinations of ratios with intra-linguistic factors, will be the most reliable predictors of a noun's relative animacy?

- Is it feasible to extend the three-category model into a more fine-grained differentiation of the three animate sub-categories (human, animal and collective human)?

Both exceptions to and special cases of the genitive ratio model are acknowledged and discussed. The extent of those exceptions should not invalidate the GR concept, provided that the corpus (Wikipedia) yields sufficient data. Pre-processing in Animyser deals with the special cases.

### 4.1 The Animyser Program

This chapter relies upon Animyser (Animacy Analyser), a program that has been developed to extract phrase-search results-counts from Wikipedia. An outline specification of the program can be found at Appendix 4.1.

The program utilises imported 'Pattern' data mining modules, provided by the Computational Linguistics and Psycholinguistics Research Centre (CLiPS) at the University of Antwerp, and is written in Python v.2.7 in order to be compatible with two modules of Pattern:

**pattern.en**

This is a natural language processing (NLP) toolkit for English, from which the program imports a 'singularizer' module and the POS (part of speech) tagger.

**pattern.web**

This provides an API (application programming interface) with Wikipedia. The pattern.web API executes searches of Wikipedia for each target noun and returns results-counts from six phrase-search templates, listed in Table 4.1.

**Table 4.1**: Phrase-search templates

| | | | |
|---|---|---|---|
| **OD** | **O**f-**D**ef | The of-genitive with the definite article | "of the army" |
| **DS** | **D**ef-**s** | The *s*-genitive with the definite article | "the army's" |
| **OI** | **O**f-**I**ndef | The *of*-genitive with the indefinite article | "of an army" |
| **IS** | **I**ndef-**s** | The *s*-genitive with the indefinite article | "an army's" |
| **ON** | **O**f-**N**ull | The *of*-genitive with no article | "of army" |
| **NS** | **N**ull-**s** | The *s*-genitive with no article | "army's" |

Pre- and post-processing of the data obtained from Wikipedia will be specified in later sections of this chapter.

### 4.2 Phrase-search

Several prior studies have used phrase-search to obtain results-count data, all from Google. By enclosing a "phrase-search" in quotation marks, Google will search for that exact phrase only. Google also provides a wildcard option, whereby the insertion of an asterisk into a search string will identify phrases that include one or more unspecified words.

Modjeska, Markert and Nissim (2003) utilised Google Search within an algorithm for resolving *other*-anaphors. Given a set of possible antecedents, the algorithm searches Google with a sequence of phrases combining each potential antecedent with the *other*-anaphor. For example, the text:

The scientists had attended Cambridge and other universities

would generate Google phrase-searches for "scientists and other universities" (3 results)

"Cambridge and other universities" (91,600 results)

thus resolving the anaphor *other universities* to the antecedent *Cambridge*. (The results-counts quoted in this and subsequent examples are the counts obtained by this author).

To acquire gender information for anaphora resolution, Bergsma (2005) combined search-pattern data obtained both from a parsed corpus and from the Web (via the Google Search API). To obtain a gender classification for *Winston Churchill*, for example, the API would submit a series of "flat pattern" phrase-searches in the format:

[4.1] "Winston Churchill * [+ possessive pronoun: *his*, *her*, *its* or *their*]"

The results-count for each pronoun-pattern is a proxy for the statistical probability of gender classification:

[4.2] "Winston Churchill * his" (160,000,000 results)
"Winston Churchill * her" (467 results)
"Winston Churchill * its" (464 results)
"Winston Churchill * their" (455 results)

The data derived from Google proved to be more accurate (90%) than the data from the parsed corpus (84%), with only a marginal improvement in accuracy (92%) if the two were combined.

Ji and Lin (2009) have emulated Bergsma's (2005) pattern-matching process to determine animacy, but with the significant advantage of direct access to Google's n-gram (n=5) corpus of (then) 207 billion tokens in English. To determine the probable animateness of a target noun, Ji and Lin use a series of relative-pronoun patterns, e.g. for the noun *linguist*:

[4.3] "*linguist* who" animate (235,000 results)

"*linguist* that" inanimate (23,700 results)

"*linguist* which" inanimate (22,300 results)

Based on Google phrase-searches of Norwegian texts, Nøklestad (2009) devised an alternative method of assessing the animacy of a noun. A range of search-patterns consist of:

[4.4] personal pronoun + verb phrase (VP) + target noun

Thus, to determine the animacy of the noun *taxi-driver*, multiple combinations of personal pronouns and VPs would be submitted to Google. The following much-simplified example combines just one of the VPs (*works as a*) with *taxi-driver*:

[4.5] "He works as a taxi-driver" (4,238 results)

"She works as a taxi-driver" (35 results)

"It works as a taxi-driver" (0 results)

The Google results-counts clearly indicate that the *he/she* animate constructions outnumber the *it* inanimate construction, and therefore that the noun *taxi-driver* is very probably animate (human). A significant overhead of Nøklestad's method is that it requires up to 3,150 phrase-searches for every noun encountered. More importantly, the results-counts provide not so much a test of animateness, but rather one that is narrowly "geared towards discovering expressions that refer to humans" (Nøklestad, 2009: 15).

Google Search was neither intended nor designed to be a tool of precise linguistic research, and Google withdrew their University Research Program for Google Search, together with the Search API, in 2012. Fortunately, there is a readily-accessible alternative resource, with precedents as a corpus for linguistic research. The massively comprehensive online encyclopaedia, Wikipedia, is sufficiently large and diverse to offer a high probability of success to any phrase-search, and incorporates its own Special Search facility (with an API). Wikipedia offers the same facilities of wildcards, exact phrase-search and results-counts as does Google Search, and with data of high quality (see next section).

## 4.3 Wikipedia as Corpus

"When seen from a corpus perspective, Wikipedia defies all definitions".
Gatto (2014: 208)

Medelyan *et al* (2009) provide an overview of Wikipedia-based research related to NLP applications. In corpus terms, they define Wikipedia as occupying the "middle ground between … quality and quantity – by offering a rare mix of scale and structure" (*ibid*: 717). These statistics for the English Wikipedia (as at October 2015) indicate its scale:

- Number of pages: 37,652,968
- Number of articles: 4,995,560

(Source: https://en.wikipedia.org/wiki/Wikipedia:Statistics)

There are arguments for and against Wikipedia as a corpus. Whilst it is API-accessible, it is in corpus terms 'unsupervised', but then so is the Google Web. Any Web-wide search will collect performance errors: mis-spellings, inaccurate punctuation, inappropriate usage, etc, particularly since a high proportion of web pages in English originate from non-native English-speakers (Rosenbach, 2007: 168). The formality of 'Wikipedia English' should minimise these problems, since colloquialisms, mis-spellings and errors of grammar and punctuation are relatively infrequent. On the other hand, Wikipedia's formality represents a genre concern, since the *s*-genitive is more frequent in conversational and informal language (Rosenbach, 2003) than in the academic language to which Wikipedia aspires.

A sample of the most commonly abbreviated words (Table 4.2) provides evidence of Wikipedia's relative formality of language, whilst confirming that both formal and informal word-forms are well-represented.

**Table 4.2**: Frequencies of informal and formal usage of commonly abbreviated words in Wikipedia

| **Informal** | **Frequency** | **Formal** | **Frequency** |
|---|---:|---|---:|
| photo | 107,315 | photograph | 34,985 |
| exam | 13,583 | examination | 37,773 |
| phone | 49,732 | telephone | 43,214 |
| fridge | 1,349 | refrigerator | 3,088 |
| gym | 16,438 | gymnasium | 22,212 |
| info | 79,196 | information | 1,019,377 |
| memo | 6,420 | memorandum | 10,403 |
| math/s | 28,248 | mathematics | 60,424 |
| flu | 5,860 | influenza | 4,672 |

Strube and Ponzetto (2006) have demonstrated the effectiveness of Wikipedia as an information source (for calculating degrees of semantic relatedness), by comparing Wikipedia with Google (as a baseline) and with WordNet. They conclude (*ibid*: 1420) that Wikipedia "consistently correlates better with human judgments than a simple baseline based on Google counts, and better than WordNet for some datasets".

Sridharan and Murphy (2012) have compared four corpora against six benchmarks of published behavioural and neuro-activity studies. The four corpora are the Google Web, Google Books, Wikipedia and Twitter. They conclude that "a corpus of high quality at a small size [Wikipedia] can perform better than a corpus of poor quality that is many orders of magnitude larger [the Google Web]. At all corpus sizes up to 1.7 billion five-grams, Wikipedia is the best choice" (Sridharan and Murphy, 2012: 64).

How reliable are the results-counts obtained from Wikipedia? Table 4.3 (with notes 1-5) presents the results-counts from five different methods of searching the English Wikipedia.

**Table 4.3**: Comparative results of phrase-search methods (20 February 2014)

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Phrase** | **Animyser** | **Wikipedia Search** | **Pattern Search** | **Google CSE** | **Google Search** |
| "cat" | 113,604 | 113,604 | 78,700 | 248,000 | 229,000 |
| "the cat" | 10,418 | 10,418 | 9,470 | 27,400 | 27,300 |
| "of the cat" | 1,488 | 1,488 | 2,660 | 132,000 | 108,000 |
| "cat's" | 2,691 | 2,691 | 5,410 | 12,700 | 12,600 |
| "the cat's" | 970 | 970 | 1,560 | 2,990 | 2,980 |
| "of the cat's" | 82 | 82 | 177 | 11,700 | 11,800 |

Notes:

1. Animyser is a Python program that uses the Wikipedia search API within Pattern.web to automate phrase-searches of the English Wikipedia.

2. Wikipedia Search is the manual application of the Wikipedia Special Search facility (http://en.wikipedia.org/wiki/Wikipedia:Special:Search).

3. Pattern Search is a Python program based on the Pattern web-mining module. This program utilises Google's Custom Search Engine (CSE). This and both of the following methods restrict their search to the English Wikipedia site by using the Google site search operator, i.e. by including "site:en.wikipedia.org" (without quotation marks) in the search field.

4. Google CSE is an implementation of Google's Custom Search Engine, set up under licence to this researcher for manual searches of Wikipedia. The count given, for this and for Google Search, is the first-page count.

5. Google Search is the standard web search facility, with site search specified.

The comparison suggests that Wikipedia Special Search (columns 1 and 2) provides the most accurate and consistent counts, whilst confirming that Animyser replicates the manual Wikipedia Special Search.

## 4.4 From Hypothesis to Algorithms

The genitive ratio is a hypothesis with theoretical and empirical support, but it is not (yet) a computational model that will support real-world applications. Those applications address different problems. In chapter 5, the problem is to identify animate nouns. In chapters 6 and 7, the problem is to differentiate concrete (which includes animate) from abstract nouns.

Given the specifics of those problems, a 'one size fits all' approach is unlikely to be as effective as developing problem-specific algorithms that, whilst they share a common foundation of theory and analysis, are more precisely targeted to deliver the optimum performance. Accordingly, two algorithms will be developed: an animateness rating (AR) that will be applied to the problem in chapter 5; and a concreteness rating (CR) that will be applied to the problem in chapters 6 and 7.

## 4.5 Construction of Datasets for Training and Testing

In order to develop the classifier models that will be deployed as the animateness and concreteness ratings, it is necessary to construct independent datasets for training and testing the models.

A test dataset of 450 nouns (Dataset A), annotated with categories and sub-categories, was constructed, initially to facilitate the selection of a classifier. All 450 nouns had a CELEX (Baayen, Piepenbrock and van Rijn, 1993) frequency greater than 100. Given a CELEX frequency range of zero to 35,351, this excluded only very infrequent nouns, though the possibility of a bias cannot be discounted. Nouns within the special case categories (discussed later – see section 4.11) were also excluded. Subsequently, a balanced dataset of 1,000 nouns (Dataset B) was constructed and similarly annotated with categories, in order to train the classifier.

The Animyser program obtained Wikipedia results-counts for the nouns in the two datasets, for each of the six phrase-search templates listed in Table 4.1. Each count was incremented by one, in order to avoid a divide-by-zero error in the ratio analyses.

**Dataset A**. A test dataset of 450 nouns was categorised as animate, concrete or abstract. Nouns in the animate category were then sub-categorised as human, animal or collective. This is the data structure:

| | | | | |
|---|---|---|---|---|
| [4.6] | Animate | 150 | | |
| | | | Human | 50 |
| | | | Animal | 50 |
| | | | Collective | 50 |
| | Concrete | 150 | | |
| | Abstract | 150 | | |

**Dataset B**. The materials for an independent training dataset of 1,000 nouns were initially drawn from Brysbaert, Warriner and Kuperman (2014) Theirs is the most extensive database (40,000 words) of concreteness ratings known to this researcher, with participants' ratings on a continuous scale from most abstract (1.00) to most concrete (5.00). However, it is important to note that their ratings do not distinguish animate from concrete: *baby* and *spaghetti* (for example) are both rated 5.00.

The concreteness ratings in Brysbaert *et al* (2014) enabled a clear differentiation of categories, by selecting the most abstract nouns (rated from 1.00 to 2.50) and the most concrete nouns (rated from 3.50 to 5.00), leaving the central ratings (from 2.51 to 3.49) unselected. In addition, a frequency filter was applied, based on the frequency ratings used by Brysbaert *et al* (2014). These are derived from SUBTLEX$_{US,}$ a database of 51 million words compiled by Brysbaert and New (2009) from a corpus of film and television subtitles, from which they calculated frequency measures for 74,000 words. They contend that their frequency data are more contemporaneous and more representative of "spontaneous" language, than those of Kučera and Francis (1967) or CELEX (Baayen *et al*, 1993). The frequency filter removed words with frequencies of 50 or below and 6,000 and above. Since the Brysbaert *et al* (2014) database includes all parts of speech, non-nouns had to be removed by hand.

Nouns that were already in Dataset A were removed. This yielded 3,511 nouns: 673 coded (by Brysbaert *et al*) as abstract and 2,838 concrete (i.e. all other categories). The 3,511 nouns were hand-coded 1-5 by category, the randomly selected within each category to provide a dataset of 1,000 nouns:

| | | |
|---|---|---|
| 1 | Human | 250 |
| 2 | Animal | 50 |
| 3 | Collective | 50 |
| 4 | Concrete | 350 |
| 5 | Abstract | 300 |

Sub-categorisation of the animate main category was necessary to ensure that the minor categories were adequately represented.

### 4.6 Statistical Test of Animate Sub-categories

From combinations of the six genitive constructions in Table 4.1, 32 different ratios were identified for initial analysis. In order to test whether the three animate sub-categories (human, animal and collective human) were differentiated from each other as well as from the concrete and abstract categories, Dataset A was reconfigured so that each of the now five categories was equal in size (i.e. 50 nouns in each category), the concrete and abstract categories having been reduced from 150 to 50 by taking a random sample.

A one-way ANOVA, with Tukey HSD post-hoc tests, indicated that no ratio would achieve a five-way differentiation. Whilst all three sub-categories could be reliably differentiated from the other two main categories, they could not be differentiated from each other. No comparison of the three sub-categories could achieve a significance better than $p$ = .87. **The conclusion from the statistical analysis is that it is not possible to extend the three-category model to differentiation of the three animate sub-categories (human, animal and collective human).**

The challenge now is to identify the ratio or ratios that will most accurately measure the relative animateness or concreteness or abstractness of a noun or a text. It will subsequently be necessary to define also a method of dealing with exceptional nouns that are not susceptible to genitive ratio analysis (see section 4.11).

## 4.7 Choosing a Classifier

Four machine learning classifiers were tested and compared for accuracy. The Weka machine learning workbench (Witten, Frank and Hall, 2011) provided the classifiers, pre-processing tools and a filtered attribute selection facility. All four classifiers were trained and tested on Dataset A against a common baseline for a three-way classification: animate vs. concrete vs. abstract (3x150).

The factors tested initially were a super-set of those flagged as significant by the Tukey HSD post-hoc statistical analysis. Salient factors for each classifier were identified initially by the filtered attribute selection facility, then by adding and subtracting factors to obtain the best fit to the data. These were the four classifiers:

**Multinomial logistic regression** predicts the outcome of a dependent variable, given a set of independent (predictor) variables. Hinrichs and Szmrecsanyi (2007: 459) claim that regression analysis is "the closest a corpus linguist can come to conducting a controlled experiment: the procedure systematically tests each factor while holding the other factors in the model constant".

**J48** is a Java implementation of the C4.8 decision-tree algorithm devised by Ross Quinlan, which was the last open-source version of his C4.5 algorithm (Quinlan, 1993).

**Naïve Bayes** is a probabilistic classifier, "naïve" because its application of Bayes Theorem is based on an assumption that each input to the model is independent of the others. Nevertheless, Naïve Bayes models generally perform well.

**SMO** is Platt's sequential minimal optimization algorithm for training a support vector classifier (Platt, 1998).

**Zero R** is the baseline classifier that simply assigns the majority category to all instances in the dataset. Since there are three equal categories in the dataset, Zero R simply assigns the same category to all three.

Table 4.4 sets out the metrics of precision, recall and combined *F*-measure obtained from each classifier, against a baseline of 33.3% accuracy (recall) from three equal categories. These results are from ten-fold stratified cross-validation of the dataset of 450 items. Cross-validation (Witten, Frank and Hall, 2011: 151) randomly divides the dataset into (normally ten) equal segments. An iterative process then 'holds out' each of the ten in turn as a test-set, whilst the other nine are used for training the model. The ten results are then averaged.

**Table 4.4**: Results (stratified cross-validation) from four machine learning classifiers, compared with baseline

| **Classifier** | **Precision** | **Recall** | ***F*-measure** |
|---|---|---|---|
| Logistic | 0.732 | 0.739 | 0.731 |
| J48 | 0.726 | 0.728 | 0.726 |
| Naïve Bayes | 0.624 | 0.575 | 0.550 |
| SMO | 0.699 | 0.704 | 0.682 |
| Zero R (baseline) | 0.111 | 0.333 | 0.167 |

Logistic regression provides the most accurate three-way classification: 73.9% compared with a baseline of 33.3%. The animate category is the most accurately classified: 140 out of 150, or 93.3%. The numbers of correct classifications in each category are shown in **bold**:

| N | **Category** | Animate | Concrete | Abstract |
|---|---|---|---|---|
| 150 | Animate | **140** | 9 | 1 |
| 150 | Concrete | 34 | **91** | 25 |
| 150 | Abstract | 6 | 31 | **113** |

The applications in chapters 5-7 demand different binary classifications of the training and test data:

| | | | |
|---|---|---|---|
| animate | vs. | (concrete + abstract) | animateness rating |
| (animate + concrete) | vs. | abstract | concreteness rating |

A regression model quantifies both the extent and the direction of each individual predictor variable in determining the outcome, as well as the sum of their contributions, i.e. how well the model as a whole predicts the outcome. Logistic regression will determine the weightings (the beta coefficients) of those attributes that provide the best-fit set of factors. The sum of the weighted factors plus or minus a constant or intercept value constitutes a binary logit model (BLM). That provides a linear value that can be calculated from the Wikipedia phrase-search results-counts obtained for any noun by the Animyser program.

It is standard practice that the larger dataset (B) should be used for training the model, with the smaller dataset (A) reserved for testing and evaluation (Kotu and Deshpande, 2014: 28). Either dataset would be adequate for purpose: the sample size for a multinomial logistic regression should equal at least ten cases for each independent variable "as a guideline" (Hosmer and Lemeshow, 2000: 347). Both datasets pass that test.

### 4.8 Training and Testing the Animateness Rating

As well as different combinations of ratios, intra-linguistic measures were considered as complements to the genitive ratios. Although only one (word-length in number of letters) was found to be significant, several others were tested. For example, on the basis of a hypothesis that abstract nouns are less likely to be preceded by an adjective, phrase-searches looked for genitive constructions with and without a word between determiner and noun. There was no significant difference between the two conditions. Another hypothesis, that abstract nouns are significantly less likely to occur in plural form, was also unsupported.

The Weka machine learning workbench again provided the pre-processing, logistic regression and attribute selection facilities. Utilising Dataset B, a process of iteratively excluding least significant factors identified the five factors (excluding an intercept value) of a predictive classification model that would offer the best combination of accuracy with parsimony. O'Connor, Maling and Skarabela (2013: 10) have cautioned that the success of a logistic regression model might be over-stated if there is a high degree of correlation between independent variables. There was no evidence of problems with multi-collinearity in the current model. These are the six factors (including the intercept value) and their weightings (coefficients) in the final AR model:

1. Ratio (OD/DS)

   Weighting -0.1715
2. Ratio (OD+OI)/DS

   Weighting +0.1214

3. Ratio OI/(DS+IS)

   Weighting -0.6608

4. IS4

   Weighting +1.1301

   This is the results-count in the indefinite poss-*s* (IS) condition, capped at a value of 4 (which was determined from Dataset B by testing a range of values). A low IS count is a strong indicator of an abstract noun, hence the significance of this predictor.

5. LEN8

   Weighting +0.3058

   This is the length of the noun, i.e. the number of letters, with the maximum length capped at 8 (again determined by testing a range of values). Abstract nouns tend to be longer than concrete nouns: the mean word-lengths in the combined datasets were 6.1 (concrete, including animate) and 7.2 (abstract).

6. Intercept (constant)

   Weighting -4.4873

In its reliance on two capped factors (IS4 and LEN8), the regression model borrows a concept from financial modelling (Fabozzi, 2013: 246-251). The function of capping is to constrain the influence of outliers that would otherwise be created by extreme values of one or more factors.

**Examples**. Two examples (*king* animate and *schooling* abstract) will illustrate the calculation of BLM scores by the AR model. A higher score generally predicts an animate noun. A lower or negative score generally predicts a concrete or abstract noun.

| Weighting | *king* | Weighted Score | *schooling* | Weighted Score |
|---|---|---|---|---|
| -0.1715 | 0.59 | -0.101 | 16.5 | -2.830 |
| +0.1214 | 0.64 | +0.078 | 18.5 | +2.246 |
| -0.6608 | 0.05 | -0.033 | 1 | -0.661 |
| +1.1301 | 4 | +4.520 | 2 | +2.260 |
| +0.3058 | 4 | +1.223 | 8 | +2.446 |
| -4.4873 | 1 | -4.4873 | 1 | -4.4873 |
| **Sum/score** | | **+1.200** | | **-1.026** |

A ten-fold stratified cross-validation test of Dataset B showed that 87.1 % of animate nouns were correctly classified, whilst the overall accuracy (*F*-measure) of the model was 87.7%. Evaluation of the model on Dataset A (450 cases) confirmed the high level of accuracy/recall (96.0%) in classifying animate nouns, with an *F*-measure of 84.2%. Although classification is the most readily-accessible test metric, it is generally more helpful to regard the genitive ratio as a predictor of relative animacy along a gradient rather than in absolute classes (see section 4.10).

### 4.9 Training and Testing the Concreteness Rating

The developmental process of the concreteness rating follows that of the animateness rating, but with the binary classification of the A and B datasets changed to (animate + concrete) vs. abstract. These are the five factors (including the intercept value) and their weightings in the concreteness rating (CR) model:

1. Ratio (OD+OI)/(DS+IS)

   Weighting -0.008

   This might be regarded as the basic genitive ratio (BGR): the ratio of poss-*of* to poss-*s* results-counts in both definite and indefinite conditions. This is capped at a maximum value of 400, as determined by an outlier analysis to exclude the 1% of extreme values over 400 that might otherwise skew the calculation of the mean.

2. Ratio ON/NS

   Weighting -0.024

   This is the ratio of null (no determiner) results-counts, capped at a maximum value of 1,200.

3. IS3

   Weighting +1.48

   This is the results-count in the indefinite poss-*s* (IS) condition, capped at a value of 3 (determined by testing a range of values).

4. LEN8

   Weighting -0.391

5. Intercept (constant)

   Weighting +2.104

**Examples**. The application of the model is illustrated by these three examples:

animate: ***infant***

2.104 - (0.008*BGR) - (0.024*ON/NS) - (0.391*LEN8) + (1.48*IS3)

2.104 - (0.008*2.05) - (0.024*1.48) - (0.391*6) + (1.48*3)

2.104 - 0.0164 - 0.03552 - 2.346 + 4.44

Score = +4.146

concrete: ***jug***

2.104 - (0.008*BGR) - (0.024*ON/NS) - (0.391*LEN8) + (1.48*IS3)

2.104 - (0.008*6.0) - (0.024*2.05) - (0.391*3) + (1.48*1)

2.104 - 0.048 - 0.049 – 1.173 + 1.48

Score = +2.314

abstract: ***diversity***

2.104 - (0.008*BGR) - (0.024*ON/NS) - (0.391*LEN8) + (1.48*IS3)

2.104 - (0.008*272.5) - (0.024*45.52) - (0.391*8) + (1.48*1)

2.104 – 2.18 – 1.0925 – 3.128 + 1.48

Score = -2.817

Evaluation of the model on Dataset A (450 cases) showed a high level of accuracy/recall (99.7%) in classifying animate/concrete nouns, but with an *F*-measure of 68.0%.

### 4.10 Evaluation of Both Models

It has been stressed from the outset of this thesis that "the genitive ratio is a relative rather than a categorical method of noun classification" (1.4). In the applications surveyed in chapters 5-7, it is the ranking of an individual noun by the GR, either relative to or in aggregate with other nouns within a text, that is important.

**Standard performance measures**. The Weka logistic regression output for each model (AR and CR) provides statistics that measure a categorical rather than a gradient classification of the test data, listing for each test item the actual classification, the predicted classification, and the probability of that prediction. Essentially, the trained model predicts for each test item the probability of a 'correct' classification, with an assumption that any test item with a probability of less than 0.5 is 'incorrect'. Three sets of related performance statistics are then provided. Whilst these statistics are relevant and useful, they are not completely adequate for the current context.

First of the given statistics is the correct/incorrect classification of the test data:

| | **AR Model** | | **CR Model** | |
|---|---|---|---|---|
| | No. | % | No. | % |
| Correct | 377 | 84 | 333 | 74 |
| Incorrect | 73 | 16 | 117 | 26 |

Second, a 'confusion matrix' provides a finer-grained analysis:

| | **AR Model** | | | **CR Model** | |
|---|---|---|---|---|---|
| Classified as: | AN | CN+AB | | AN+CN | AB |
| AN<br>(150) | 144 | 6 | AN+CN<br>(300) | 299 | 1 |
| CN+AB<br>(300) | 7 | 233 | AB<br>(150) | 116 | 34 |

Third are standard performance measures of classification:

- *Precision* measures the 'exactness' of a classifier, with a relatively low score indicating the probability of a high number of false positives.
- *Recall* measures the 'completeness' of a classifier, with a relatively low score indicating the probability of a high number of false negatives.
- The *F-measure* (or *F1* Score) is the "harmonic mean of recall/sensitivity and precision: *F*-measure = (2 x Precision x Recall) / (Precision + Recall)"

  (Cios, Pedrycz, Swiniarski and Kurgan, 2007: 480).

**AR Model**

| Category | Precision | Recall | *F*-measure |
|---|---|---|---|
| AN | 0.682 | 0.960 | 0.798 |
| CN+AB | 0.975 | 0.777 | 0.865 |
| Weighted average | 0.877 | 0.838 | 0.842 |

**CR Model**

| Category | Precision | Recall | *F*-measure |
|---|---|---|---|
| AN+CN | 0.720 | 0.997 | 0.836 |
| AB | 0.971 | 0.227 | 0.368 |
| Weighted average | 0.804 | 0.740 | 0.680 |

In summary, these standard performance measures from the test Dataset A (*N*=450) indicate very high levels of recall for both the AR and CR models' primary categories (AN 0.973 and AN+CN 0.997 respectively), but with low recall (particularly in the CR model) for the secondary categories.

**Supplementary analysis**. Whilst the provision of standard measures is useful in comparing the performance of different models, they rely upon a probability calculation with a 'cut-off' of 0.5 that assumes a binary/categorical classification. This supplementary analysis will now address two questions. First, are the aggregated scores (AN+CN vs. AB) produced by the CR model sufficiently differentiated to support the analysis required by the applications in chapters 6 and 7? Second, do the recall scores adequately represent both models' accuracy?

In answer to the first question, a *t*-test of the CR model indicates a significant difference between the two categories (AN+CN vs AB) as represented by the test data (*N* = 450):

| | | | |
|---|---|---|---|
| Mean | AN+CN | (300) | 3.943 |
| | AB | (150) | 1.260 |
| SD | AN+CN | (300) | 0.846 |
| | AB | (150) | 1.029 |
| *t* (448) | | | 15.925 |
| *p* | | | <.0001 |
| *g** | | | 2.945 |

*Hedges' *g* is a measure of effect size appropriate to unequal samples. 2.945 is a "strong effect" (Ferguson, 2009: 533).

**A ranking test**. In both models, the GR scores should generally rank the test items from animate (high) to abstract (low or negative). A simple test of the models' accuracy is therefore to rank the 450 test items by their GR scores, high to low. In a perfect world, the top-ranked 300 scores in the CR model would all have been classified in the test data as animate or concrete (AN+CN). The percentage of items correctly classified therefore provides a measure of performance:

**AR Model**

| Category | *N* | Correct | % |
|---|---|---|---|
| AN | 150 | 146 | 97.3 |
| CN+AB | 300 | 222 | 74.0 |
| Total | 450 | 388 | **81.8** |

**CR Model**

| Category | *N* | Correct | % |
|---|---|---|---|
| AN+CN | 300 | 270 | 90.0 |
| AB | 150 | 119 | 79.3 |
| Total | 450 | 450 | **86.4** |

### 4.11 Pre-Processing of Special Cases

Three categories of nouns are not susceptible to the genitive ratio method: temporal nouns, plural-only nouns and singular plurals.

**Temporal nouns**. In 2006, an analysis of the Oxford English Corpus found that three temporal nouns ('time words') were ranked in the 25 most frequent nouns: *year*, *day* and *week*. The previous chapter (s3.7) presented an analysis of temporal nouns. The conclusion is reproduced here for convenience:

> "Genitive constructions of *time* follow pragmatic conventions unrelated to animacy... Because the vocabulary of *time* is limited and well-defined, it should be possible to treat *time* phrases as a special case in any computational analysis, by reference to a look-up table of *time* words".

**Plural-only nouns and singular plurals**. Given a plural noun, Animyser will convert it to its singular form, but there are plural-only nouns that end in –*s* and thus confound the simplistic singularizer module imported from Pattern.

Huddleston and Pullum (2002: 341-345) call these nouns "bipartites", because they most often signify objects that consist of two parts, such as items of clothing (*jeans, pants*, *pyjamas*), and tools (*pliers*, *shears*, *secateurs*). They are designated by the Concise Oxford Dictionary as "pl.n" (plural noun). Singular plurals are nouns that appear to be plural but are treated grammatically as singular, e.g. *billiards*, *mathematics* (Huddleston and Pullum, 2002: 345-348). The Concise Oxford Dictionary designates them as either "usually treated as singular" or "plural same".

**Pre-processing**. The treatment of these special cases will rely on Brysbaert, Warriner and Kuperman (2014) as an independent source of ratings for concreteness (see chapter 2.8). The Animyser program contains dictionaries for temporal and other exceptional nouns, with each dictionary entry as in this example (using the Python notation):

```
{"century" : 2.83}
```

The key (the noun *century*) is 'defined by' its value (its Brysbaert *et al* concreteness rating of 2.83). Table 4.5 lists the current dictionary content, by category.

If the target noun matches a noun in the dictionary, the program converts the concreteness rating into a BLM score. The Brysbaert *et al* ratings are specified to two decimal places, with minimum and maximum values of 1.00 and 5.00, hence a span of 400 data-points and a median of 3.00. The conversion formula is therefore:

(Brysbaert rating-1.00) / 4.00

So, for example, *species* with a concreteness rating of 3.36 would be assigned a CR score of (3.36 - 1) / 4 = 0.59. Whilst this approximation is inferior to a GR rating, it does reflect the relative concreteness of the noun. Nouns for which Brysbaert *et al* do not provide a rating (e.g. *arrears, bellows, doldrums, secateurs* in Table 4.5) have been assigned the median rating of **3.00** (in bold).

**Table 4.5**: Words that are not susceptible to the genitive ratio method, with concreteness ratings from Brysbaert *et al* (2014)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Temporal** | | moment | 1.61 | earnings | 4.39 | waterworks | 4.07 |
| Sunday | 3.32 | month | 4.20 | evens | **3.00** | **Singular** | |
| Monday | 2.25 | week | 3.48 | forceps | 4.79 | **Plurals** | |
| Tuesday | 3.29 | morning | 3.44 | gasworks | **3.00** | acoustics | 3.12 |
| Wednesday | 3.14 | night | 4.52 | genitals | 4.96 | acrobatics | 4.36 |
| Thursday | 3.03 | daytime | 4.31 | goggles | 4.93 | athletics | 3.65 |
| Friday | 3.28 | nighttime | 3.86 | goods | 4.26 | barracks | **3.00** |
| Saturday | 3.07 | season | 3.32 | grassroots | 2.63 | billiards | 4.61 |
| January | 3.13 | year | 3.25 | hustings | **3.00** | ceramics | 4.62 |
| February | 2.40 | nightfall | 3.79 | jeans | 5.00 | crossroads | 4.67 |
| March | 4.03 | lunchtime | 3.79 | knickers | 4.54 | diabetes | 3.83 |
| April | 3.33 | dinnertime | 3.62 | molasses | 4.84 | economics | 1.77 |
| May | 3.00 | dusk | 4.24 | munitions | 3.82 | forensics | 2.13 |
| June | 3.40 | daybreak | 4.21 | odds | 2.24 | headquarters | 4.14 |
| July | 2.93 | noon | 2.67 | panties | 4.90 | innings | 3.41 |
| August | 3.04 | midnight | 3.14 | pants | 4.86 | linguistics | 2.63 |
| September | 3.81 | instant | 2.70 | peelings | 4.27 | electronics | 4.37 |
| October | 2.81 | twilight | 4.11 | pliers | 4.93 | gallows | 4.35 |
| November | 2.93 | yesterday | 3.00 | proceedings | 2.89 | gymnastics | 4.04 |
| December | 3.00 | tonight | 2.93 | proceeds | 2.59 | mathematics | 2.52 |
| Spring | 3.89 | today | 2.57 | pyjamas | 4.73 | measles | 4.69 |
| Summer | 3.64 | tomorrow | 2.04 | remains | 3.46 | mews | **3.00** |
| Autumn | 3.27 | past | 1.70 | shears | 4.61 | mumps | 4.10 |
| Winter | 3.84 | future | 1.86 | shorts | 4.82 | news | 3.41 |
| Christmas | 3.41 | **Plural Only** | | specifics | 1.97 | particulars | 1.50 |
| Easter | 2.83 | alms | **3.00** | surroundings | 3.88 | phonetics | 2.38 |
| afternoon | 3.70 | amends | 1.92 | thanks | 2.15 | physics | 3.07 |
| century | 2.83 | arrears | **3.00** | tights | 4.62 | politics | 2.66 |
| day | 3.92 | bellows | **3.00** | tongs | 5.00 | rabies | 3.83 |
| decade | 3.19 | belongings | 3.88 | scissors | 4.85 | rickets | **3.00** |
| second | 3.30 | binoculars | 5.00 | secateurs | **3.00** | semantics | 1.70 |
| evening | 3.26 | cahoots | 1.96 | slacks | 4.57 | series | 2.92 |
| fortnight | 2.71 | cattle | 4.64 | suds | 4.59 | species | 3.36 |
| hour | 3.10 | clothes | 4.76 | trousers | 4.93 | summons | 2.97 |
| weekend | 3.83 | doldrums | **3.00** | tweezers | 4.96 | | |
| millennium | 2.63 | dregs | 3.60 | underclothes | 4.66 | | |
| minute | 3.04 | droppings | 4.48 | valuables | 3.07 | | |

**'No score' default**. If all of the actual results-counts for a noun are zero (augmented to one by the Animyser program, to avoid 'divide by zero' errors), then the program omits that item from the analysis. This discounts words that have either been mis-spelt or (more probably) are very infrequent in any genitive construction: actual incidences, encountered in the analysis of authors (chapter 6), are *footmark* and *impassivity*.

### 4.12 Limitations of Phrase-search Analysis

Player: We are tied down to a language which makes up in obscurity what it lacks in style.
Tom Stoppard: *Rosencrantz and Guildenstern are Dead*

Calculations of the genitive ratio based on phrase-searches of Wikipedia (or of any other unsupervised corpus) will be neither definitive nor exact**.** There are inherent limitations of phrase-search that constrain the analysis of genitive constructions. For these exceptions, there are no ready solutions. They are all the product of natural language. Application-based tests of the genitive ratio's reliability will follow in chapters 5-7. The exceptions discussed here simply explain why the reliability of phrase-search analysis can never be 100 percent, although their actual impact is arguably less than might be expected.

**Auxiliary verb contractions**. Phrase-search will not discriminate between the poss-*s* genitive and the –'*s* auxiliary verb contraction (of *is* or *has*). For example:

[4.6] This drink's main ingredient is soda [possessive]

This drink's not fit for human consumption [auxiliary *is*]

This drink's gone flat [auxiliary *has*]

Although this is potentially a problem for the genitive ratio, auxiliary verb contractions should be moderated by the formal language of Wikipedia articles (see Table 4.2). To test that assumption, ten high-frequency animate and inanimate nouns were submitted to Wikipedia Special Search in the null-'*s* (NS) phrase-search format. The first 50 results for each noun were then analysed. The data in the right-hand column of Table 4.6 show the percentage of auxiliary verb contractions identified.

**Table 4.6**: Auxiliary verb contractions that would be counted as *s*-genitives

| **Test Word** | **NS Score** | **Auxiliary** |
|---|---:|:---:|
| | **May 2014** | **Verb -'*s* %** |
| child | 8,555 | 0 |
| woman | 15,577 | 0 |
| government | 22,682 | 0 |
| student | 5,345 | 0 |
| family | 21,184 | 6 |
| hand | 396 | 0 |
| eye | 530 | 0 |
| head | 656 | 4 |
| door | 179 | 2 |
| word | 19 | 0 |

**Partitives and pseudo-partitives.** Partitive constructions consist of a quantifier followed by an *of*-phrase, e.g.

[4.7] Most of the boys
Two of the chairs

Eligible quantifiers listed by Huddleston and Pullum (2002: 539 [31] and [33]) include *all*, *both*, *certain*, *several*, as well as cardinal numbers. Rutkowski (2007: 337) cites Koptjevskaya-Tamm's (2001) definition of partitives as "a part/subset of a (definite) superset", and of pseudo-partitives as "expressions referring to an amount/quantity of some (indefinite) substance".

Pseudo-partitives are often conventionalised phrases in which one noun acts as a measure of another noun, e.g.

[4.8] A cup of tea
A slice of bread

Both partitive and pseudo-partitive constructions potentially skew the results of a genitive phrase-search by increasing the poss-*of* count, because in neither case is there a poss-*s* alternative, e.g.

[4.9] *The chairs' two
*A cup's tea

The examples in [4.7] suggest that partitive *of*-phrases are mostly plural (an exception would be a fractional quantifier, e.g. *half* of the house), and the pseudo-partitive examples in [4.8] suggest that most are null-determiner phrases (e.g. *a side of bacon*). Since the Animyser program collects only singular genitive constructions, the potential for a 'partitive confound' is mitigated, if not eliminated.

**Variant spellings**. The principal language of the English Wikipedia is American English. Whilst there are variant (i.e. equally acceptable) spellings in British English, such as *judgment* and *judgement,* the main source of variation in Wikipedia is between British and American English. These are examples, with their respective Wikipedia frequencies:

| [4.10] | British | | American | |
|---|---|---|---|---|
| | *colour* | 63,131 | *color* | 151,519 |
| | *centre* | 293,871 | *center* | 467,616 |
| | *catalogue* | 60,373 | *catalog* | 77,449 |

Variant spellings are only a problem for the genitive ratio if the frequency of the word-form within Wikipedia is very low. The frequencies in [4.10] would be adequate for a GR analysis, in either spelling.

**The third apostrophe**. The poss-*s* form of singular English nouns that end in 's' selects either an apostrophe-*s* or simply an apostrophe. It seems probable that the selection is biased by prosody, in that *Thomas's* is easier to articulate than *Moses's*, for example. The poss-*s* form of *Moses* would normally be *Moses'*, but neither Google Search nor Wikipedia Special Search (both enclosed in double inverted commas) would recognise that form. This is because both treat a phrase-search ending in three inverted commas ("*Moses*'") as if it ended in just two inverted commas: phrase-search does not recognise a 'third apostrophe'. The same constraint would apply to the much more frequent poss-*s* plural, which is most regularly formed by adding an apostrophe after the plural-*s* ending (e.g. "*lasses*'"), except that the Animyser pre-processing 'singularizes' target nouns before constructing the phrase-search templates.

**Syntactic polysemy**. A noun adjunct fills the syntactic role of an adjective, by modifying another (head-) noun. For example, the phrase-search "of the baby" will count *baby* in its noun adjunct role as well as in its role as a head-noun:

[4.11] The weight of the baby elephant was 400 kilos [adjunct]
The weight of the baby was four kilos [head-noun]

Because Wikipedia is an unsupervised corpus, situations arise where the target word has been correctly identified (by the POS tagger) as a noun, but most of the phrases harvested by Animyser contain the target word in a different syntactic category, most commonly as an adjective. High-frequency examples of such words are *kind* and *cold*:

[4.12] It was nothing of the kind [noun target]
It was the act of the kind man [adjectival hit]
[4.13] The fever was the result of a cold [noun target]
It was the end of a cold day [adjectival hit]

A polysemous word such as *major* will generate outlier results-counts, since it carries head-noun, adjunct, proper noun, and adjectival senses:

[4.14] The valour of the major deserved a VC [head-noun]
It is the site of the Major Smith memorial [adjunct]
It was the end of the Major premiership [proper noun]
This was the site of the major battle [adjective]

Other outliers encountered in the course of analysis have been *polish* (Polish), *silver*, *emergency*, *health* and *peace*. These are all infrequent as poss-*s* nouns, but much more frequent as adjectives or adjuncts: *Polish government*, *silver coin*, *emergency action*, *health centre*, *peace conference*.

**Non-possessive *of*-phrases**. These are also *of*-constructions with no *–'s* counterpart. As the post-head complement of an adjective or adverb phrase, a preposition phrase (PP) with 'of' corresponds to the poss-*of* form and as such will be identified by Wikipedia Special Search as a genitive construction. Adjectival examples are:

[4.15] She was mindful of the risk

He was afraid of a repetition

(See Huddleston and Pullum, 2002: 542-544 for a full discussion and a comprehensive list of adjectives that take the preposition 'of'.) Some adverb phrases similarly license a PP complement, e.g.

[4.16] They acted independently of the government

(*ibid*: 571) and *of*-PPs can form the complement of a verb:

[4.17] He died of a heart attack

(*ibid*: 731).

With some differences in usage between British and American English, a number of prepositions license 'of' as the head of their complement. Huddleston and Pullum (2002: 639) give examples, such as:

[4.18] because of  ahead of  instead of

**Compound nouns**. Wikipedia phrase-search distinguishes hyphenated from unhyphenated compound nouns, with their frequencies varying accordingly:

| | | | | |
|---|---|---|---|---|
| [4.19] | *flower-head* | 57 | *flowerhead* | 370 |
| | *summer-house* | 58 | *summerhouse* | 514 |

The Animyser POS tagger processes unhyphenated compound nouns separately, so that an input text containing *cottage door* and *death certificate* would trigger separate searches for *cottage*, *door*, *death* and *certificate*. However, when the same compounds are encountered in the Wikipedia corpus by the poss-*of* phrase-searches, the noun modifier of the phrase will be counted, but not the head-noun:

[4.20] of the cottage [noun modifier] door [head-noun]

of the death [noun modifier] certificate [head-noun]

**The 'big data' trade-off**. A supervised, parsed and annotated source-corpus might identify and resolve these exceptions. Wikipedia has scope, scale and diversity of vocabulary. This is the essential trade-off underlying the argument for 'big data': the method might be sub-optimal, "but the vast amount of data … more than compensate[s] for the imperfections" (Mayer-Schonberger and Cukier, 2013: 187).

A "tagged and cleaned Wikipedia" (Artiles and Sekine, 2009) has been made available, based on a download of the English Wikipedia in June 2008. This could be utilised in future work, but the authors acknowledge that the files are "not tagged 100% accurately and are not 100% cleaned", as well as omitting Wikipedia user discussions that were included in the current analysis.

### 4.13 Caveats and Conclusions

The chapter overview (section 4.0) posed three questions, to which the answers are:

- The three-category model is robust
- Specific factors have been identified as components of ratings for animateness and concreteness

- The current model will not differentiate the animate sub-categories (though future research should not rule out the possibility)

This chapter is the 'pivot' of the thesis. The previous chapter built a case for the genitive ratio, based on linguistic theory and analysis of an independently annotated database of genitive constructions. The products of this chapter are two computational models developed from that analysis. In the next chapter we will test the animateness rating, at the level of an individual noun.

The achievement of Brysbaert *et al* (2014), in gathering reliable concreteness ratings for nearly 40,000 words, is all the greater when set against the comparatively limited datasets of previous rating studies, but it is still 'just' 40,000 words, only a subset of which are nouns. By contrast, the genitive ratio method is able to access the entire vocabulary of Wikipedia, affording significant advantages of scale and contemporaneity.

Reliance on Wikipedia as the corpus of choice might in future be mitigated by gaining direct access (i.e. not via their public search engines) to the databases of Google Search or Microsoft's Bing; or by the further development of "very large linguistically processed web-crawled corpora" such as ukWaC (Baroni *et al*, 2009) that might provide the benefit of scale whilst capturing the syntactic differences (e.g. auxiliary-*s* vs. poss-*s*) that have eluded the current analysis. The GR models presented in this chapter are not definitive, but they perhaps provide a template for the development by future researchers of algorithms with features engineered towards a specific problem.

The genitive ratio is not a perfect classifier. Its primary role is as a measure of relativity in a gradient from animate to abstract. If the foregoing discussion of the model's limitations seemed rather negative, we should not lose sight of the fact that those are exceptions, that the model is sufficiently robust and serviceable to have potential applications in several fields. The following chapters will demonstrate some of those applications.

# CHAPTER 5

# The Animate Language of Salience

*The test of all knowledge is experiment. Experiment is the sole judge of scientific 'truth.*
Richard Feynman

## 5.0 Overview

The central question addressed by this chapter is: will the genitive ratio model, in the form of an animateness rating (AR), reliably identify the animate nouns within a text?

There is a consensus in the relevant literature that animacy is a determinant (probably the principal determinant) of discourse salience: the more animate an entity is, the more prominent or focused its referent is likely to be in the mental model of a discourse, and the more likely it is to be the co-referent of a subsequent pronoun or other referring expression. Computational systems for co-reference resolution have evolved from models based on syntactic ranking and weighted factors, to more complex modular architectures, which might in future accommodate a genitive ratio module.

An innovative online sentence production experiment will link the AR to a test of animacy as a determinant of salience. That experiment will have two other purposes. It will provide further evidence that proper nouns have a particular salience and do not conform to the genitive ratio model. It will also provide a benchmark for a replication of the experiment in Italian, to test whether the genitive ratio might be applied to languages that lack a possessive-*s* construction, by using English as a 'bridge'.

A further experiment will simulate the real-world application of the AR, by testing co-referent sentences extracted from a corpus of blog postings.

### 5.1 Why Salience Matters

The most salient entity (or set of entities) in a discourse is defined as that which constitutes the centre of attention at that particular juncture of the discourse. (For the purpose of the ensuing discussion, 'discourse' encompasses both written and spoken language).

The experiments reported in this chapter rest upon two assumptions. The first is that animateness is a key determinant of salience. The second is that the most salient entity in a sentence will most often occupy the initial position in that sentence, usually as its grammatical subject. Both assumptions are supported by a significant body of prior research.

Before we consider the results of those experiments, we need to reach at least a broad understanding of what makes an entity salient, and why salience matters. Almor and Nair (2007) provide a comprehensive overview of discourse salience theories. Whilst different academic disciplines have proposed their own theories and models of salience, and in their own terminology, there has been significant cross-disciplinary influence, particularly between computational and psychological models. What motivates these theories is the problem of anaphora resolution.

Anaphors are "words or phrases which relate new information to ideas or objects that have been mentioned previously" (Coulson, 1995: 93). Anaphora resolution is the linguistic mechanism that bridges sentence boundaries in order to maintain the coherence of a discourse. Specifically, it is a problem of how to resolve what a noun phrase (NP) or a pronoun is referring to, what is its *co-referent*. Theories that attempt to track co-references through the sequence of clauses or sentences in a discourse are motivated by potential applications of automated anaphora resolution in machine translation, text summarisation and question answering (Mitkov, 2003).

Mitkov (2002: 33-34) has observed that "anaphora resolution offers an ideal illustration of the complexity of natural language understanding … Many real-life examples of anaphors require world knowledge for their resolution". One such example is presented by Winograd (1972: 33), who asks us to consider these two sentences (co-referents are underlined):

[5.1] The city councilmen refused the demonstrators a permit because they feared violence

[5.2] The city councilmen refused the demonstrators a permit because they advocated revolution

Winograd (*ibid.*) concludes that "no set of syntactic or semantic rules could interpret this pronoun reference without using knowledge of the world".

The genitive ratio will not solve the problem of co-reference, but it might complement some of the established computational algorithms.

## 5.2 Computational Algorithms of Co-reference

"A major component of any discourse algorithm is the prediction of which entities are salient" (Walker, 1989: 254).

The key difference between a psychologically plausible model and a computational model of salience is that the former can rely upon world knowledge and commonsense inference, whereas the latter - in all practical respects - cannot.

Kaiser (2006: 150) concludes that salience is determined by a "competition-based system sensitive to multiple factors which are weighted differently". That description relates equally to psychological and computational models of salience. A computational "stack" of potential co-referents finds its psychological analogue in the contents of working memory.

In a computational model of salience, the availability of previous referents is potentially unlimited, with all referents equally activated, but that is not psychologically plausible. A psychological model assumes that entities within working memory (WM) are immediately accessible, but must accept that WM capacity is constrained (Baddeley, 1986), with significant individual differences that affect both syntactic and semantic processing (Traxler *et al*, 2005: Experiment 2). Kibrik (1999: 48) concludes that WM for discourse typically has a capacity of "three strongly activated referents" - it is not just the number of salient referents, but their degree of activation, that fills the WM buffer.

Many of the claims made about salience factors in discourse are actuarial – based on the probability (weighting) of an event's occurrence given a particular set of circumstances (factors). Stevenson (2002: 373-376) offers a series of tests as a "vote-winner" (essentially a weighting) approach. The entity with the most votes is the winner, i.e. most salient. Stevenson acknowledges that her psycholinguistic tests might better be implemented as an activation-based or probabilistic rules-based system.

Ng (2002: 10) has noted that the trend in research on co-reference resolution has been towards knowledge-lean approaches; away from the utilisation of hand-coded heuristics and towards probabilistic corpus-based approaches that rely on classification and machine learning. This trend is illustrated by a review of influential process models and algorithms.

**Hobbs (1978).** Hobbs describes a "naïve" algorithm that searches through the "surface parse tree" of the sentences in a text, first to identify antecedent noun phrases of the appropriate gender and number, then to rank them according to their syntactic "focus": subject, then object, then the nominal component of a preposition phrase, for example. In effect, the algorithm proceeds through a series of syntactic filters.

In a performance comparison of pronoun resolution algorithms by Tetreault (2001: 515, Tables 2 and 3), Hobbs' (1978) algorithm performed well against later efforts such as BFP (Brennan, Friedman and Pollard, 1987) and S-list (Strube, 1998). Dagan and Itai (1991) enhanced Hobbs' algorithm with a statistical pre-analysis of all co-occurrences in the target discourse, for example how frequently specific entities occurred as either the subject or object of specific verbs.

**Lappin and Leass (1994).** The RAP (Resolution of Anaphora Procedure) algorithm developed by Lappin and Leass (1994) "relies on measures of salience derived from syntactic structure and a simple dynamic model of attentional state to select the antecedent noun phrase (NP) of a pronoun from a list of candidates" (*ibid*: 535). The  algorithm calculates a salience weighting for candidate antecedents, from a set of salience factors. Examples of these are set out in Table 5.1, with their initial weights.

**Table 5.1**: Salience factors (Lappin and Leass, 1994: 541)

| **Factor Type** | **Initial Weight** |
|---|---|
| Sentence recency (current sentence) | 100 |
| Subject emphasis | 80 |
| Accusative emphasis (direct object) | 50 |
| Indirect object and oblique complement emphasis | 40 |

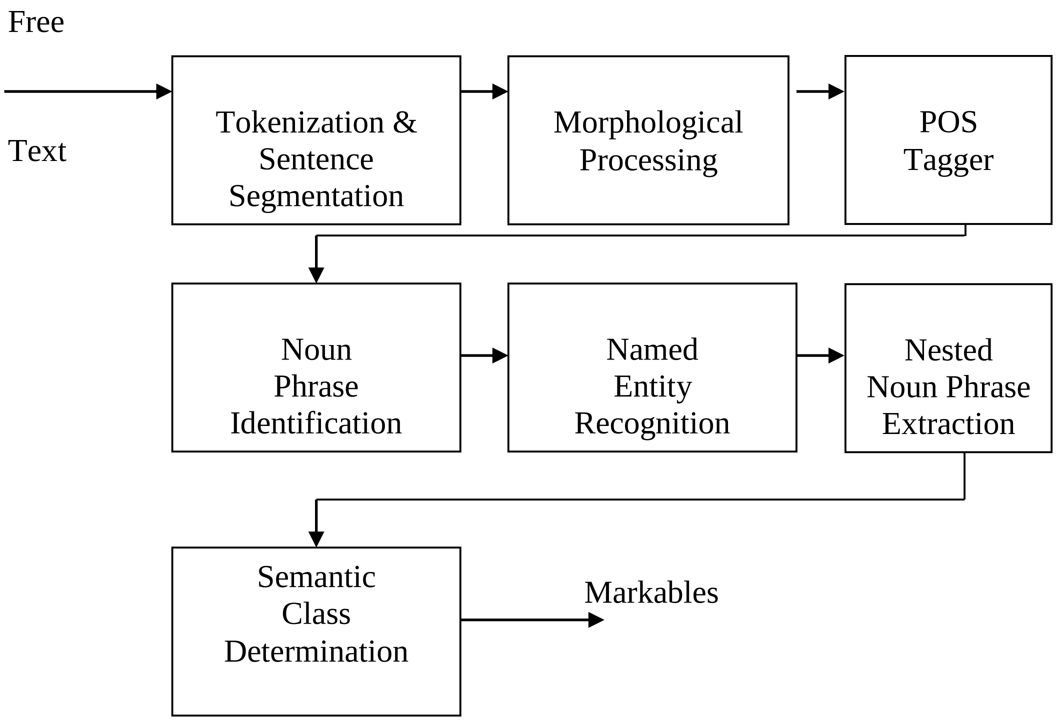
The Lappin and Leass algorithm incorporates neither semantic (e.g. animacy) nor real-world knowledge constraints, but still achieved an 86% success rate, albeit when applied to the specialist domain of computer manuals.

**MARS (Mitkov, Evans and Orăsan, 2002).** Mitkov (2000) has advocated what he calls a "practical" alternative to the "traditional" methods of anaphora resolution that have relied on linguistic (particularly syntactic) analysis and on knowledge of at least a specific domain. Mitkov's alternative approach relies on a combination of heuristics and statistical measures, thus attempting to eliminate both the need for domain knowledge (which is expensive to implement) and pre-processing by a parser - though the approach does employ a part-of-speech tagger and "simple noun-phrase rules".

MARS (Mitkov's Anaphora Resolution System) applies weights to 14 tests of salience for each candidate co-referent of a pronoun, in order to select the most likely NP. The weights "were decided in accordance with corpus observation and certain tenets of centering theory on an intuitive manual basis" (Evans, 2002: 2). A branching algorithm creates a separate branch for each different type of pronoun. A genetic algorithm finds the combination of salience factors and weights that optimises the resolution of that pronoun with its correct antecedent.

Training and testing of the genetic algorithm relies on an annotated corpus consisting of eight computer manuals and the annual report of Amnesty International, a total of 263,168 words (Evans, 2002). Within the 'sublanguage' of computer and technical manuals (which surely constitute a "specific domain"), Mitkov claims a success rate of 85% of anaphors correctly resolved. However, with (for example) just two occurrences of the pronoun *him*, Evans (2002) has conceded that the corpus needs to be at least 100 times larger, which would represent a prohibitive cost of annotation.

**Soon, Ng and Lim (2001)**. The MARS algorithm (Mitkov, Evans and Orăsan, 2002: 182) is dependent upon a POS tagger (judged to be more reliable than a parser) and effective pre-processes for number agreement; animacy determination (Orăsan and Evans, 2001); named entity recognition; and gender agreement (based on a "proper names list"). Any process of co-reference resolution also requires a method of identifying both anaphoric and *non*-anaphoric NPs, i.e. which NPs actually need to be resolved (Ng, 2010: 1401). A typical system architecture of a natural language processing 'pipeline' is that of Soon, Ng and Lim (2001: 522):


Free
Text
Tokenization & Sentence Segmentation
Morphological Processing
POS Tagger
Noun Phrase Identification
Named Entity Recognition
Nested Noun Phrase Extraction
Semantic Class Determination
Markables

Soon *et al*'s pipeline (pre-processor) designates "markables" (annotated entities) that are possible co-referents. Co-reference resolution is then achieved by training a classifier, based on the C5 decision-tree learning algorithm (Quinlan, 1993), on a subset of a corpus before applying it to the test documents.

The training input to their classifier is a vector of 12 features for every possible pairing of potentially co-referent entities (Soon, Ng and Lim, 2001: 522-530). The authors conclude that "a learning approach using relatively shallow features can achieve scores comparable to those of systems built using nonlearning approaches" (Soon, Ng and Lim, 2001: 532).

**BART (Versley et al, 2008).** BART (the Beautiful Anaphora Resolution Toolkit) is a set of tools for co-reference resolution, assembled from many different sources. BART's system architecture comprises four principal sets of modules:

**Pre-processing** comprises named entity recognition, "chunking" of noun phrases, POS tagging and syntactic analysis, and delivers an inventory of "mention objects" that incorporates type, number and gender.

**Feature Extraction** creates a PairInstance object for every possible combination of an anaphor and any previously-mentioned antecedent. These objects are enriched by feature sets drawn from an XML description file.

**Learning** assigns the pre-processed and feature-enriched data to one of several machine learning toolkits capable of co-reference resolution.

**Training and Testing** are achieved by an encoder/decoder component which interacts with the machine learning system.

BART's structure is highly modular, and therefore able to absorb new methods and new information sources with relative ease. It also escapes the over-reliance on fixed salience weightings that constrain its competitor algorithms. A modular architecture such as BART might well accommodate a genitive ratio module, as an additional constraint factor.

### 5.3 Perspectives on Salience

**Mental models.** Garnham (1987: 152) defines a mental model as a "mental representation of part of the real or imaginary world". Psychologists generally attribute mental models theory to Johnson-Laird (1983), though he himself dates the concept back to Craik (1943). In Johnson-Laird's account, a discourse is represented, beyond its physical realisation as a sequence of phonemes or graphemes, at both a propositional level (capturing its sense) and more fundamentally as a mental model (capturing its significance). The construction of a mental model immediately follows the (relatively transient) surface encoding of a discourse.

Oakhill, Garnham and Vonk (1989: 264) list three crucial assumptions of mental models theory. Mental models are:

- Real- (or imaginary-) world, rather than linguistic, representations.
- Constructed dynamically and incrementally.
- Limited in their sets of possible tokens (and hence referents).

Glenberg and Langston (1992: 147) define mental models as "constructions in working memory", with the capacity constraint that implies. As we endeavour to comprehend a discourse, these non-linguistic models are activated spontaneously, using processes that are stored in semantic memory.

The activation-based account of a discourse is conceptually quite simple (Grüning and Kibrik, 2005: 166-168). At any stage in a discourse, a number of referents are held in the working memory (WM) of the discourse participants. Each referent has a level of activation, determined by *activation factors* that are either intrinsic to the referent (e.g. its degree of animacy) or extrinsic (e.g. deriving from the referent's role and/or position in the structure of the discourse).

There is an inverse relationship between a referent's degree of activation and the likely semantic content of a subsequent anaphor: low activation makes a full NP more likely; high activation will bias a pronominal anaphor. The *threshold* of activation that determines selection of (say) a full NP versus a pronoun is not fixed, but is relative to the activation levels of other referents within the discourse model in WM. At any point in time the entity that is most prominent is the most salient referent, and its relative salience determines its linguistic representation. For example, an indefinite description carries the implicit assumption that a new entity is being introduced to the mental model. A pronoun or definite description carries the implicit assumption of reference to an entity already represented in the mental model, e.g.

[5.3] Ian picked up a fine [new] when it/the parking ticket [established] expired.

The mental model concept has been particularly influential, possibly because it provides a relatively simple paradigm of the cognitive structure that finds its linguistic realisation in a discourse. Garnham (2001: 36) points out that both Sanford and Garrod's (1998) *scenario mapping and focus* approach and Gernsbacher's (1997) *structure building framework* have "much in common with mental models theory", as do the *situation models* of Van Dijk and Kintsch (1983). The entities represented in each of these 'cognitive maps' acquire and then lose their relative prominence as the discourse progresses.

**Gestalt psychology**. Osgood and Bock (1977: 90) propose three "production principles" that affect the relative salience of an entity: Naturalness, Vividness, and Motivation-of-Speaker. *Naturalness* expresses how clearly the linguistic cognition of an entity wholly reflects its pre-linguistic perceptual cognition. Gestalt perceptual concepts, of the dominant *figure* set against the more general *ground*, are thus assigned linguistic equivalents. *Vividness* expresses the "affective intensity" that derives from an entity's semantic features: the specificity of *vampire* compared with the generality of *man* is the example given. *Motivation-of-Speaker* expresses the idea that the speaker's own focus will affect an entity's salience and hence its (earlier) position in a sentence-production.

In an implicit reference to gestalt psychology, Levelt (1989: 266-267) relates the salience or "foregrounding" of a discourse entity to the degree of "human interest". Levelt cites as empirical evidence experiments that demonstrate the effects of animacy as well as of foregrounding: experiments by Osgood (1971), Osgood and Bock (1977), Flores d'Arcais (1987) and, in cross-linguistic studies, by Sridhar (1988).

**Pragmatics**. From the perspective of conversational pragmatics, Smith, Jucker and Müller (2000) propose a "scale of salience" with three levels, described within a theatrical metaphor. At the highest level are the "star" entities. Their role is mutually accepted and understood, and they remain immediately accessible to both speaker and listener for as long as they remain the focus of the conversation's main "plot". At the second level are the "supporting cast" of entities that need to be individuated and accessible only for as long as their particular sub-plot is important to the conversation. In the background, at the third level, are the "stage props", generic entities that give a context to the plot.

**Accessibility theory**. Ariel (1988) explains co-reference in terms of "accessibility", which is determined primarily by distance (between antecedent and anaphor) and by the salience (or "topicality") of the antecedent. Thus, a pronoun is preferred when the antecedent is highly accessible, whereas a less accessible antecedent is more helpfully recalled by a definite description. Ariel contends that this finding is broadly consistent across most languages.

We might think of a sentence as having an underlying conceptual network, formed by the semantic relationships between the lemmas (the base forms) of the words in that sentence. The richness or otherwise of that network derives from the extent of the possible relationships, which in turn is a product of the lemmas' semantic features. Bock and colleagues (Bock and Warren, 1985; McDonald, Bock and Kelly, 1993) argue that it is the "centrality" of animate (and concrete) concepts within such networks that is the basis of their accessibility (i.e. their salience). Bock and Warren (1985) characterised this accessibility in terms of the relative number of predicates to which an entity might be related (its "predicability").

Branigan, Pickering and Tanaka (2008: 174) observe that "highly predicable entities tend to be both concrete and prototypical". They are also likely to have been acquired at a relatively early stage of language development (Keil, 1979). The thesis proposed by Branigan *et al* (2008) is that animacy is "one of a constellation of graded conceptual features that contribute to an overall index of conceptual accessibility" (*ibid*: 187).

## 5.4 Thematic Roles

Stevenson (2002: 370-373) has concluded that salience is the product of three factors: animacy, thematic role, and recency, but that "both recency and thematic role seem to depend on animacy".

The semantic content of a sentence or clause is usually termed its proposition. That proposition consists of a predicate (specifying an activity or state or event) and one or more arguments (specifically the participants and their roles). Thematic roles classify those arguments. Stevenson *et al* (1994: 545) suggest that thematic roles may provide a bridging function "between the argument structure at the syntactic level, the thematic structure at the semantic level, and the event structure at the non-linguistic level".

Table 5.2 presents a subset, sufficient only for the current discussion, of the main thematic roles. The table is my own summary, though heavily reliant on Radford (2004) and Haegeman (2006).

**Table 5.2**: Semantic roles of arguments

| **Role** | **Description** | **Example** |
|---|---|---|
| AGENT/ACTOR | Intentionally initiates action | He kissed her |
| PATIENT | Undergoes effect of action | He kissed her |
| THEME | Undergoes change of state | He cured her |
| EXPERIENCER | Experiences some state | He is in love |
| GOAL | Destination of another entity | He reached Lincoln |

Fillmore's (1968: 24) Case Grammar specifies an agent (or "agentive") as "typically animate", because the role carries an obligation of intentionality. Klaiman (1991: 113) characterises agency in terms of intention, awareness and volition: all animate attributes.

The thematic roles of agent and experiencer, both requiring sentience, are most commonly associated with the syntactic roles of subject and indirect object (Dahl and Fraurud, 1996). Arnold (1998: 160) speculates about how this relationship between subjecthood and the thematic role of agent evolved: "The salience of subjects may derive from the fact that they are often used to indicate the agent role of a proposition. If agents are salient, salience could have become grammaticized into the grammatical subject". Branigan (1995: 27) speculates that the association of agent with subject might be due to a perceptual (or attentional) prominence conferred by their animacy: "agents tend to be perceptually prominent because they move".

Cruse (1973) contends that inanimate entities can "acquire a temporary 'agentivity' by virtue of their kinetic (or other) energy" (Cruse, 1973: 16), e.g.

[5.4] The computer updated the spreadsheet.

### 5.5 Focusing and Centering

Psycholinguistic accounts of anaphora resolution (e.g. Sanford, Garrod and colleagues) have been influenced by computational linguists such as Sidner (1983) on focus and by the subsequent development of Centering Theory (Grosz, Joshi and Weinstein, 1995; Walker, Joshi and Prince, 1998).

Focusing is most simply defined as "the movement of the focus of attention of the discourse participants as the discourse progresses" (Grosz, Pollack and Sidner, 1989: 446). The theory of focusing (Grosz, 1978; Sidner, 1979; Grosz and Sidner, 1986) accommodates two levels – global and local. Global focus is conceptualised as a "pushdown" stack of "focus spaces". Each focus space represents a segment of discourse, its purpose and the entities referenced in it. With each new segment, a new focus space is pushed on to the stack, to be "popped" when no longer salient (Grosz, Pollack and Sidner, 1989: 442). Global focus affects the candidates for co-referent definite descriptions. Local (or immediate) focus, which affects the candidates for co-referent pronouns, operates within discourse segments. The distinction between *global* and *local* focus is often related to the distinction between long-term and short-term (or working) memory (Carter, 2000: 268).

Sanford and Garrod (1981) explain anaphoric reference in terms of focus. As we comprehend a text, some items will be focused and in the foreground of our attention (and therefore easier to refer back to), and others will be in the background. Sanford and Garrod contend that four memory modules are required for successful textual comprehension. *Episodic memory* contains our representation of the text. *Semantic memory* holds the world-knowledge that we might need to call upon in order to understand the text. *Explicit focus* holds token representations of those items explicitly referenced by the text – people, objects, locations, etc. *Implicit focus* draws on semantic memory for what Sanford and Garrod call *scenarios* – representations of things that are implied by the text and that therefore facilitate a contextual interpretation.

Stevenson (1996) identifies three classes of focusing models. All three are defined (broadly) as mental models, and they all propose that entities or tokens represented within the model will have different weightings, or degrees of accessibility, or focus, which determine the most readily available co-referent of a pronoun. Stevenson (1996) classifies the three models as knowledge-based (e.g. Sanford and Garrod, 1981); semantic/pragmatic (e.g. Stevenson, Crawley and Kleinman, 1994); and structural (e.g. Grosz, Joshi and Weinstein, 1983; 1995).

What differentiates these three models is their supposed mechanism of focusing. Semantic/pragmatic focusing is considered by Stevenson (1996) to be the product of two principal factors: causal bias (of a verb) and thematic roles. Structural focusing derives from research in computational linguistics, and particularly from Centering Theory (see below). Having reviewed a wide range of candidates for "focusing effects", Stevenson (1996: 71) opts for a "dynamic" model of focusing, in which top-down effects (thematic role, first-mention) interact with bottom-up effects (heuristic strategies such as the role of connectives, parallel function).

Centering Theory possesses a number of attributes that might explain its appeal to researchers: conceptual simplicity; the relatively intuitive nature of its basic assumptions; and supportive evidence from both cognitive and computational studies. Centering Theory (CT) "attempts to predict which entities will be most salient at any given time" (Poesio *et al*, 2000). CT provides a theoretical framework in the form of propositions that proceed logically to link the coherence of a discourse to its "center" (the American spelling is the convention). The center is determined by its salience, relative to other discourse entities; and by the likelihood of it being instantiated as a pronoun in the subsequent discourse.

Poesio, Stevenson, Di Eugenio and Hitzeman (2004) present the propositions of CT as an ordered series of hypotheses:

- The coherence of a discourse segment is determined by repeated references to the same entities within successive utterances.
- Subsequent utterance is connected to a prior utterance by a specific discourse entity, a unique "backward-looking center" (CB).
- Each discourse participant maintains in "local focus" a dynamic set of "forward-looking centers" (CFs), i.e. entities that might feature as the CB in the next utterance. Stevenson (2002: 362) associates the CF with salience and the CB with coherence.
- The relative prominence of the available CFs is based on a ranking process that determines which CF is the "preferred center" (CP) and therefore which CF becomes the next CB.
- That ranking process is based on the relative salience of each CF.

- The center, or more specifically the backward-looking center (CB), is the single most salient entity, within the shared mental model of a discourse, at a particular point in time.
- That most salient entity (the CB) is more likely to be pronominalised than is any other entity.

These propositions have support, both theoretical and empirical, from independent research in linguistics, psycholinguistics, and computational linguistics (see Poesio *et al*, 2004, for references).

Strube and Hahn (1996: 271) have claimed that "the most important single construct of the centering model is the ordering of the list of forward-looking centers". The forward-looking centers in an utterance (the CF list) are ranked according to an "ordered set of features" or "CF template" (Cote, 1998: 56) which is specific to the language. The English-language CF template put forward by Brennan, Friedman and Pollard (1987) relies on a hierarchy of grammatical roles:

subject > direct object > indirect object > adjuncts

though they acknowledge that other factors might be significant. Cote (1998) presents counter-examples from everyday discourse to demonstrate that grammatical or syntactic features are inadequate to constitute a workable CF template for English, arguing that "lexical conceptual structures", based around thematic roles, provide essential scaffolding for the CF template.

It seems likely that a number of factors interact to determine the preferential ranking of CFs as potential co-referents for pronominal CBs, principally position (e.g. first-mention, or the center of the previous clause), grammatical role, and syntactic parallelism (i.e. entities with the same syntactic function as the pronoun). However, these surface factors are not absolute and may be superseded by semantic, pragmatic and/or contextual constraints.

## 5.6 Subjecthood, Salience and Animacy

The experiments documented in this chapter rely on an assumption that the first-mentioned or subject referent in a sentence will usually be the most salient referent. The link between salience and subjecthood was recognised in 1900 when Wundt (translation 1970: 29) postulated that "words follow each other according to the degree of emphasis on the concepts. The stronger emphasis is naturally on the concept that forms the main content of the statement. It is also first in the sentence".

In other words, "people put what they want to talk about … in the beginning of the sentence" (Clark, 1965: 369). In an experiment reported by Clark (1965), participants were asked to insert nouns in both active and passive sentence-frames that each consisted of "actor, verb, and object". Actor-nouns generated were generally animate (81.5% in active sentences and 68.3% in passive sentences), whereas object-nouns were generally inanimate (73.3% and 54.2% respectively). Clark also observed that, in both active and passive sentence-frames, participants tended to use an animate noun at the beginning of the sentence, and that this was a reflection of salience.

The correlation of animateness with subjecthood is explained by the canonical word-order of English which places the subject in first position, and by the high frequency of English verbs that license or demand an animate subject (McDonald, Bock and Kelly, 1993: 220). Arnold (1998: 112) finds that the salience of an entity is enhanced if it is in subject position, but does not distinguish this effect from the highly correlated (in English) "advantage of first mention". Crawley and Stevenson (1990: 193) cite prior studies that have identified structural features that correlate with salience, including first-mention (Sanford and Garrod, 1981) and using a proper name instead of a noun phrase (Sanford, Moar and Garrod, 1988).

The claimed advantage of first mention derives primarily from an influential paper by Gernsbacher and Hargreaves (1988). Their argument, that the salience of first-mention is a cognitive rather than a linguistic phenomenon, should be seen within the wider context of Gernsbacher's Structure-Building Framework (Gernsbacher, 1997), which posits that the developing foundation of a comprehender's mental model is based upon the information first received.

Itagaki and Prideaux (1985) provide evidence that frequency, alongside animacy and concreteness, is a significant independent variable in subject selection, and by extension in salience. From a review of psychology experiments, Itagaki and Prideaux (1985: 138) conclude that "concrete, animate and/or frequent nominals may be more salient and more available in the mental lexicon than abstract, inanimate, and/or infrequent ones". On this evidence they base their *Animacy-Concreteness Hypothesis*: The more animate and/or concrete a nominal referent is, the more likely it is to be realized as a subject (Itagaki and Prideaux, 1985: 139).

There is substantial cross-linguistic evidence (summarised in Branigan, Pickering and Tanaka, 2008: 173) of a correlation between animateness and syntactic prominence. In European languages, corpus studies have demonstrated that an animate NP is more likely to occupy the subject position in a sentence, and an inanimate NP is more likely to occupy the object position. In some non-European languages, such as Mam-Maya, there is a well-documented Animate First principle, according to which it is a grammatical requirement that an animate entity should always occupy sentence-initial position (see de Swart, Lamers and Lestrade, 2008, for an overview of cross-linguistic studies).

An ERP (event-related potential) study of object-relative clauses by Weckerly and Kutas (1999: 566) found "unequivocal" evidence that the animacy of a noun causes a rapid and specific activation upon being encountered, and evidence of "surprise" when a sentence-subject is not animate. They conclude (*ibid*: 569) that "the timing of the various animacy effects, in particular, the fact that they occur early in the sentence and at multiple locations even before any verbs, suggests that syntactic, semantic, and perhaps other types of information interact early and continuously to influence the incremental formation of a sentence-level representation".

A difficulty in discriminating animate from inanimate entities has been advanced as one of the defining characteristics of people with autism spectrum disorders (ASDs). This has been attributed to their relative resistance to social stimuli. Lake, Cardy and Humphreys (2010) tested word-order preferences in a picture description task for an ASD group and a control group. Overall, they found no significant difference between the two groups. Both the ASD and control groups placed animate referents in subject positions, but with a single exception: when the inanimate entity was a clock.

ASD individuals are often fascinated by intricate mechanisms. Lake *et al* (2010) contend that this demonstration of "personal salience" argues against the view that a preference for an animate subject is ruled simply by convention. If that were the case, then the ASD individuals would not have prioritised the clock. These findings support an account of animacy as a salience factor with a strong cognitive basis, even in individuals with perhaps an atypical perspective on the normal animacy hierarchy.

### 5.7 The Web as a Psychology Lab

Questionnaire-based studies are now routinely and preferentially administered via dedicated websites; psycholinguistic studies less so, reflecting a contemporary preference for physiological measures such as ERP and eye-tracking. The experiments reported in this chapter have been conducted exclusively online. They do not rely on timed responses (other than as a control check) and they utilise a dedicated website. The key questions addressed in this section are:

- Will a web-based experiment replicate the validity and reliability of a laboratory-based experiment?
- What constitutes best practice for the conduct of a web-based experiment?

A review of the relevant literature will address these questions.

Reips (2007) has highlighted three advantages of web-based experiments. First, there is the potential to obtain data from significant numbers of participants within short timescales and at relatively low cost. Second, a web-based participant pool replaces the usual, exclusively student, population (Foot and Sanford, 2004) with one that is more heterogeneous. Alternatively, the population for the experiment might be more specifically defined, e.g. by their interest in a particular web-based forum. Third, by taking the experiment to the participant rather than the participant to a laboratory, and by capturing experimental data in machinereadable form, internet-based experiments are cost-effective for the researcher and time-efficient for both participant and researcher.

Whilst some studies have successfully replicated laboratory-based results, there is not a complete consensus. Keller (2000: chapter 5) tested the reliability of web-based versus laboratory-based experiments by comparing the web-obtained data with data from more conventional experimental methods, conducted in a controlled environment and with the same materials. Keller found that his web-based experiments (grammar acceptability judgment tasks) achieved "near-perfect replication" (*ibid*: 230).

Corley and Scheepers (2002) used the WebExp software platform to replicate an earlier syntactic priming experiment by Pickering and Branigan (1998), using exactly the same materials. The method required participants to type sentence completions (128 per participant). The results of the web-based experiment were statistically comparable to those of the earlier, laboratory-controlled experiment.

However, Barenboym, Wurm and Cano (2010) have cast doubt on those findings. They compared two sets of data from online psycholinguistic rating experiments, where the only difference was the environment of the test: a remote location (e.g. the participant's home) vs. laboratory conditions. The results were sufficiently different to affect the statistical conclusion drawn from a regression analysis of each dataset. They also noted a gender difference: male participants' data correlated less well than females'.

Given this lack of consensus, the design of a web-based experiment must incorporate adequate checks of reliability and validity. Reips (2002: 248-249) provides a checklist of best practice, intended to maximise good data and minimise dropout (i.e. non-completion of the experiment). These include:

1. Stress the seriousness of the study, its scientific nature, and the need for good data.
2. Locate most text up-front, then progressively reduce it.
3. Stress that any reward will be for full compliance.
4. Ask questions up-front that identify ineligible participants (previous participation, expert status, language skills, etc), then accept or reject the participant.
5. Report both the dropout rate and the point at which most participants dropped out.

These recommendations have been implemented.

### 5.8 Sentence Continuation vs. Sentence Production

In a sentence continuation experiment, participants are presented with a stimulus sentence and invited to write a following sentence that continues the sense of the first. The rationale is that participants' choice of subject referent or pronominal referent will normally indicate the most salient entity from the preceding stimulus sentence. Sentence continuation has been deployed in many studies (e.g. Sanford and Garrod, 1981; Stevenson, Crawley and Kleinman, 1994; Arnold, 1998; Prat-Sala and Branigan, 2000; Pearson, Poesio and Stevenson, 2001; Koh and Clifton, 2002).

There are several issues with this methodology, however. First, it is difficult to devise stimulus sentences that both meet the specific requirements of the experiment and have ecological (i.e. real-world) validity - sentences that might actually be encountered in everyday discourse. For example (from Pearson, Poesio and Stevenson, 2001):

[5.5] The shop obtained Ann from the agency

[5.6] The club loaned Peter to Jane

The problem is that both the syntactic structure and the semantic content of each sentence have been dictated by the objective of the experiment.

The second issue, well-illustrated by the above examples, is that the stimulus referents (e.g. *shop*, *Ann* and *agency*) have to be forced into specific sentence structures in order to test (in that instance) different word orders and thematic roles. A hypothetical example in [5.7] illustrates how a sentence continuation experiment might need to be structured to test triplets of animate **[A]** and inanimate **[I]** referents. There are eight possible orders of the three conditions:

[5.7]

| | |
|---|---|
| Kim gave a friend a cat | **AAA** |
| Kim gave a friend a book | **AAI** |
| Kim gave the book to a friend | **AIA** |
| Kim gave the book to a library | **AII** |
| The book was given by Kim to a friend | **IAA** |
| The book gave Kim pleasure | **IAI** |
| The book gave pleasure to Kim | **IIA** |
| The book gave pleasure and information | **III** |

The example in [5.7] also illustrates the third issue, or rather constraint – the need to control for the semantics of the verb. That has been achieved by using the same verb (*gave/given*) throughout, but that would be difficult to sustain across multiple examples whilst preserving any semblance of ecological validity.

Pearson *et al* (2001) used opposing verbs of transfer such as *obtained* and *loaned*, the consequences of which are readily apparent in [5.5] and [5.6].

The fourth issue concerns verb bias created by the stimulus sentence. Fillmore (2003: 191-199) has argued that the choice of verb confers *perspective* on an event, e.g. the choice of *sell* or *spend* or *pay* or *cost* or *buy* in the description of a commercial event. Any one of these verbs will evoke the same schema (Fillmore prefers "scene"), but each verb brings a different entity/actor into perspective. Because it is inherently more natural to adopt the perspective of a human being (say, *buyer* or *seller*) than of an inanimate object (*cash*), the verb in the stimulus sentence will prime the continuation sentence accordingly.

The final issue concerns a potential fatigue effect on participants (Rasinger, 2008: 43-44), though this applies to both sentence continuation and production. The task of devising then writing or typing multiple sentences is extremely tedious and risks the production of poor-quality data, as well as a high dropout rate of participants. Itagaki and Prideaux (1985: 140) assigned 40 student participants to either a sentence or passage production task, with a single noun as the stimulus in both conditions. Of each group of 20, 17 completed the sentence task (average time 70 minutes) whilst only 14 completed the passage task (average time three hours). The design of the experiment that is documented here seeks to minimise any fatigue effect.

### 5.9 Online Sentence Production Experiments

The online experiments are designed to test whether the more animate of two referents is also the more salient of the two. For this investigation, the preferred alternative to sentence continuation is the experimental method of sentence production. Although this has been used most often with pictorial stimuli, there are precedents for presenting participants with lexical stimuli, e.g. Itagaki and Prideaux (1985), Nordquist (2004) and Gilquin (2010). The specific methodology of the sentence production experiments that are reported in this chapter differs from those precedents in two significant respects. First, the experiments have been conducted online, with participants recruited via social media. Second, those participants, presented with two nouns in a randomised order, have been asked to produce a sentence that contains both nouns, but not (in most cases) to write out the complete sentence. They were asked to think of a sentence and then simply to specify which noun occurs first in that sentence. The working assumption is that the first-occurring noun is most salient (see section 5.6).

In the sentence production task, both the context and the verb are of the participant's own devising. This eliminates the problem of devising ecologically valid stimulus sentences, and the participant is not presented with a context (and a verb) that might bias a particular word order. The recruitment and retention of large numbers of participants is facilitated, because the experiment can be completed in a relatively short time. Finally, this particular method of sentence production is more easily, and more objectively, assessed. In sentence continuation experiments, human judges are required to assess the output, since participants might use a pronoun as the subject of their sentence, and a judge must verify the referent of that pronoun. In this investigation's method of sentence production, there is a forced binary choice that is easily marked by computer.

There is a specific risk that participants answer at random. A counter-measure was therefore implemented. Participants (informed in advance) were periodically prompted to type in full the sentence that they had in mind. Such requests were apparently made at random, but in fact involved the same stimuli (presented within a randomised order) for each participant, so that their timed responses could be compared. Relatively slow responses would be assumed to indicate that they had not already thought of a sentence, and would be a criterion for excluding them from the results.

## 5.10 Materials for the Experiments

The initial set of materials was taken from the categorical analysis of the Denison *et al* (2008) database: 25 nouns from each category. Three values were assigned to each noun: CELEX wordform frequency (Baayen, Piepenbrock and van Rijn, 1993); length in number of letters; and number of syllables. The Match program (van Casteren and Davis, 2007) identified the best-fit pairings of nouns.

The selection process reduced, but did not eliminate, duplicated nouns, since a noun might be the best-fit with more than one other noun, e.g. *husband* with both *machine* and *kitchen*. A review eliminated pairs that would be difficult for participants to form into a sentence with any external validity (e.g. *kitchen* and *species*); and pairs in which one word would more likely be used in its verb sense (e.g. *fish* and *cast*). It is accepted practice to eliminate words on the basis that they are polysemous across different semantic categories (e.g. Anderson, Murphy and Poesio, 2014: 660). Deleted noun-pairs would still be available as examples and practice items.

**Names.** The analysis in chapter 3 would lead us to predict a high degree of salience for human names, but less so for collective names and place names. In the current experiments, the matching of one category of names to another sought to test whether different categories of names would be differentiated by their relative salience. Since the CELEX frequency data do not provide a comprehensive coverage of proper nouns, the matching of names to non-name items was based on frequency counts obtained from Wikipedia. This again necessitated some duplication of items. For example, *Adam* was the best match with both *Amnesty International* and *Cyprus*, whilst *Amnesty International* and *Cyprus* were the best match with each other.

Based on the matching process, 54 noun-pairs were selected: 18 for the pilot experiments and 36 for the main experiment, with six in common to both, in order to facilitate comparison of the results obtained:

| | **Pilot** | **Main** |
|---|---|---|
| Names v Names | 3 | 6 |
| Non-Names v Names | 6 | 12 |
| Non-Names v Non-Names | 9 | 18 |

Appendix 5.1 contains the lists of materials.

## 5.11 Pilot Experiments

**Objectives.** The first objective of the pilot experiments was to test and optimise the design and method of the main experiment. Recruitment of participants; information, instructions and examples provided; incentive of a charitable donation; structure of the experiment; operation of the online software and data capture; all were tested by the pilot experiments. The second objective was to test the hypothesis that participant data obtained from a sentence production experiment based on word choice would be reliably comparable to data obtained from an experiment based on participants typing every sentence in full.

**Method.** Two pilot experiments (A and B) were each completed by 40 participants, recruited through social media. No personal reward was offered, but a payment to a specified charity was made for every fully completed response. All participants certified that they were aged 18 or over and were native speakers of British English.

Participants were presented online with two sets of nine noun-pairs, and asked to think of a short sentence that included both nouns in any order. For the first set of nine they had to type their sentence. For the second nine they simply had to think of a sentence and then indicate which of the nouns came first in that sentence. To keep them 'honest' they were then asked to type out their sentence in two of those nine cases (having been told in advance that would be required).

The software randomised both the order of presentation of the noun-pairs and the order of the nouns within each noun-pair. The set of noun-pairs that came first in Pilot A came second in Pilot B. The combination of the two experiments thus provided 40 results for each noun-pair in each condition (typing a sentence vs. nominating a word).

**Results.** The data obtained from typed sentence production and the data obtained from word choice, for the same noun-pairs, were compared by Pearson's correlation. The resulting correlation coefficient was 0.80, *p* < .001. In order to locate this result in a meaningful context, it is necessary to ask: what would be the corresponding correlation within the 'gold standard' of all-typed sentence production? This was answered by comparing the typed responses to the same 18 noun-pairs in Pilot A and Pilot B. The correlation coefficient was 0.89, $p < .001$. The conclusion is that the two methods give results that are sufficiently reliable and comparable for the main experiment to proceed on the basis of word choice (though with a number of additional verification measures).

**Participant feedback.** The average completion time for the pilot experiments was 11 minutes. Only one participant commented that the instructions were not clear, and most completed the experiment correctly. The *collective human* category caused a few problems, with younger participants unfamiliar with organisations such as *Amnesty International*, and one commented that she "wasn't sure if *Labour Council* was a place or an institution or a person". Otherwise, aside from some ribald responses to the noun-pair of *husband* and *machine*, there were only isolated instances such as *park* used as a verb instead of a noun, *Bob* used as a hairstyle in spite of its capitalisation, and an objection that *god* had (deliberately) not been capitalised. These observations informed the selection of materials for the main experiment.

## 5.12 Design of Main Experiment

A full script of the experiment is at Appendix 5.2.

**Block design.** The main experiment was conducted as three parallel experiments. By organising the 36 noun-pairs into six blocks of six, it was possible to test all 36 equally, whilst limiting the task for any one participant to 24 noun-pairs, to minimise any fatigue effect or practice effect (Rasinger, 2008: 43-44). Each block in the six-block structure contained the same mix of six items from three sectors:

| **Sector** | **Items** |
|---|---|
| Names vs. Names | 1 |
| Non-Names vs. Names | 2 |
| Non-Names vs. Non-Names | 3 |

The blocks were allocated to the three parallel experiments in a simple Latin square design, such that each of the six blocks was presented twice across the three experiments. The presentation order of the noun-pairs was randomised for each participant by the software, as was the order of the nouns in each pair.

**Software.** The software platform selected for the online experiment was provided by Qualtrics (under licence to the University of Essex), following a comparison of the facilities provided by several similar providers. Qualtrics was equal or superior to its rivals on all comparisons, but in particular on its facilities for randomisation and on its ability to time participants' responses.

Citation: Version 37,892 of the Qualtrics Research Suite. Copyright © 2013 Qualtrics. Qualtrics and all other Qualtrics product or service names are registered trademarks of Qualtrics, Provo, UT, USA: http://www.qualtrics.com.

**Ethical approval.** The sequence of experiments (both pilot and main) received formal ethical approval from Professor Berthold Lausen on behalf of the

University of Essex Faculty Ethics Committee (Science and Engineering). Approval was granted under Annex B to the Procedure for Making an Application for Ethical Approval, which covers "well-established, ethically non-controversial and commonly-used types of test or experimental procedure".

**Participants**. The sample of participants recruited for both the pilot and main experiments might be regarded as an "opportunity sample" (Sapsford and Jupp, 1996: 38) or "convenience sample" (Bryman, 2008: 183-184). The recruitment method might best be described as "snowball" sampling (Bryman, 2008: 184-185; Rasinger, 2008: 51). A small number of recruiters promoted participation via social networks – Facebook, Twitter and e-mail. They stressed the serious nature of the experiment and asked recipients to forward the message to others, hence the "snowball" description. To be eligible to take part in the experiment, participants had to confirm that they were aged 18 or older and were native speakers of British English. They were informed that £1 would be donated to charity, up to a maximum of £200, for every completed response (a donation of £200 was duly made).

At the end of the experiment, participants optionally stated their age (within bands) and gender, and 95.8% of the 216 participants did so. This was the demographic of participation (percentages):

| | **Female** | **Male** | **Total** | **Withheld** |
|---|---|---|---|---|
| **18-30** | 23.6 | 15.7 | 39.3 | |
| **31-50** | 23.6 | 9.7 | 33.3 | |
| **Over 50** | 14.4 | 8.8 | 23.2 | |
| **Total** | 61.6 | 34.2 | 95.8 | 4.2 |

Sapsford and Jupp (1996: 38) define an opportunity sample as "whoever happens to be available from the population of interest", precluding any planned balance of genders and ages. Although the gender distribution is predominantly female, the age distribution is arguably more representative of the general population than the samples drawn exclusively, as in so many psychology experiments, from a population of students (Foot and Sanford, 2004).

### 5.13 Cheaters, Speeders and Dropouts

The experimental design incorporated counter-measures to detect those participants who are described in the Qualtrics user manual as "cheaters and speeders", i.e. those who respond at random and without any thought, aiming simply to complete the experiment in the shortest possible time. Three checks were implemented.

The first check was based on the elapsed time for completion of the experiment. This eliminated any participants with an elapsed time that was less than half of the mean time for all participants.

The second check examined the four sentences that participants were required to type out, having already nominated the first-occurring stimulus word. The participant was eliminated if more than one sentence was nonsense (one responded "bananas" to every request), or did not contain both stimulus words, or did not begin with the nominated word.

The same single item in each block (i.e. four items in each experiment) prompted participants to "type here the sentence you thought of". The third check took account of the elapsed time between onset of the request for a typed response and the 'first click' by the participant, to begin typing in the text box. This was summed for all four sentence requests. Any participant with more than twice the mean time for all participants was eliminated, on the assumption that a consistently long thinking time indicated that they had probably not actually formed a sentence prior to nominating the first word in that sentence.

The combined effect of these counter-measures was to eliminate 28 participants:

| | **Exp A** | **Exp B** | **Exp C** |
|---|---|---|---|
| **Elapsed time** | 2 | 2 | 2 |
| **Bad sentences** | 3 | 2 | 1 |
| **First click** | 4 | 6 | 6 |

In each of the three experiments there was one participant who was caught by both the 'bad sentence' and the 'first click' counter-measures.

The few comments by participants were mostly complimentary about the ease of completing the experiment. Only two elected to contact the researcher by email for more information. No comments indicated the adoption of a particular strategy to complete the experiment.

The average dropout rate across the three experiments, i.e. those who joined but did not complete an experiment, was 24%. Nearly all of those dropped out after progressing through the instructions to the first set of examples. Anecdotal evidence would suggest that at least some of these dropouts were simply taking a first look, and later returned to complete the experiment.

### 5.14 Results and Analysis

Management of the quota facility in the Qualtrics software achieved an equal number of completed responses from each of the three parallel experiments, to meet the target level of 72 per experiment, or 5,184 data-points (144 per noun-pair). The results for each block of items are shown in Table 5.3. Each heading indicates which of the three experiments incorporated that block, and its order: Block 1 was presented first in Experiment A (hence A1) and third in Experiment B (B3). The right-most columns are the first choice percentage for each item. H, C and P represent the three categories of names – human, collective and place.

**Table 5.3**: Results of sentence production experiments – Blocks 1-6

**Block 1: Experiments A1 and B3**

| | | | | | |
|---|---|---|---|---|---|
| H | Jack | 1 | husband | 52.1 | 47.9 |
| 1 | teacher | 3 | building | 84.7 | 15.3 |
| 1 | boy | 3 | car | 91.7 | 8.3 |
| 1 | husband | 4 | kitchen | 93.1 | 6.9 |
| H | Billy | C | Aston Villa | 95.8 | 4.2 |
| H | Michael | 1 | wife | 76.4 | 23.6 |

**Block 2: Experiments A2 and B4**

| | | | | | |
|---|---|---|---|---|---|
| 1 | wife | 5 | road | 84.7 | 15.3 |
| P | York | 1 | baby | 25.0 | 75.0 |
| 1 | girl | 4 | book | 86.8 | 13.2 |
| 1 | doctor | 5 | window | 87.5 | 12.5 |
| H | Bob | C | BBC | 95.1 | 4.9 |
| P | Egypt | 1 | girl | 11.8 | 88.2 |

**Block 3: Experiments A3 and C1**

| | | | | | |
|---|---|---|---|---|---|
| 2 | lord | 4 | tree | 79.9 | 20.1 |
| C | Co-op | 2 | dog | 13.2 | 86.8 |
| C | CIA | 2 | politician | 42.4 | 57.6 |
| 2 | politician | 4 | territory | 84.7 | 15.3 |
| 2 | dog | 5 | seat | 88.2 | 11.8 |
| H | Adam | P | Cyprus | 95.1 | 4.9 |

**Block 4: Experiments A4 and C2**

| | | | | | |
|---|---|---|---|---|---|
| C | Labour Party | 3 | weapon | 81.2 | 18.8 |
| H | Billy | P | London | 95.1 | 4.9 |
| 2 | person | 6 | action | 75.0 | 25.0 |
| 2 | horse | 5 | fire | 79.9 | 20.1 |
| C | BBC | 3 | town | 84.7 | 15.3 |
| 2 | son | 6 | view | 66.0 | 34.0 |

**Block 5: Experiments B1 and C3**

| | | | | | |
|---|---|---|---|---|---|
| H | Arthur | 4 | book | 63.2 | 36.8 |
| 3 | hotel | 6 | value | 73.6 | 26.4 |
| 3 | building | 5 | village | 72.9 | 27.1 |
| 3 | town | 5 | wall | 34.7 | 65.3 |
| H | Adam | 4 | ball | 89.6 | 10.4 |
| C | Aston Villa | P | London | 83.3 | 16.7 |

**Block 6: Experiments B2 and C4**

| | | | | | |
|---|---|---|---|---|---|
| 4 | room | 6 | word | 36.8 | 63.2 |
| P | Cyprus | 4 | palace | 24.3 | 75.7 |
| 3 | paper | 6 | table | 87.5 | 12.5 |
| 4 | glass | 6 | floor | 78.5 | 21.5 |
| C | BBC | P | India | 84.0 | 16.0 |
| P | Africa | 4 | tree | 20.1 | 79.9 |

**Names.** We have already reviewed evidence in chapters 3 and 4 that the genitive ratio is not reliable with proper nouns/names. The experiment nevertheless tested the relative salience of different categories of names (human, collective and place) when competing with common nouns (both animate and inanimate) and with each other. The results have been summarised in a precedence matrix (Figure 5.1). Each number in the matrix represents the percentage of participant responses that opted for the category in the left-hand column when it was competing against the category in the top row. So, given the choice of a human name versus an animate noun, 64.3% of responses selected the human name as the first noun in the sentence they produced.

| | Human Name | Animate Noun | Collective Name | Inanimate Noun | Place Name |
|---|---|---|---|---|---|
| Human Name | ▨ | 64.3% | 95.5% | 76.4% | 95.1% |
| Animate Noun | ▨ | ▨ | 72.2% | 88.2% | 81.6% |
| Collective Name | ▨ | ▨ | ▨ | 83.0% | 83.7% |
| Inanimate Noun | ▨ | ▨ | ▨ | ▨ | 77.8% |

**Figure 5.1**: Precedence matrix of names and nouns

The matrix shows that, as discussed in section 3.6, human names take precedence over all other categories, followed by animate nouns and then collective names. Place names are not salient, even when competing with inanimate nouns, and have therefore been excluded from the matrix. This might be an unintended consequence of the experimental method. Presented with *York* and *girl*, it seems likely that most respondents will produce a sentence that places

*York* in a prepositional phrase, usually preceded by *in, from, of* or *to*. The competing noun, even if inanimate or abstract, is therefore more likely to take sentence-initial position:

[5.8] The car was made in York

The view of York is nice

It has also been observed that only the names of collectives are compound nouns, or the initials of compound nouns (e.g. *BBC*, *CIA*). A consequent bias cannot be discounted.

### 5.15 The Animateness Rating of Salience

**Table 5.4**: Animateness rating (AR) as predictor of salience

| | | **AR** | | | **AR** | **Predicted?** |
|---|---|---|---|---|---|---|
| **Animate-Inanimate** | | | | | | |
| teacher | 84.7% | +2.017 | building | 15.3% | +2.093 | **N** |
| boy | 91.7% | +0.740 | car | 8.3% | +0.432 | Y |
| husband | 93.1% | +1.871 | kitchen | 6.9% | +0.273 | Y |
| girl | 86.8% | +1.020 | book | 13.2% | +0.826 | Y |
| wife | 84.7% | +0.863 | road | 15.3% | -0.605 | Y |
| doctor | 87.5% | +1.758 | window | 12.5% | -4.286 | Y |
| lord | 79.9% | +1.129 | tree | 20.1% | -0.235 | Y |
| politician | 84.7% | +1.978 | territory | 15.3% | +2.053 | **N** |
| dog | 88.2% | +0.588 | seat | 11.8% | -1.730 | Y |
| horse | 79.9% | +1.054 | fire | 20.1% | -2.642 | Y |
| person | 75.0% | +1.768 | action | 25.0% | -9.517 | Y |
| son | 66.0% | +0.117 | view | 34.0% | -2.703 | Y |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Both Inanimate** | | | | | | |
| building | 72.9% | +2.093 | village | 27.1% | +1.624 | Y |
| wall | 65.3% | -1.729 | town | 34.7% | +1.071 | **N** |
| hotel | 73.6% | +1.238 | value | 26.4% | -15.078 | Y |
| paper | 87.5% | +1.371 | table | 12.5% | -1.588 | Y |
| word | 63.2% | -1.638 | room | 36.8% | +0.037 | **N** |
| glass | 78.5% | -29.898 | floor | 21.5% | -1.805 | **N** |

The animateness rating (AR) scores were computed after the experiment was completed. The AR algorithm has been optimised to predict whether a target noun is animate. It is less effective at differentiating inanimate concrete from inanimate abstract nouns. A high positive score generally indicates that a noun is likely to be animate. A lower or negative score generally indicates a concrete or abstract noun. The animateness rating is judged to be animate (predicted: Y) if it places the nouns in the same order as the participants' ratings.

**Common nouns**. Table 5.4 shows the experimental results from the non-name pairs of nouns, together with their respective AR scores. Although the sample sizes are small, some tentative conclusions might be drawn. There are 12 pairings of animate vs. inanimate nouns. In all 12 cases, participants' sentence productions strongly (by a mean of 83.5%) specified the animate noun as sentence-initial (and therefore assumed to be salient). The animateness rating (AR) correctly predicted the more animate/salient noun in 10 out of 12 cases: 83.3%. The two exceptions, *building* and *territory,* both scored unusually high (for concrete nouns) in the DS condition, i.e. *the building's*, *the territory's*).

Only six inanimate pairings were tested. With the benefit of hindsight, it is probable that the selection of the inanimate nouns was ill-advised. Five of the six noun-pairs contain what might broadly be defined as location words, words that often feature in preposition phrases preceded by *in* (*village*, *town*, *room*) or *on* (*table*, *floor*), and are therefore unlikely to be sentence-initial in the iconic sentence structure of English.

The minimal conclusion that can be drawn from this experiment is that the AR is a reliable (above chance) predictor of relative salience when tested on animate/inanimate noun-pairs. That conclusion will be tested more rigorously by the corpus-based analysis in section 5.17.

## 5.16 Produzione di frasi in italiano

Google Translate will translate between (say) Hindi and Catalan by using English as a 'bridge' (Mayer-Schönberger and Cukier, 2013: 38). An acknowledged limitation of the genitive ratio model is that it is specific to English, in that it relies upon the "Saxon or Germanic genitive" (see section 3.2). The experiment reported in this section tests a hypothesis that English-language genitive constructions will provide a proxy for their equivalent referents in other languages. The specific prediction is that there should be no significant difference in the salience judgments of English and Italian native speakers. If both select an animate noun as salient, then the relative animacy of the Italian noun can be measured by the animateness rating of its English equivalent.

**Experiment**. This experiment replicated the online sentence production experiment reported above, though on a smaller scale (of both items and participants) and in a different language (Italian).

**Materials**. Twelve of the 18 pairs of common nouns tested in the previous experiment (see Table 5.4) were tested. Names were not tested.

**Participants**. Participants were again recruited mainly via social media, with an additional appeal to the Dante Alighieri Society (established in the UK to promote Italian language and culture). The experiment was completed by 62 participants, all aged 18 or over and native speakers of Italian. This was the demographic of participation (percentages):

| | **Female** | **Male** | **Total** | **Withheld** |
|---|---|---|---|---|
| **18-30** | 45.2 | 19.4 | 64.6 | |
| **31-50** | 12.9 | 12.9 | 25.8 | |
| **Over 50** | 1.6 | 0 | 23.2 | |
| **Total** | 59.7 | 32.3 | 92.0 | 8.0 |

Eighteen respondents dropped out, mostly at the introduction/instructions stage. In some cases they will not have matched the criteria for participation. Six dropped out after completing the first practice sentence.

Verification checks to detect cheaters and speeders were again applied. Only one participant was eliminated after completing the experiment, on the grounds that none of her sentences contained the given words. One participant requested feedback.

**Results**. Table 5.5 shows the results of the experiment. The twelve noun-pairs are listed in English and Italian in the four left-hand columns, with the first-listed noun having the highest AR score (as in Table 5.5). The columns headed 'English' and 'Italian' list the mean percentage of participants in each language who selected that first-listed noun. Thus, *lord* was placed in sentence-initial position by 79.9% of English speakers and 71.0% of Italian speakers. The prediction that there should not be a significant difference between the English and Italian scores was tested by a paired two sample for means *t*-test: *t* (11) = 1.81, *p* (two-tailed) = .097. The test therefore indicates that, as predicted, there is not a significant difference between the two sets of scores.

**Table 5.5**: Comparison of English and Italian preferences for noun-pairs

| | | | | **English** | **Italian** |
|---|---|---|---|---|---|
| lord | tree | signore | albero | 79.9% | 71.0% |
| politician | territory | uomo politico | territorio | 84.7% | 80.6% |
| dog | seat | cane | posto | 88.2% | 41.9% |
| horse | fire | cavallo | fuoco | 79.9% | 74.2% |
| person | action | persone | azione | 75.0% | 75.8% |
| son | view | figlio | veduta | 66.0% | 58.1% |
| building | village | edificio | paese | 72.9% | 59.7% |
| wall | town | muro | città | 65.3% | 74.2% |
| hotel | value | albergo | valore | 73.6% | 53.2% |
| paper | table | carta | tavolo | 87.5% | 66.1% |
| word | room | parola | stanza | 63.2% | 50.0% |
| glass | floor | bicchieri | pavimento | 78.5% | 96.8% |

**Inference**. Asked which of two nouns they would place first in a sentence of their own devising, English and Italian native speakers make similar choices. Those choices seem to be primarily cognitive rather than linguistic – *horse* and *cavallo* are two referents for the same entity, in a common mental model which is independent of language. This small sample in just one other language suggests the possibility at least that genitive ratio analysis might be extended to languages other than English.

### 5.17 A Corpus Test of Salience

"Conclusions that are common to studies using several techniques are almost always more secure than those based on a single technique" (Garnham, 2001: 62).

Tasks such as sentence continuation or production require participants to simulate their linguistic responses to materials not of their own choosing, and within the artificial context of an experiment. By contrast, an analysis based on naturally-occurring usage within a corpus is arguably a more reliable and more objective (though less controlled) measure of linguistic behaviour. Nordquist (2004: 211) cites several independent studies that have compared the data elicited from participants in sentence completion and sentence production experiments unfavourably with the patterns of language found in corpora, and particularly in conversational corpora.

The sentences analysed here would present no great challenge to any of the anaphora resolution systems discussed in section 5.2. The co-referent pronouns are gender-based, there is only one animate candidate, and that is generally in subject/sentence-initial position. The purpose of this experiment is simply to assess the animateness rating's reliability in identifying animate nouns, based on semi-conversational samples of American and British English. The diverse combinations of referents simulate the challenges of natural language to any real-world application of the genitive ratio model.

**Materials**. From the Birmingham Blog Corpus (Kehoe and Gee, 2012) and its concordance, thirty sentences were extracted, all preceding a sentence beginning with the personal pronoun "He …" or "She …". Each sentence contains one referent which is both animate and salient and which is the co-referent of the pronoun, together with 1-5 other referents. Some original wording has been edited, either to make the examples more succinct, or in some cases to combine wording from two preceding sentences. Each noun is underlined and followed by its AR score, with the salient co-referent noun in **bold**.

[5.9] The **president** [+2.377] needs to turn this page [-1.019] in a hurry [-6.941]. He

[5.10] The **president** [+2.377] worships power [-11.851], money [-4.626] and war [-0.321]. He

[5.11] The **singer**'s [+1.803] appeal [-9.447] is a strong voice [-1.598]. He

[5.12] This **anglophile** [-2.607] has an obsession [-20.731] that borders on psychosis [-3.234]. He

[5.13] The **founder** [+1.971] was charged with cruelty [-5.593] and theft [-19.287]. He

[5.14] There's a **commenter** [+1.256] who can recite a passage [-10.890] from the report [+0.064] in his sleep [-9.799]. He

[5.15] My **neighbour** [+2.335] even relaxed his torture [-6.618] regime [+1.687]. He

[5.16] The **farmer** [+1.729] does not make much profit [-31.975] per field [-0.617]. He

[5.17] A **composer** [+2.424] has to start at some point [-6.431] and then build on success [-142.867]. He

[5.18] How does a **quarterback** [+2.182] fail a drug [-0.626] test [-2.850] and continue to play football [-20.928]? He

[5.19] The **governor**'s [+2.404] amateurism [-1.320] and liabilities [-14.372] are badges [-3.969] of honour [-1.834]. She

[5.20] The movie [+1.224] **character**'s [+2.315] general challenge [-12.418] in life [-4.359]: finding happiness [-6.698] somewhere besides work [+0.024]. She

[5.21] This **director** [+2.382] of an investment [-9.557] firm [+1.183] lives in a waterside home [-0.545]. She

[5.22] Does the **writer** [+1.773] of this review [-2.089] even like the show [+1.181]? He

[5.23] Look at the **spouse** [+0.824] or romantic interests [-26.346]. He

[5.24] The **prime minister** [+2.378] commented on the cabinet [-1.344] reshuffle [-5.901]. He

[5.25] The **speaker** [+2.066] is leading the race [+0.245] to be the next CM. He

[5.26] The **man** [+0.436] has made millions [-34.606] upon millions [-34.606]. He

[5.27] The **man** [+0.436] had returned to the window [-4.227] for another slice [-16.165]. He

[5.28 My **husband** [+1.871] heard the commotion [-2.178]. He

[5.29] I see her as a **victim** [+1.778] who has the consolation [-4.251] of success [-142.867]. She

[5.30 What should a **woman** [+1.301] with a history [-78.639] of breast [-5.861] cancer [-6.297] do? She

[5.31] The **woman** [+1.301] has a serious case [-2.415] of hoof [-5.474] and mouth [-13.098] disease [-1.052]. She

[5.32] I found my **mother** [+1.668] up, in her robe [-3.984], with the radio [-2.293] on. She

[5.33] Our **babysitter** [+2.374] was a high school [+1.778] junior [-3.358]. She

[5.34] The **actress** [+1.136] is ready for a new relationship [-11.856] status [-47.427]. She

[*5.35*] The **mayor** [+1.390] runs a small company [+2.110] that specializes in coffee [-5.316], tea [-6.464] and spices [-12.045]. She

[*5.36*] The FAA **spokeswoman** [-0.090] could not confirm reports [+0.062] that the plane [+1.139] struck turbulence [-5.313]. She

[*5.37*] The **police officer** [+1.604] sued the department [+1.974], claiming that it engaged in sexual discrimination [-7.518] and other illegal activities [-2.527]. She

[*5.38*] A **terrorist** [-0.886] with a bomb [+0.237] in his underwear [-3.576] got on a flight [+0.174]. He

**Results**. The animacy ratio successfully identified the salient animate noun in 26 of these 30 examples (86.7%). The four failures are listed at the end in [5.35] – [5.38].

**Error analysis**. Two of the nouns that scored higher than the human animate co-referent noun were collective animate nouns: *company* [5.35] and *department* [5.37]. The salient noun *spokeswoman* [5.36] is infrequent in Wikipedia, scoring only one actual hit in any possessive construction. The frequent use of *terrorist* [5.38] as an adjective (e.g. *terrorist threat*, *terrorist organisation*) in Wikipedia affects the genitive ratio in favour of poss-*of*. The same point could be made about *anglophile* [5.12], which is more common as an adjective and which has the higher AR only because it is competing with two abstract nouns (*obsession* and *psychosis*).

**Conclusion**. The four 'failures' illustrate the limitations of the genitive ratio that were discussed in chapter 4. The 26 successes illustrate its potential. This is a tougher test than (for example) the MARS tests (see section 5.2) that were based on the relatively well-defined and less-ambiguous language of computer and technical manuals.

The aim of this experiment has been to demonstrate the potential of the AR, not to propose a new model of co-reference resolution that would compete with BART (section 5.2) or even MARS. Those are multi-factor models with pre-processing, whereas this has been a single factor test: an appropriately-weighted gender recognition pre-process would have resolved all four errors.

## 5.18 Caveats and Conclusions

Ariel (1990: 108) acknowledges that "subjects, humans, topics, etc are more salient than non-subjects, non-topics, etc." It is not the claim of this thesis that animateness is always sufficient, in and of itself, to determine the discourse salience of a referent. Rather, it is claimed that animateness, in interaction with other factors, will augment that determination, even though Koh and Clifton (2002: 841) rather unhelpfully "suspect that the factors that determine salience are unlimited in number".

Neither is it the aim of this thesis to make exaggerated claims for the genitive ratio. On the contrary, its limitations as well as its potential have been and will be acknowledged. In common with the succeeding chapters on possible applications, this chapter has offered a proof of concept, a *prima facie* case for a method that has evident value in its own right, but might best be applied in combination with other factors and processes.

Taken together with the findings of the preceding chapters, these experiments complete a 'chain of evidence' that links the genitive ratio to relative animacy, and relative animacy to relative salience. The empirical findings support the case for animateness as a determinant of salience and for the animateness rating as a reliable (above chance) predictor of co-reference with an animate pronoun.

Within the framework of the experiment, new thinking has been applied to the process of obtaining experimental data. Whilst there are precedents for components of the process, the whole represents an original paradigm: recruitment of participants via social media; online delivery of the experiments to those participants; a method of sentence production that minimises participant fatigue and facilitates automated assessment; verification checks that test the commitment of participants.

The Italian-language experiment suggests that the genitive ratio might be applied to languages other than English, by using English as a bridge that connects the animacy of an entity to its linguistic referent. This would significantly extend the genitive ratio's range of applications.

Accuracy of 86.7% in identifying the most animate referent (in the Birmingham Blog Corpus experiment) is a creditable result, particularly for a single-factor model. As we have seen (in section 5.2), co-reference algorithms are typically multi-factor and pre-processed. A viable role for the genitive ratio might be to supply an additional module to an established system such as BART (Versley *et al*, 2008).

CHAPTER 6

# The Concrete Language of Alzheimer's

Lear: I fear I am not in my perfect mind.

William Shakespeare: *King Lear,* Act 4, Scene 7

## 6.0 Overview

By means of the animateness rating, the genitive ratio method has been applied (in chapter 5) to the differentiation of individual nouns within a text or discourse – animate from inanimate (concrete and abstract). This chapter shifts the focus of analysis from the individual noun to the text as a whole, and utilises the concreteness rating (CR) that was developed in chapter 4. By obtaining CR scores for all of the common nouns (i.e. all of the nouns except proper nouns) within a text, and then calculating the mean of those scores, it should be possible to derive a concreteness rating for that text, whether it be a letter, a novel or poem, a diary or blog, or spontaneous speech.

Aggregated within a concreteness rating, the genitive ratio becomes a possible tool for text comparison. The hypothesis is that the CR will provide a measure of a writer's or speaker's cognitive bias towards relative concreteness or abstractness, thus enabling (for example) two speakers' responses to the same interview question to be compared; or two letters or novels written by the same author at two different stages of their life. Such a comparative measure has potential relevance because a change or difference in the semantic bias of linguistic usage, from abstract to concrete or from concrete to abstract, either within or between subjects, could be a significant factor in the assessment, and potentially the treatment, of mental illness.

This chapter will examine the application of the concreteness rating to an illness that dominates the current public health debate – dementia, and specifically Alzheimer's disease – an illness that is characterised by thinking and language that are atypically concrete. The next chapter will examine the other mental illness with which 11% of adults in England were diagnosed in 2009 (source: Office of

National Statistics, Social Trends 41, 2011) - depression, an illness that is characterised by thinking and language that are atypically abstract.

There is no doubt that the onset of dementia is marked by changes in language production. Although the nature of those changes is affected by both individual differences and by the specific subtype of dementia, there is a consensus from prior studies that, in patients with Alzheimer's disease (AD) in particular, the vocabulary of the patient becomes progressively less abstract and more concrete.

Several studies have attempted to construct a model of language change that will provide both early diagnosis and monitoring of the progression of AD. Nearly 200 separate linguistic features have been tested in those models. The most interesting study tested just two: the Nun Study (Snowdon *et al*, 1996) found that measures of idea density and grammatical complexity, manually computed from texts written at an average age of 22, were above-chance predictors of the development of AD 60 years later.

A key problem for research in this area is lack of data. Projects such as the Nun Study are expensive and slow to deliver up their findings. This has prompted a more accessible line of research: retrospective analysis of texts written by authors with a diagnosis of AD (though not always clinically confirmed).

This chapter re-visits three published studies of linguistic change in cases of dementia: two novelists and one politician. Both one of the novelists (Iris Murdoch) and the politician (Harold Wilson) had clinical diagnoses of AD. The other novelist (Agatha Christie) had no formal diagnosis but the anecdotal evidence strongly suggests AD or possibly vascular dementia. Two novelists, both recently deceased writers of crime fiction (P.D. James and Ruth Rendell) provide 'healthy controls'. A diachronic analysis of Harold Wilson's language illustrates a limitation of GR analysis: the constraints of a particular genre or linguistic register (in this case, parliamentary language) as a data source.

Autopsy is still considered by many clinicians to be the only certain diagnosis of AD. Advances in brain imaging have made available new diagnostic tools, but these are limited in their application by the high costs of provision. It is possible that a risk assessment model might combine genetic biomarkers with a GR analysis and other non-invasive diagnostic factors, to detect AD at an early stage, as well as to monitor the degenerative course of AD.

## 6.1 Benefits of Early Diagnosis

"Alzheimer's pathology likely begins many years and perhaps decades before the onset of symptoms; therefore, there is an opportunity for prevention once future advances make it possible to diagnose the disease through the use of biomarkers before symptom onset" (Lyketsos, 2009: 249).

The 2014 report *Dementia UK: Second Edition*, by King's College London and the London School of Economics, estimates the total cost of dementia in the UK at £26.3 billion (£32,250 per person with dementia). Since the implications, and particularly the costs, of a burgeoning 'dementia problem' in an ageing population became apparent to governments around the world, there has been a long-overdue increase in international cooperation and in funding for research into dementia in general, and Alzheimer's disease (AD) in particular.

This chapter will focus mainly on AD, as it is by far the most common of the subtypes of dementia. Research into AD is broadly aligned with four objectives:

1. In the long term, to find a 'cure' that would reverse the effects of the disease, and
2. To develop treatments that would prevent the onset of the disease.
3. In the shorter term, to establish courses of therapy that will slow down the progression of the disease.
4. More immediately, to identify at an early stage those who are most at risk of developing the disease.

The research findings presented in this chapter are aligned with that fourth objective. To the well-established cognitive tests and questionnaires used by clinicians have been added serious efforts to apply computational linguistic analysis to the diagnosis of AD and other subtypes of dementia. A review of prior research findings will give examples of these applications, most of which are based on machine learning.

Early diagnosis of incipient AD carries potentially significant benefits for the cost and delivery of public health. It facilitates both the most efficient allocation of resources and the deployment of interventions to inhibit the progression of AD. Interventions aimed at slowing down the progression of the disease might include some form of cognitive training as well as medication. At-risk patients could also be counselled to address the lifestyle risk factors that have been associated with AD (e.g., lack of exercise, smoking, alcohol consumption, diabetes, poor diet, and obesity), and to make better-informed decisions for themselves and their families. Early diagnosis might prompt them (for example) to move to a more suitable type of accommodation, or perhaps simply to bring forward the pursuit of their 'bucket list' objectives, before the more severe effects of AD begin to become apparent.

These potential benefits are, however, accompanied by the ethical considerations associated with a probability-based prognosis, particularly the psychological harm that might be caused by a 'false positive'.

## 6.2 Neurocognitive Disorders

DSM-5 (*Diagnostic and Statistical Manual of Mental Disorders*, 5th edition, American Psychiatric Association, 2013) classifies dementia as a "neurocognitive disorder" (NCD). The NCD classification includes only disorders in which there is a severe cognitive deficit that is both core and acquired (i.e. not developmental). DSM-5 identifies the 13 "etiological subtypes" of NCD set out in Table 6.1. These subtypes are not discrete. The *Dementia UK* report (Alzheimer's Society, 2007) cites studies indicating that "mixed pathologies are much more common than 'pure'" (*ibid*: 20).

**Table 6.1**: Subtypes of neurocognitive disorder (source: DSM-5, 2013: 591)

| NCD | Subtype |
|---|---|
| | due to Alzheimer's disease |
| | vascular |
| | with Lewy bodies |
| | due to Parkinson's disease |
| | frontotemporal |
| | due to traumatic brain injury |
| | due to HIV infection |
| | substance/medication induced |
| | due to Huntingdon's disease |
| | due to prion disease |
| | due to another medical condition |
| | due to multiple etiologies |
| | unspecified |

The Alzheimer's Society UK estimates that there will be 850,000 people with dementia in the UK by 2015, projected to exceed a million by 2025. They estimate that at least half of those cases are undiagnosed. The *Dementia UK* report (2007) estimates that Alzheimer's disease (AD) accounts for 62% of dementia cases, with vascular dementia (VaD) and 'mixed' (Alzheimer's and vascular dementia) cases constituting a further 27%. Estimated to account for just 8% of cases are, in frequency order, dementia with Lewy bodies (DLB), frontotemporal dementia (FTD), and Parkinson's dementia. There is no estimate given for semantic dementia (SD), but since it is a subtype of FTD the numbers might be assumed to be very small, relative to cases of AD. Worldwide, by far the most common diagnosis of dementia is Alzheimer's disease, accounting for 50-75% of cases (source: Alzheimer's Disease International). Two-thirds of dementia patients are women, possibly due to their longer life expectancy. In 2013, dementia was the most common cause of death of women in England and Wales (source: Office of National Statistics).

### 6.3 DAT: Dementia of the Alzheimer's Type

"Dementia of the Alzheimer's Type" (DAT) was the collective term used in DSM-4 (American Psychiatric Association, 4th edition, 2000) to encompass the most common forms of dementia. Though the term is absent from DSM-5, it has been widely adopted by researchers. The discussion in this chapter will assume that prior research into DAT is equally relevant to AD.

There is no clearly defined boundary between the (usually gradual) onset of dementia and the normal effects of ageing. A challenge for clinicians is to distinguish diagnostically between 'full blown' dementia and the mild cognitive impairment (MCI) that is a common feature of ageing but might also be a precursor of dementia. MCI is characterised by an evident decline in some (but not necessarily all) cognitive abilities, typically affecting memory, language, attention, judgment and/or decision-making, but without serious functional impairment of day-to-day activities.

Although AD accelerates the loss of linguistic function, there is an underlying component of that decline that is related to age and that applies generally to 'healthy' adults. In a longitudinal study of participants' spontaneous speech, Kemper, Thompson and Marquis (2001: 610) observed an age-related decline in both grammatical complexity and propositional content (the latter is equivalent to the measure of idea density in the Nun Study, see below). "A period of relative stability is followed by a period of accelerated decline and by a third period of more gradual decline", even in healthy adults (*ibid*: 610). This normal, age-related decline should be borne in mind when considering the methodologies and results of linguistic analysis.

The typical progression of AD is from a relatively early impairment of episodic memory (the recollection of events that are personal to the patient), to later impairment of semantic memory for vocabulary and concepts (Harnish and Neils-Strunjas, 2008: 50). The decline of semantic memory characteristically follows a gradual deterioration but then experiences a severe drop in performance, a "semantic cliff" (Strain *et al*, 1998). Chertkow *et al* (2008) have provided a comprehensive literature review of studies relating to the impairment of semantic memory in AD patients.

Whilst the dominant pathology of Alzheimer's disease (AD) is loss of memory, there are also impairments of other cognitive abilities, including language: "Essential to the diagnosis of dementia is the presence of cognitive deficits that include memory impairment and at least one of the following abnormalities of cognition: aphasia, agnosia, apraxia, or a disturbance in executive function" (First and Tasman, 2004: 275).

Although DSM-5 makes no distinction between the different subtypes of dementia based on linguistic factors, aphasia, the broad term for impairment of language production and comprehension, is a diagnostic criterion in all forms of dementia (First and Tasman, 2004), though the specific manifestations of the aphasia are diverse. Symptoms include loss of vocabulary and difficulties with semantic processing and word-finding (Thomas *et al*, 2005: 1570). Tests for AD that rely upon patients' spelling of words, homophones and pseudo-words are widely used, but they are dependent for their validity on accurate knowledge of the patients' pre-diagnosis level of literacy (Harnish and Neils-Strunjas, 2008: 53), and such data are often unavailable.

Ahmed *et al* (2013: 3735) found that "subtle changes in spoken language may be evident during prodromal stages of Alzheimer's disease", i.e. during the periods between the first indications of possible symptoms and the progressive onset of the disease. The detection of these linguistic changes preceded the clinical diagnosis of AD by an average of 12 months, although "the abnormalities found were heterogeneous rather than conforming to a common profile" (*ibid*).

## 6.4 Semantic Dementia

We have seen (in 6.1) that there are many subtypes of 'dementia'. The application of the genitive ratio set out in this chapter is specific to DAT. This discussion of semantic dementia will support the proposition that 'all dementias are not the same'.

Semantic dementia is a clinical variant of frontotemporal dementia (FTD; Hodges, 2007: 2), also known as Pick's disease, a chronic neuro-degenerative condition caused by atrophy of the temporal lobe of the brain. The DSM-5 diagnostic criteria for major or mild FTD specify two variants: behavioural and language. The diagnosis of language variant FTD looks for "Prominent decline in language ability, in the form of speech production, word finding, object naming, grammar, or word comprehension" (American Psychiatric Association, 2013: 615). Language variant FTD is known to psycholinguistic and neurolinguistic researchers as semantic dementia (SD), and to clinicians as semantic variant primary progressive aphasia (svPPA).

Hoffman, Meteyard and Patterson (2013) have analysed transcriptions of 'autobiographical memory interviews' with patients who have been diagnosed with SD/svPPA. These patients retained their cognitive abilities (until the later stages of the disease) and their linguistic fluency, but experienced a progressive loss of the capacity to remember concepts, faces, objects, and specific words. The range of the vocabulary deployed by the patient narrows over time, resulting in an increasing reliance on more familiar words, and on superordinate 'light nouns', such as *person*, *place*, *stuff*, *thing* and *type* (Hoffman *et al*, 2013: 9). These 'whatsit words' are non-specific and can be applied to a wide range of contexts in which the more precise word cannot be recalled. Hoffman *et al* (2013) conclude that "SD is not really a word-*finding* difficulty so much as a word-*knowing* difficulty" (*ibid*: 2).

### 6.5 Can Linguistic Analysis Predict Dementia?

The Nun Study is a longitudinal study of ageing and dementia, based on two communities of Roman Catholic nuns from convents of the School Sisters of Notre Dame in the USA (Snowdon *et al*, 1996). Between 1991 and 1993, researchers gained the consent of 678 nuns, all born before 1917, to participate in a study of ageing. Two factors distinguish the Nun Study from other studies. First, many potentially confounding lifestyle and environmental factors have been eliminated by studying a population of nuns. Second, the participants consented not only to annual cognitive tests and physical examinations, but also to bequeath their brains for autopsy, thus facilitating both pre-mortem cognitive and post-mortem neuro-pathological diagnoses of dementia (Danner, Snowdon and Friesen, 2001: 806).

Some time after the inception of their study, the researchers discovered that many of the participant nuns had been required to write a short autobiographical essay, just before taking their final vows (at an average age of 22). By selecting only those essays that were hand-written (thus verified as the nun's own work) by native speakers of American English, a corpus of 180 autobiographies was constructed (Snowdon *et al*, 1999; Snowdon, Greiner and Markesbery, 2000).

Snowdon and colleagues transcribed a subset of 74 of these autobiographies and manually computed two measures of linguistic ability - idea density and grammatical complexity - as correlates of the nuns' cognitive abilities at that young age: "Prior studies suggest that idea density is associated with educational level, vocabulary, and general knowledge, whereas grammatical complexity is associated with working memory, performance on speeded tasks, and writing skill" (Snowdon *et al*, 1996: 529).

Chand *et al* (2012) define idea density as "a measure of the efficiency in which one communicates ideas". Idea density is also known as propositional density, or P-Density. Engelman *et al* (2010) define a propositional density score as quantifying "the extent to which a person is connecting ideas (via assertions, questions, etc.) rather than merely referring to entities". Farias *et al* (2012) have found that late-life measures of idea density continue to be "associated with steeper subsequent decline in cognitive function" (*ibid*: 683).

Snowdon and colleagues found a strong correlation between the scores from the nuns' annual cognitive tests in old age and the early measure of idea density, less so for grammatical complexity (Snowdon, 2001: 112-114). Idea density therefore became the hub of their analysis. In cases confirmed by brain autopsies, Snowdon (2001: 114) reports an accuracy of "about 80 percent" for idea density in the early-life written samples as a predictor of AD.

The findings of the Nun Study linguistic analysis call into question the widely-held belief that the causes as well as the symptoms of AD typically develop only in later life. They suggest that relatively low idea density, at the age of only 22, predicts (significantly above chance) a relatively higher risk of developing AD by the age of 80. There is neuro-pathological evidence to support the linguistic analysis. Ohm, Müller, Braak and Bohl (1995) autopsied the brains of 887 adults aged from 20 to 100 years, and found pathological evidence of the progressive development of the neurological characteristics of AD "that may even extend into adolescence" (*ibid*: 209).

If we accept the findings of Ohm *et al*, there are two possibilities. Early-life evidence of relatively low cognitive ability might indicate early development of the lesions, the 'tangles and threads', found post-mortem in the brains of Alzheimer's patients. Or, relatively high cognitive ability from an early age (whether innate or acquired) might slow down the development of those lesions. Snowdon *et al* (1996: 532) conclude that "Regardless of the mechanism, our findings indicate that low linguistic ability in early life is a potent marker of both Alzheimer's disease risk and the extent of Alzheimer's disease lesions present at death".

Access to the Nun Study archive is currently restricted, although there is a stated intention to "make the materials from the Nun Study an international scientific and teaching resource" (Lim, 2011). An article about the project in the

New York Times (7 May, 2001) quoted very brief extracts from two of the autobiographical essays that were analysed by Snowdon and colleagues (Riley *et al,* 2005). The first is from a nun then (in 2001) "in her late 90's", who showed strong cognitive evidence of late-stage AD:

"After I left school, I worked in the post-office".

The second is from Sister Nicolette, then aged 93 and showing no signs of cognitive decline:

"After I finished the eighth grade in 1921 I desired to become an aspirant at Mankato but I myself did not have the courage to ask the permission of my parents so Sister Agreda did it in my stead and they readily gave their consent".

Whilst this is in no way a representative sample, the differences between the two accounts of the same life-event are stark. The first is minimal and concrete, as well as low in idea density. The second is much richer in detail and vocabulary, and uses abstract nouns (underlined).

An associated study by Danner, Snowdon and Friesen (2001) analysed the emotional content (positive and negative words) of the nuns' autobiographical essays. Given that the young novitiates were presumably very positive about following their vocation when they wrote the essays, the positive emotional content was predictably high. Nevertheless, Danner *et al* found that the degree of positivity was inversely associated with the onset of AD 60 years later.

## 6.6 Linguistic Factors in the Diagnosis of Alzheimer's Disease

Mehl and Gill (2010: 109) point out that there are several alternative labels for what they term automatic text analysis – computer content analysis, computerassisted content analysis, and computerised text analysis. Because the ensuing discussion will cover a broad range of methods and approaches, this thesis will adopt a more general term: computational linguistic analysis.

Whatever the labels, the processes they describe are similarly reliant on statistical models, based on the frequency of particular lexical or syntactic units within a text. Where these models differ is in their selection of the specific variables for analysis. Table 6.2, reproduced from Le *et al* (2011: 439), tabulates the patterns of change in language usage that are typical in normal ageing, compared with cases of dementia. To facilitate computational linguistic analysis, these patterns must be translated into specific measures.

**Table 6.2**: Patterns of linguistic changes expected in normal ageing and in dementia

| **Linguistic marker** | **Normal ageing** | **Cases of dementia** |
|---|---|---|
| *Lexical* | | |
| Vocabulary size | Gradual increase, possible slight decrease in later years | Sharp decrease |
| Repetition | Possible slight decrease/increase | Pronounced increase |
| Word specificity | Possible slight decrease/increase | Pronounced decrease |
| Word class deficit | Insignificant change | Pronounced deficit in nouns: possible compensation in verbs |
| Fillers | Possible slight increase | Pronounced increase |
| *Syntactic* | | |
| Complexity | Gradual or no decline, possibly rapid around mid-70's | Sharp decline |
| Use of passive | Possible slight decrease | Pronounced decrease |
| Auxiliary verb | *Be*-passives are dominant | *Get*-passives are dominant |
| Agentless passive | Moderate decrease | Greater decrease |

Drawing on previous work by Bucks *et al* (2000) and by Thomas *et al* (2005), Baldas *et al* (2011: 108-109) have compiled a useful summary of lexical measures, of rates and ratios that might be combined to monitor the development and progression of AD. The object of their analysis is a transcribed passage of continuous speech, containing the total number of words spoken (*N*), and the number of different words, representing vocabulary size (*Voc*). A part-of-speech (POS) tagger identifies nouns, pronouns, adjectives and verbs, facilitating the computation of seven measures, in three of which AD patients typically perform at a higher rate than 'neurotypical' controls (AD ≥ nC), with the opposite direction (AD ≤ nC) in the other four measures. Table 6.3, adapted from Baldas *et al* (2011: 108-109), summarises the metrics, their method of calculation and direction.

**Table 6.3**: Lexical measures of Alzheimer's disease (AD) vs. 'neurotypical' controls (nC).

| **Metric** | **Calculation** | **Direction** |
|---|---|---|
| Pronoun rate | pronouns/*N* | AD ≥ nC |
| Adjective rate | adjectives/*N* | AD ≥ nC |
| Verb rate | verbs/*N* | AD ≥ nC |
| | | |
| Noun rate | nouns/*N* | AD ≤ nC |
| Type-Token-Ratio (TTR) | *Voc*/*N* | AD ≤ nC |
| Brunét's index (W) | *N*^*Voc*-0.165 | AD ≤ nC |
| Honoré's statistic (R) | 100 log*N*/(1-$V_1$/*Voc*)* | AD ≤ nC |
| *$V_1$ = words used only once | | |

Two of these measures perhaps require explanation:

- Brunét's index (W) is a version of the type-token ratio (TTR – see below) that is not sensitive to the length of the text (Brunét, 1978).

- Honoré's statistic (R) is a measure of the richness of vocabulary (Honoré, 1979), which is indicated by a comparatively higher value of R. The premise for this statistic is that the count of words that occur only once ($V_1$) within a text provides a measure of the richness of the originator's lexicon.

Probably the most commonly deployed index of lexical diversity is the type to token ratio (TTR). The 'type' counts the first occurrence of every word in a text. The TTR is calculated by dividing the number of types by the total number of words (tokens) used, as exemplified in Table 6.4.

**Table 6.4**: Calculation of the 'type to token ratio' (TTR)

| **Text** | **Types** | **Tokens** | **TTR** |
|---|---|---|---|
| *This is cold* | 3 | 3 | 1.00 |
| *This is hot and this is cold* | 5 | 7 | 0.71 |

The TTR has been applied not only to English-language texts, but also to the Dutch author Gerard Reve (1923-2006), who was diagnosed with AD shortly after completing his final novel. Van Velzen and Garrard (2008) compared the mean TTRs for three novels by Reve, written in the early, middle and final stages of his career. They found not only a significantly lower TTR in Reve's final novel, but also a "dramatic" drop between the first and second halves of that final novel (a within-text difference that was not detected in the other novels).

From an analysis of speech samples obtained from 15 patients with AD (diagnosis confirmed post-mortem), Ahmed, Haigh, de Jager and Garrard (2013: 3735) concluded that micro-measures of linguistic performance were not sufficiently reliable markers for AD. They derived and tested five "more robust" composite measures. Of these, three were found to be significant in tracking the progression of AD: lexical content, syntactic complexity, and semantic content. The other two, speech production and fluency, were not significant. Table 6.5 sets out the measures included in each of the three significant composites.

**Table 6.5**: Components of three composite markers of Alzheimer's disease progression (adapted from Ahmed, Haigh, de Jager and Garrard, 2013: Figure 1)

| **Lexical content** | **Syntactic complexity** | **Semantic content** |
|---|---|---|
| Pronouns | Syntactic errors | Total units |
| Verbs | Words in sentences | Subjects |
| | Nouns with determiners | Objects |
| | Verbs with inflections | Actions |
| | Mean length of utterance | Idea density |
| | | Efficiency |

The studies published to date, and sampled in this chapter, are essentially exploratory. They test a wide range of linguistic factors and attempt to identify those that should form the critical mass of a linguistic model for the diagnosis of dementia. There is a converging consensus that linguistic analysis has the potential to be an effective indicator or predictor of possible dementia. There is no consensus on the specific factors that should constitute such a model.

### 6.7 Machine Learning Approaches to Diagnosis: Four Studies

The application of computational linguistic analysis to the diagnosis of AD is exemplified by four recent studies. The first three of these have the aim of differentiating between different subtypes of dementia, and rely on speech transcription as their principal data source. The fourth (Pakhomov and Hemmy, 2013) focuses on the early diagnosis of AD, and relies on cognitive test data obtained from the Nun Study archive.

**Jarrold, Peintner, Yeh, Krasnow, Javitz and Swan (2010)**

The longitudinal Western Collaborative Group Study (WCGS; now closed) collected data on patients with cardiovascular problems over more than 40 years. From the WCGS population, Jarrold *et al* (2010) obtained a sample of participants who had recorded a structured interview (15 minutes) in or around 1988; who had at the same time completed the Iowa Screening Battery for Mental Decline (ISBMD), a cognitive test that included word-list generation (Eslinger *et al*, 1984); and whose subsequent cause of death had been clinically classified as AD.

The ISBMD test scores classified participants (with a mean age of 73.13 at the time of the structured interview) as either cognitively normal (score of 0) or cognitively impaired (score >8). Jarrold *et al* (2010: 302) thus identified three distinct participant groups, as set out in Table 6.6. This enabled both pre-symptomatic AD (pre-AD) and cognitively impaired (CI) participants to be compared with controls. The structured interviews of these three participant groups were then transcribed.

**Table 6.6**: Classification of WCGS participants by Jarrold *et al* (2010)

| **Cognitive test result** | **Cause of death** | **Participant group** |
|---|---|---|
| Normal | Alzheimer's disease | Pre-symptomatic AD (pre-AD) |
| Impaired | Other | Cognitively impaired (CI) |
| Normal | Other | Controls |

From the transcriptions of the structured interviews, lexical features were extracted to a vector via three analysers: a part-of-speech tagger, LIWC (see chapter 7), and CPIDR (Computerized Propositional Idea Density Rater, pronounced 'spider'). The individual vectors, labelled with the appropriate diagnosis, constituted the training set for a classifier (machine learning) module. The linguistic analysis by Jarrold *et al* (2010: 303) correctly identified 73% of the participants who had achieved a 'normal' score on the ISBMD cognitive test, but who subsequently received a clinical diagnosis of Alzheimer's disease. Accurate prediction of cognitive impairment was 82.6%. The study is a relatively successful proof of concept, though the authors readily acknowledge that their sample sizes are small.

**Jarrold, Peintner, Wilkins, Vergryi, Richey, Gorno-Tempini and Ogar (2014)**

Jarrold *et al* (2014) exemplify the application of machine learning algorithms to a diagnostic model that aims (in their case) to distinguish AD from the three clinical subtypes of frontotemporal dementia (FTD): semantic dementia (SD), the behavioural variant of FTD (bvFTD), and progressive nonfluent aphasia (PNFA). Their data-source was a set of speech-samples (each 3-5 minutes) obtained from a semi-structured interview and picture description task. Their system then extracted acoustic features (41 duration-based measures) and lexical data obtained from transcriptions. Part-of-speech and LIWC feature profiles (14 and 81 categories respectively) were then extracted from the lexical data. From these feature profiles (acoustic and lexical), the system constructed a vector of the features that represented an individual speaker, for input to the classifier, which had been trained on vectors with diagnostic labels attached.

Their participant pool was relatively small: nine AD patients, nine with bvFTD, thirteen with SD, eight with PNFA, all undergoing treatment at the UCSF Memory and Aging Center, and nine "age-matched healthy" controls.

Nevertheless, the model's diagnostic accuracy compares reasonably well with questionnaire or inventory-based diagnostic instruments:

| | | | | | |
|---|---|---|---|---|---|
| AD | vs. Controls | | | | 88% |
| AD | vs. FTD | | | | 88% |
| AD | vs. FTD | vs. Controls | | | 80% |
| AD | vs. SD | vs. bvFTD | vs. PNFA | vs. Controls | 61% |

Jarrold *et al*(2014) are careful to put these results into perspective. The system should be trained and tested on much larger numbers of participants, and should complement rather than replace other sources of information that would be assimilated by clinicians in order to form a diagnosis. Their method is, in their own words, "fast, inexpensive, and non-invasive", and "may show most promise as a screening tool to decide which patients need deeper evaluation" (*ibid*: 34). However, this ignores the most evident limitation of the system: that it relies upon a multitude of different measures and features.

**Fraser, Hirst, Graham, Meltzer, Black and Rochan (2014)**

Studies based on machine learning classifiers are differentiated principally by the number and range of different feature sets that they apply. Perhaps the most comprehensive to date is that of Fraser *et al* (2014), who defined seven feature sets with 189 different features, in a speech-based analysis of progressive aphasia:

| | | |
|---|---|---|
| 13 | POS | Parts of speech features (counts and ratios) |
| 11 | C | Complexity features (e.g. mean length of sentence and clause) |
| 134 | CFG | Context-free grammar production rule features |
| 5 | F | Fluency features (e.g. words per minute) |

|  |  |  |
|---|---|---|
| 5 | P | Psycholinguistic features |
| 4 | VR | Vocabulary richness features (e.g. Honoré's statistic) |
| 17 | A | Acoustic features (e.g. pauses) |

The aim of Fraser *et al* (2014) was to use computational linguistic analysis to distinguish between two subtypes of progressive aphasia: semantic dementia (SD) and progressive nonfluent aphasia (PNFA); and to distinguish both from controls. A standard feature selection method involved the iterative deletion of the least significant feature set, i.e. the set that contributed least to the measure of accuracy, for each comparison. In each case, a combination of two feature sets was found to yield the highest accuracy:

|  |  |  |  |
|---|---|---|---|
| SD | vs. controls | P + POS | 1.00 |
| PNFA | vs. controls | P + A | 0.97 |
| SD | vs. PNFA | P + CFG | 0.92 |

The constant is the psycholinguistic feature set (P), a set that contains measures of frequency, familiarity, imageability, age of acquisition, and 'light verbs', i.e. verbs such as *have, go, do, get, put*, that generalise more specific verbs. As with Jarrold *et al* (2014), the disadvantage of this approach lies in the scale of the individual features that would need to be analysed. The three comparisons respectively deploy 18, 22 and 139 different features.

**Pakhomov and Hemmy (2013)**

For those given access to the data, the Nun Study has provided a rich data-pool for parallel research, including this application of computational linguistics by Pakhomov and Hemmy (2013). Their data were longitudinal sets of the semantic verbal fluency (SVF) tests administered to participants in the Nun Study

(Snowdon, 2001). The SVF test requires participants to name as many words as possible from a given semantic category, within a time limit (usually one minute). The responses elicited from participants typically form clusters. So, the category *birds* might stimulate clusters of garden birds (*robin*, *sparrow*, *starling*), birds of prey (*eagle*, *owl*, *hawk*), and birds bred for food (*chickens*, *ducks*, *geese*). Analysis of the results obtains two measures: the size of the clusters, and the number of transitions ('switches') between the clusters.

Prior studies (see Pakhomov and Hemmy, 2013: 2 for references) have correlated each of these measures with activity in a particular area of the brain: the left temporal lobe (cluster size) and the frontal lobe (switching). Based on these correlations, the same prior studies suggest that the combination of these two SVF measures might "index the strength of associations in the patient's lexical-semantic networks" (*ibid*: 2). Furthermore, these cluster measures (size and switching) are known to decline progressively in patients diagnosed with Alzheimer's disease.

Pakhomov and Hemmy (2013) automated and objectified the process of quantifying the cluster measures, using Latent Semantic Analysis and comparative clustering data obtained from the Wikipedia corpus. Their results suggest that the trend of the SVF data constitutes a significant predictor of the development of dementia, in patients who were otherwise "cognitively intact" (*ibid*: 6) at the baseline of their SVF assessments.

## 6.8 The Impairment of Abstract Thinking in Alzheimer's Disease

Findings of a significant impairment of abstract thought and language in cases of dementia, leading to a consequent dominance of concrete thought and language, are well-established (e.g. Bayles and Tomoeda, 1983: 111; Murdoch, 1990: 179); Baudic *et al*, 2006: 18; and Amanzio *et al*, 2008: 2).

Lezak *et al* (2004: 569) identify 'concrete thinking' as "the most common sign of impaired conceptual functions". Concrete thinking is associated with impairment of the ability to manipulate concepts, to categorise and to generalise, all of which are features of abstract thinking. In cases of dementia, not only the patient's thinking but also their language becomes progressively more concrete (Jacques and Jackson, 2000: 275).

Although the decline of abstract problem-solving abilities is typical of ageing, it is much more pronounced in AD patients, and from an early stage of the disease: "Impairments on tests of memory and abstraction are often the only deficits in mild AD" (Zec, 1993: 21). In a series of 'free recall' experiments, Rissenberg and Glanzer (1987) tested both 'old normal' controls and patients with a diagnosis of dementia of Alzheimer's type (DAT). In the DAT group, they found "a significant recall advantage for concrete over abstract words" (Rissenberg and Glanzer, 1987: 322).

Crutch and Warrington (2006: 487) concluded, from a synonym comprehension test, that both SD and AD patients "have conceptual knowledge systems that, to a greater or lesser degree as a consequence of brain disease, appear to be damaged in such a way that specific or detailed information about abstract concepts is more affected than cruder broad-sense information." Given that such "crude" concepts will be simple and in common use, it is possible that the concreteness effect observed in dementia is, at least to some degree, a frequency effect (Bates *et al*, 1995: 495).

In a wide-ranging review of prior studies, Chertkow *et al* (2008) found substantial evidence of a category-effect impairment to semantic memory in AD patients. The categories that patients find more problematic than others are fruits and vegetables, biological items such as animals, and abstract nouns. In one such study, Fung *et al* (2000) compared the performance, on a semantic association judgment test, of nine DAT patients and eleven controls matched by age, gender and level of education. They tested 150 target words from six semantic categories: animals, clothing and furniture, fruits and vegetables, tools, verbs, and abstract nouns. The overall performance of the DAT patients was significantly worse (73%) than that of the controls (88%), but their errors showed an "effect of category" for abstract nouns, animals, and fruits and vegetables. The categories of verbs, tools, clothing and furniture showed no significant difference. Chertkow *et al* (2008: 404) conclude that, whilst the evidence of a category effect is reliable, "the theoretical explanation for this phenomenon appears difficult to pin down".

A number of studies have identified a "reverse concreteness effect" in cases of semantic dementia (SD), i.e. the opposite of the general finding of a concreteness effect in Alzheimer's disease (see Reilly, Troche and Grossman, 2014, for references). However, Jefferies *et al* (2009: 493) point out that these findings represent only a small number of patients. Other studies (Crutch and Warrington, 2006; Pulvermüller *et al*, 2008) have found greater impairment of abstract than concrete words in SD patients. The reverse concreteness effect could be caused by individual differences such as the level of educational attainment, but is more likely to result from atypical atrophy to specific areas of the brain that are associated with the visual representation of objects and their features. Jefferies *et al*'s own study, based on a synonym judgment task, found a consistently greater impairment of abstract than of concrete nouns, in all of their SD-diagnosed participants.

Although there is converging psycholinguistic and clinical evidence that abstract thinking and language are progressively impaired in cases of AD, those measures do not generally feature in clinical diagnostic tests. The Mini-Mental State Exam (MMSE), or Folstein test, was introduced to clinical diagnosis in 1975 and is still widely used (e.g. Ahmed *et al*, 2013). It is a simple test of cognitive ability (30 questions, taking about ten minutes), that does not require specialist expertise to administer it. Whilst the MMSE does not constitute a comprehensive assessment, it does provide an initial evaluation of possible cognitive deficits, and it potentially differentiates between different types of dementia, including mild cognitive impairment. No component of the test assesses the patient's capacity for abstract thought.

Ahmed *et al* (2013: 3729, Table 1) analysed transcriptions of connected speech samples obtained from two matched groups: 15 healthy controls, and 15 participants who progressed from mild cognitive impairment (MCI) to a next-stage diagnosis of mild AD. All 30 participants completed the CAMCOG battery of cognitive tests, which does include a test of abstract thinking with a maximum score of 8. The scores (mean and standard deviation) for each group were significantly different:

| | | | |
|---|---|---|---|
| [6.1] | Controls | 7.6 | sd 0.9 |
| | MCI | 6.6 | sd 1.5 |
| | Mild AD | 5.7 | sd 2.0 |

A separate study by Ahmed *et al* (2013), of only nine participants per group, extended the analysis to a third clinical stage (moderate AD). With such small samples, this failed to reach significance in the measure of abstract thinking, but the progression of decline is in the anticipated direction (Ahmed *et al*, 2013: 3729, Table 2):

| | | | |
|---|---|---|---|
| [6.2] | Controls | 7.3 | sd 1.1 |
| | MCI | 7.0 | sd 1.3 |
| | Mild AD | 5.8 | sd 2.1 |
| | Moderate AD | 5.1 | sd 2.9 |

## 6.9 Linguistic Evidence of Dementia from Published Texts

Harnish and Neils-Strunjas (2008: 44) observe that research on the language (reading and spelling) of Alzheimer's patients has been predominantly at the level of the single word. A widely reported exception to that was a study by Garrard, Maloney, Hodges and Patterson (2005) that looked at whether evidence of the onset of dementia (specifically Alzheimer's disease) might be derived from a computational linguistic analysis of published works of fiction. Their analysis encompassed structural measures, vocabulary, syntactic differences and lexical characteristics. Their subject for a "within-patient comparison" was the novelist and philosopher Iris Murdoch, who died in 1999 at the age of 76.

The materials selected for analysis were three of Murdoch's novels: her first, *Under the Net* (1954); her most acclaimed work *The Sea, The Sea,* written in mid-career (1978); and her final published work *Jackson's Dilemma* (1995). Based on cognitive tests, Iris Murdoch had received a diagnosis of the onset of dementia in 1996. Post-mortem examination subsequently confirmed the diagnosis of Alzheimer's disease (AD).

Perhaps because her diagnosis of AD was clinically beyond doubt, Iris Murdoch's novels have provided the data for a number of studies following in the wake of Garrard *et al* (2005). A Computerized Linguistic Analysis System (CLAS), developed by Pakhomov, Chacon, Wicklund and Gundel (2011) to provide metrics of the syntactic complexity of a text, has been applied to Murdoch's novels. Their finding of "clear patterns of decline in grammatical complexity" (*ibid*: 136) supported the findings of Garrard *et al* (2005).

Also with the study by Garrard *et al* (2005) as their precedent, Lancashire and Hirst (2009) conducted a comparable analysis of novels written by Agatha Christie (1890 – 1976). Christie published over 80 novels and plays between 1920 and 1973. Lancashire and Hirst used concordance software to analyse and compare elements of the language of 16 Christie novels. As far as is known there was never a clinical diagnosis of dementia in Christie's case. The evidence is therefore anecdotal, but also based on a literary critique: "her last novels reveal an inability to create a crime solvable by clue-deduction according to the rules of the genre that she helped to create" (Lancashire and Hirst, 2009: 2). The anecdotal evidence derives from her biographers, one of whom (Morgan, 1984: 374) notes that in her final years Christie became childlike, eccentric and "often difficult", at one point hacking off her own hair. By 1975, just prior to her death, "her beautiful brain was fragmenting" (Thompson, 2007: 483), she was experiencing sudden and significant changes of behaviour, and her conversation often "made no sense".

Le *et al* (2011) have added to the 16-novel Christie analysis by Lancashire and Hirst (2009), with a dissection of 20 novels by Iris Murdoch and 15 novels by P.D. James, who was selected as a novelist born in 1920 and still productive in her 90s, with no evidence of cognitive decline. Pre-processing of all 51 novels was extensive, involving (*inter alia*) separation of punctuation marks and clitics from word-tokens, lemmatisation, parsing and part-of-speech tagging. The analysis by Le *et al* encompassed "vocabulary size, repetition, word specificity, word-class deficit, fillers, grammatical complexity, and the use of passive" (*ibid*: 438), a total of seven syntactic and lexical linguistic markers.

The syntactic analysis by Le *et al* (2011) was largely inconclusive, but their lexical analysis identified, as markers of AD in the novels of both Murdoch and Christie, a loss of vocabulary, increased word repetition, and a word-class deficit that was characterised by a loss of noun tokens with some compensation by verb tokens. They found that these markers were either absent or marginal in the novels of P.D. James.

These text-based studies are not without their critics. Van Velzen, Nanetti and de Deyn (2014) applied data modelling techniques of permutation testing and Akaike Information Criterion to assess the quality of the statistical analyses in the preceding authorial studies of Murdoch, Christie and James. They looked specifically at the factors of lexical diversity (as measured by the type-token ratio), and the noun-pronoun ratio. They consider the latter to be a "finer-grained" analysis of lexical diversity (*ibid*: 194), since it measures the progressive displacement of proper names and precise nouns in favour of less explicit pronouns. Their "more sophisticated analysis" (*ibid*: 200) supports the findings of Garrard *et al* (2005) and of Van Velzen and Garrard (2008), but not those of Le *et al* (2011). Their own analysis of Christie's novels supports an admittedly speculative diagnosis of semantic dementia, rather than Alzheimer's.

In pursuit of what Garrard (2009) has termed "cognitive archaeology", the computational linguistic analysis of historical cases of dementia has been extended beyond literary works, to the language of politics (Garrard, 2009) and even to the letters of a king (Williams *et al*, 2003).

Using transcripts from the parliamentary record (Hansard) of unscripted exchanges in the House of Commons, Garrard (2009) has compared the language of prime minister Harold Wilson (1916-1995) to that of other contemporary politicians, and has presented linguistic evidence of the cognitive decline which most likely led to Wilson's surprise resignation in 1976.

Williams *et al* (2003) have analysed 57 letters written between 1604 and 1624 by King James I of England (James VI of Scotland; 1566-1625). Their analysis indicates a decline in the letters' syntactic complexity which is not matched by an AD-typical semantic decline, leading them to a tentative diagnosis of vascular dementia, though tempered by acknowledged "gaps in the data, lack of a cohort for comparison, and the unknown applicability of modern linguistic analysis to Elizabethan writing style" (*ibid*: 44). Given these rather significant caveats, it is difficult to give much credence to their analysis.

With the other prior studies as benchmarks, this chapter will present genitive ratio analyses of texts sampled from a politician (Harold Wilson) and from five authors of fiction, writing in English. Three of those authors (Agatha Christie, P.D. James and Ruth Rendell) have written predominantly crime fiction.

Iris Murdoch wrote 'literary' fiction. Only one (Murdoch) had a clinical diagnosis of AD, confirmed post-mortem. The evidence for Christie's dementia is strong but anecdotal, and it might not have been AD. James and Rendell are included as healthy controls, with no signs of cognitive decline and still writing in their 90s and 80s respectively.

These four authors thus prompt two hypotheses, that the concreteness of their language in their later works will:

| | |
|---|---|
| Increase | (Murdoch and Christie) |
| Not change | (James and Rendell) |

All of the genitive ratio analyses have been facilitated by the Animyser program (see chapter 4). The sample sizes are similar to those of Garrard *et al* (2005), though the method of sampling is arguably more efficient and more economical.

### 6.10 The Dialogue Debate

"It's very hard to write a novel and not reveal a great deal about yourself. Even if it's not autobiographical, it's always personal".
(David Nicholls)

The question of whether dialogue should be included in the analysis of fictional works is of some consequence. The exclusion of dialogue adds significantly to the pre-processing of textual data for analysis. The decision by Garrard *et al* (2005) was to exclude dialogue, on the basis that such text might not be in the author's own 'voice'. Pakhomov *et al* (2011: 139) followed Garrard *et al* (2005) in excluding dialogue from their own analysis of Iris Murdoch's novels, because it "constitutes a different type of discourse". Le *et al* (2011) take the opposite view. They argue that "fiction authors' … characters' conversational styles arguably reflect, to some extent, their own styles" (*ibid*: 449). Perhaps because she was a playwright as well as a novelist, Agatha Christie's crime fiction relies very heavily on dialogue.

This prompts a question: does the exclusion of dialogue justify the increased pre-processing cost? Based on statistical analysis, the answer is 'possibly'. A Student *t*-test compared two samples, both extracted from Agatha Christie's *The Mysterious Affair at Styles* (1920). The first sample is of the CRs of 400 nouns (excluding dialogue) randomly selected for the analysis reported in 6.12. The second sample is of the CRs of 8,292 nouns extracted by the Animyser program's POS tagger from the complete text (i.e. with dialogue included) of the same novel, which is available in digitised form from Project Gutenberg. The null hypothesis, that there would be no significant difference between the two samples, was not rejected [*t* (8690) = 1.38, *p* = .084]. This finding supports the decision by Garrard *et al* (2005) to exclude dialogue and direct quotation from their analysis.

Whether the exclusion of dialogue justifies the significant additional pre-processing cost is ultimately down to the judgment of the researcher. A novelist will typically try to present a wide variety of characters of different ages, genders, social classes and occupations, and must invent dialogue for each character accordingly. If they are all in the novelist's 'own voice', they will not be convincing and the novelist will have failed. Clearly, none of the novelists analysed here has failed.

The sampling method adopted for the analyses reported in this chapter is simple, low-cost, and the manual identification of dialogue is facilitated. The genitive ratio analysis will therefore follow Garrard *et al* (2005), and will exclude dialogue.

### 6.11 Three Novels by Iris Murdoch

"[Our GP] asked Iris who the Prime Minister was. She had no idea but said to him with a smile that it surely didn't matter".
(John Bayley: *Iris*, 1998: 151)

"Language style" has been defined by Pennebaker and King (1999) as an individual difference. The vocabulary deployed by individuals normally remains "remarkably reliable across time and situations" (*ibid*: 1308), and Iris Murdoch was noted for her adamant rejection of any editorial revisions to her work. The analysis by Garrard *et al* (2005) encompassed the diversity of vocabulary, syntactic complexity, and lexical differences (word length and frequency) identified across three novels.

Their main findings concerned the reduced range and variety of vocabulary and "lexical selection" in *Jackson's Dilemma*, when compared with the earlier novels (*ibid*: 258). They found that Murdoch's decline into dementia was preceded by a decline in lexical diversity, resulting in "a smaller, higher frequency vocabulary" (*ibid*: 259). Le *et al* (2011) have gone further, claiming to have found a sharp decline in Iris Murdoch's writing, not simply in *Jackson's Dilemma*, but actually in the course of writing that novel.

The sampling method of Garrard *et al* (2005) involved converting each of three Iris Murdoch novels into a text file, with dialogue and direct quotation flagged for exclusion from their analysis. They then used the Concordance (v3.0) software to create a word list for each novel, in alphabetical order and with a frequency count of each word. The Concordance software then selected five random samples of 100 words from each novel: a total of 15 word lists. Since this process resulted in some duplication of words for each novel, the actual totals of different 'word types' for each novel were:

| | |
|---|---|
| *Under the Net* | 379 |
| *The Sea, the Sea* | 373 |
| *Jackson's Dilemma* | 352 |

These words (total 1,104) covered the full range of grammatical categories, and there is no information as to the number of nouns included in that total.

The much simpler process designed for the genitive ratio (GR) analyses in this chapter  randomly identified 400 nouns for analysis from each novel. A random number generator selected ten page numbers from each novel, and ten line numbers (in a range 1-30) to be applied consecutively to the ten page numbers. Starting at the designated line on each selected page, the next 40 nouns were manually identified and listed. Proper nouns were not listed. As in Garrard *et al*, dialogue and direct quotation were excluded. The Animyser computer program was then used to obtain CR counts from Wikipedia for each of the nouns listed (see chapter 4 and Appendix 4.1 for the program and method). A mean CR score for each novel was then calculated. This method was replicated for each of the authors who are case-studied in this chapter.

Table 6.7 shows the results of the concreteness rating analysis. Mean word length is included in this and in subsequent analyses, to show that it was not a significant factor, and neither was it in the study by Garrard et al (2005).

**Table 6.7: Iris Murdoch: Genitive Ratio Analysis**

**Prediction**: The mean CR will be significantly higher in the later novel, indicating an increase in the use of concrete language

| **Novel** | **Date** | **Word Length** | **Mean CR** |
|---|---|---|---|
| *Under the Net* | 1954 | 5.8 | +1.545 |
| *The Sea, the Sea* | 1978 | 5.4 | +1.398 |
| *Jackson's Dilemma* | 1995 | 5.4 | +2.197 |

CR comparison of the three Murdoch novels by t-tests shows no significant difference between *Under the Net* and *The Sea, the Sea* (*t* (798) = 0.42, *p* (one-tailed) = .338). As predicted, *Jackson's Dilemma* is significantly different from both *Under the Net* (*t* (798) = 2.07, *p* (one-tailed) = .019) and *The Sea, the Sea* (*t* (798) = 2.31, *p* (one-tailed) = .011).

## 6.12 Watching the Detectives: Christie, James and Rendell

### Agatha Christie

Table 6.8 is an extract of the results from the concordance analysis by Lancashire and Hirst (2009). Of the 16 Christie novels that they analysed, the extracted analysis focuses on three: two early works and one late work.

**Table 6.8**: **Linguistic analysis of Agatha Christie novels** (extracted from Table 1, Lancashire and Hirst, 2009)

|  |  |  | **Repeated** | **Indefinite** |
|---|---|---|---|---|
| **Novel** | **Christie's age** | **Word-types** | **phrase-types** | **words** |
| *Styles* | 28 | 5027 | 7623 | 0.27 |
| *Ackroyd* | 34 | 5576 | 7320 | 0.39 |
| *Elephants* | 81 | 3762 | 8821 | 1.02 |

*The Mysterious Affair at Styles* (1920) was her first published work, written at the age of 28. *The Murder of Roger Ackroyd* (1926) is generally considered to be one of her most accomplished works: "the one that changed her reputation for ever … the supreme, the ultimate detective novel" (Thompson, 2007: 155). *Elephants Can Remember* (1972) was her penultimate novel, "the last novel Agatha wrote before her powers really declined" (Morgan, 1984: 370), and was probably the last book that was all her own work.

Her final novel was *Postern of Fate*, published in 1973. One of her biographers (Morgan, 1984: 371) states that the manuscript was "tidied up" by her husband and her secretary, possibly also by her literary agent and her publisher. It is impossible to determine how representative that final novel might be of her own language production towards the end of her life, and it has therefore been excluded from this analysis.

The data in Table 6.8 track a decline in the "richness" of Christie's vocabulary, as measured by the number of different word-types used in the first 50,000 words of each text. There is a reduction of 21% between the means of the early novels and the late work. This is matched by a 13% increase in phrasal repetition, and a near-fourfold increase in the use of indefinite words (*thing*, *anything*, *something*), from 0.27% to 1.02%. All three measures are symptomatic of a writer who is struggling to match the lexical diversity of her earlier work.

**Table 6.9: Agatha Christie: Genitive Ratio Analysis**

**Prediction**: The mean CR will be significantly higher in the later fiction, indicating an increase in the use of concrete language.

| **Novel** | **Date** | **Word Length** | **Mean CR** |
|---|---|---|---|
| *The Mysterious Affair at Styles* | 1920 | 6.0 | +1.668 |
| *The Murder of Roger Ackroyd* | 1926 | 6.1 | +1.581 |
| *Elephants Can Remember* | 1972 | 5.7 | +2.342 |

Table 6.9 indicates a similar decline in the use of abstract language, as measured by the mean CR of each novel. A CR comparison of the three Christie novels by *t*-tests shows no significant difference between *The Mysterious Affair at Styles* and *The Murder of Roger Ackroyd* [*t* (798) = 0.28, *p* (one-tailed) = .391]. As predicted, *Elephants Can Remember* is significantly different from both *The Mysterious Affair at Styles* [*t* (798) = 2.29, *p* (one-tailed) = .011] and *The Murder of Roger Ackroyd* [*t* (798) = 2.63, *p* (one-tailed) = .004].

### P.D. James

P.D. (Phyllis) James, Baroness James of Holland Park (1920-2014), is best known for her crime fiction featuring Adam Dalgliesh, a Scotland Yard detective who is also a published poet, in 14 novels written between 1962 and 2008. She rebuffed any comparisons with Christie, who was "such a bad writer". James's most recent book was published in 2011. In 2010, aged 90, she won a national journalism prize for the "best broadcast interview of the year", with the then Director-General of the BBC. The three novels sampled for this analysis all feature Adam Dalgliesh.

See Table 6.10. Although the mean CR for the final Dalgliesh novel (The Private Patient) is noticeably higher, indicating a relative increase in concrete language, a one-way ANOVA shows no statistically significant difference between the three novels (*F* (2, 1197) = 2.36, *p* = .095). It is possible that the mean CR is simply an early indicator of a progression that is concomitant with ageing. Flynn (2007: 64) recounts the case of Richard Wetherill, a keen chess player who became alarmed when he could think not eight but only four moves ahead. Cognitive tests revealed no clinical evidence of dementia, but a brain autopsy just two years later revealed evidence of severe AD. Flynn (*ibid*: 65) concludes that we should "hold fast to the image of the brain as a muscle. At any age, an athlete is better off for training; but however hard you train, your times will get slower as you age".

**Table 6.10: P.D. James: Genitive Ratio Analysis**

**Prediction**: There will be no significant differences in the use of concrete language, as measured by the CR.

| **Novel** | **Date** | **Word Length** | **Mean CR** |
|---|---|---|---|
| *Cover Her Face* | 1962 | 6.2 | +1.405 |
| *Devices and Desires* | 1989 | 5.9 | +1.598 |
| *The Private Patient* | 2008 | 5.8 | +2.046 |

## Ruth Rendell

Ruth Rendell, Baroness Rendell of Babergh (1930-2015), was the author of 76 novels, collections of short stories, and works of non-fiction. Her final novel was published in 2014. In 2013 she published her last novel featuring Chief Inspector Wexford (in 24 novels written since 1964). As was her close friend Phyllis James, Rendell was an active member of the House of Lords. The three novels sampled for this analysis all feature Chief Inspector Wexford. As measured by a one-way ANOVA, there were no significant differences between the three Rendell novels (*F* (2, 1197) = 0.28, *p* = .758). See Table 6.11.

**Table 6.11: Ruth Rendell: Genitive Ratio Analysis**

**Prediction**: There will be no significant differences in the use of concrete language, as measured by the CR

| **Novel** | **Date** | **Word Length** | **Mean CR** |
|---|---|---|---|
| *From Doon with Death* | 1964 | 5.8 | +2.095 |
| *The Veiled One* | 1988 | 5.8 | +1.922 |
| *No Man's Nightingale* | 2013 | 5.8 | +1.905 |

### 6.13 Parliamentary Responses: Harold Wilson

Harold Wilson (1916-1995) served as the Prime Minister of a Labour government in two terms of office, from 1964 to 1970, and from 1974 to 1976. With an outstanding academic record (he was appointed lecturer in economic history at Oxford at age 21), he was renowned for his prodigious memory and powers of concentration. His resignation, announced in March 1976, came as a complete surprise to both party and country, and was beset by conspiracy theories. More likely is the speculation that he was becoming aware of a mental decline which would later be diagnosed as AD. His demise at age 79 was attributed on his death certificate to colon cancer and Alzheimer's disease.

Using transcripts from the parliamentary record (Hansard) of unscripted exchanges in the House of Commons, Garrard (2009) compared Wilson' parliamentary responses to those of other politicians, and found evidence of cognitive decline. Garrard's (2009) study was followed up by Cantos-Gómez (2010), who criticised the quality and depth of Garrard's analysis, and conducted his own analysis of Wilson's spontaneous statements as recorded in Hansard, over two specific time-periods, at the beginning and end of Wilson's two terms of office as Prime Minister:

16 October 1964 to 31 December 1964

1 January 1976 to 4 April 1976

Garrard had conducted both a within-subject and a between-subjects analysis, examining Wilson's own language change over time and also comparing his language with that of other Members of Parliament. Cantos-Gómez focused on the "intra-speaker" materials, as offering a more valid assessment of possible linguistic decline, and used discriminant function analysis (DFA) to build a statistical model with AD as the categorical dependent variable and no less than 49 independent variables, including measures of frequency and repetition.

Cantos-Gómez's DFA model identified just two 'best predictors': person deixis, specifically a reduction in the use of *our*, and syntactic complexity, specifically a 566% increase (between the 1964 and 1976 samples) in the use of the conjunction *so that*. Cantos-Gómez (2010: 185) concedes that such a precise set of predictors might not extrapolate to other subjects, and that the same linguistic markers might not apply to patients who are native to languages other than English.

The genitive ratio analysis (table 6.12) is based on speeches by Wilson in 1964, 1976 and 1986, transcribed and published by Hansard. Wilson's first parliamentary appearance as the newly-elected prime minister was on 3 November 1964, in a combative debate on the Queen's Speech (setting out the government's legislative plans). Hansard records 12 contributions to the debate, several responding to others' interventions, a total of 7,753 words.

**Table 6.12: Harold Wilson: Genitive Ratio Analysis**

**Prediction**: The mean CR will be significantly lower in the later speeches, indicating an increase in the use of concrete language

| **Parliamentary Speeches** | **Date** | **Word Length** | **Mean CR** |
|---|---|---|---|
| *House of Commons* (3 November) | 1964 | 6.8 | +0.627 |
| *House of Commons* (13 December) | 1976 | 7.2 | +0.166 |
| *House of Lords* (25 June) | 1986 | 7.2 | +1.068 |

Wilson's resignation in April 1976 was (contrary to convention) not marked by a speech to the House of Commons. He is absent from the parliamentary record until 13 December 1976, when he made seven contributions as a back-bencher to a debate on the Scotland and Wales Bill, a total of 2,597 words. Wilson left the Commons in 1983. His last parliamentary contribution (as Baron Wilson of Rievaulx) was a prepared speech to the House of Lords, in a debate on Marine Pilotage, on 25 June 1986.

The three transcripts were analysed by the Animyser program. The program's POS tagger identified 1,583 nouns in the 1964 transcript, 451 nouns in the 1976 transcript, and 108 nouns in the 1986 transcript. The apparently significant increase in the concreteness of the 1986 sample is not reflected in the statistical analysis. As measured by a one-way ANOVA, there were no significant differences between the three Wilson speeches ($F$ (2, 2139) = 2.03, $p$ = .132). At an alpha level of 0.5, the difference between the 1976 and 1986 samples only approached significance [$t$ (557) = 1.44, $p$ = .075]. See Table 6.13.

In addition to Harold Wilson, two other contemporary politicians are known to have succumbed to dementia: Ronald Reagan and Margaret Thatcher.

Ronald Reagan was formally diagnosed with AD, and famously made a moving declaration of his illness in an open letter to the American people. Margaret Thatcher was never formally diagnosed with AD (or at least no such diagnosis was ever made public), though as with Agatha Christie there is extensive anecdotal evidence. Examination of their private correspondence, early and late, might have provided further support for the GR method. Whilst both the Ronald Reagan Presidential Library and the Margaret Thatcher Archive were very helpful and could provide examples of early correspondence, in both cases the later personal correspondence is currently embargoed.

Berisha, Wang, LaCross and Liss (2015) have analysed Reagan's spontaneous responses to journalists in his press conferences, from 1981 to 1988. Their regression analysis tracked the diversity of Reagan's vocabulary, together with his usage of non-specific nouns and "conversational fillers" (e.g. *well*, *actually*, *um*, *ah*). Compared with a similar analysis of the press conferences of President George H.W. Bush, they found significant indicators of decline in Reagan's language. From the perspective of a concreteness rating, a cursory study of the press conference material supports the contention that political discourse is essentially abstract. Politicians typically anticipate questions and rehearse answers (Gottschalk, Uliana and Gilbert, 1988). Even if it is supposedly spontaneous, their language is constrained by the context.

## 6.14 Summary of Case Study Findings

With the exception of Harold Wilson, and with some reservations (see section 6.16), the results of these case studies offer qualified support to the deployment of the mean-CR analysis that might be applied as a diagnostic of Alzheimer's disease or as a metric of its progression.

In one confirmed case of AD (Iris Murdoch), there is a statistically significant increase in the use of concrete language in the final novel, compared to its two predecessors. The same statistically significant pattern is seen in the novels of Agatha Christie, but not in those of the 'control' authors of detective fiction (P.D. James and Ruth Rendell). Perhaps the most interesting finding comes from a one-way ANOVA comparison of the first two novels (i.e. pre-AD in the cases of Murdoch and Christie) of all four novelists. The ANOVA detected no significant difference [$F$ (7, 3192) = 1.19, $p$ = .304]. This suggests at least the possibility of a normative level of the GR.

Sociolinguists distinguish different "registers" of a language, i.e. the changes in language production that are influenced by both the purpose and the context of a discourse. The ANOVA finding suggests that, given a common register of testing and a sufficiently large test population, it might be possible to establish a norm of concreteness for that register, against which individuals' test results could be measured. In the current context that measure is simply 'fiction'. Any suggestion that crime fiction and literary fiction are different, with crime fiction perhaps favouring more concrete language, is not supported by the statistical analysis. In another context, the register might be a set of standard questions posed within a structured interview.

The failure to detect any significant change across the three speeches of Harold Wilson perhaps demonstrates a limitation of the GR analysis. The register of parliamentary language is both formal and formulaic, with the added difficulty of distinguishing prepared from spontaneous speech in the context of a debate.

Political discourse is characterised by a relatively high incidence of abstract compound nouns (e.g. *tax cut, spending increase, executive order*). It is also noteworthy that mean word length, which is not significant either within or between the novelists, is significantly greater in the political discourse of Harold Wilson.

A 400 noun sample of each novel might seem modest, but is actually larger than the samples analysed by Garrard *et al* (2005). Only one novel (Christie's *The Mysterious Affair at Styles*) was readily accessible in digital form for analysis, hence the sampling method employed here.

The genitive ratio analysis, in the form of the concreteness rating, follows the predicted course of linguistic change in each of the case studies. It does so with a method that is simpler and more economical than the alternative methods of computational linguistic analysis reviewed in this chapter.

## 6.15 P-Density: CPIDR vs. CR

As previously discussed, propositional idea density (P-density) is the number of propositions in a text divided by the total number of words. As a single measure, P-density is arguably the principal competition that the genitive ratio has to match. The CPIDR program (version 3.2) relies on 37 proposition-counting rules to automate this measure (Brown *et al*, 2008: 542). See Chand *et al* (2012) for a critique of CPIDR and its limitations.

Jarrold *et al*'s (2010) inclusion of idea density was prompted by the Nun Study finding of low idea density as an early predictor of AD. Their finding, that idea density was significantly lower in pre-AD participants than in the controls, supports the Nun Study finding, but also generalises it: to spoken as well as written language, to men as well as women, and to later as well as early life. Ahmed *et al* (2013: 3734) found a significant difference in the measure of idea density across three clinical stages: MCI, mild AD, and moderate AD (*N* = 9, *p* = .02).

What might be termed the 'Nun Study hypothesis', that early-life language might be a predictor of later-life AD, was tested on a different cohort by Engelman, Agree, Meoni and Klag (2010). Membership of the longitudinal Precursors Study consists of entrants to the Johns Hopkins School of Medicine between 1948 and 1964, with their responses to detailed annual questionnaires and verification of their subsequent causes of death. Although drawn from a demographic (91% male, predominantly white and middle-class) that is very different from the nuns', the Precursors Study participants also curtail the confounding variables. They have generally followed the same career path, and they share a similar level of intellectual and academic attainment. Remarkably, samples of their writing (at an average age of 22, the same as in the Nun Study) were available for linguistic analysis, in the form of the personal statements written when seeking admission to the elite medical school.

Engelman *et al* (2010) examined the personal statements of 18 participants with verified clinical diagnoses of AD, each matched on age and gender with two "non-cognitively impaired" controls: a total sample of 54 participants. They analysed the last 10 sentences of each statement (or the whole statement if less than 10 sentences), using the CPIDR (version 3) software. Measured by propositions per 10 words, they found a statistically significant difference between the AD cases and the controls (4.70 vs. 4.99, *p* = .01).

**Table 6.13**: Comparison of CPIDR and mean CR ratings

<table>
<tr><td colspan="2"><b>TEXT</b><br>My extra-curricular activities include a social fraternity, Phi Gamma Delta; the Biology Club which I helped organize last year and in which I serve at present as vice president; and the Chemistry Club.  I have participated in intramural sports, basketball, touch football, softball and tennis.</td></tr>
<tr><td><b>CPIDR</b><br>20 propositions/ 45 words<br>(20/45)*10 = 4.44 density</td><td><b>CR</b><br>+1.143</td></tr>
<tr><td colspan="2"><b>TEXT</b><br>My deepest interests are, however, satisfied by my college courses.  Chemistry is intriguing as well as the courses included under zoology.  I find foreign languages engrossing and I hope to continue my study of Spanish and German in the future, with the addition of perhaps French and Latin and other languages.  Although my practice time is limited I enjoy music from the standpoint of interpretation as well as appreciation.</td></tr>
<tr><td><b>CPIDR</b><br>41 propositions/ 69 words<br>(41/69)*10= 5.94 density</td><td><b>CR</b><br>-4.203</td></tr>
</table>

In an appendix to their paper, Engelman *et al* (2010) provide two brief examples of CPIDR coding, applied to extracts from personal statements. These extracts and their CPIDR density scores are reproduced in Table 6.13, together with the comparative mean CR ratings.

The CPIDR analysis identifies the author of the first text as more likely to develop AD than is the author of the second text. The genitive ratio analysis makes the same distinction, with a marginally significant difference between the sets of nouns extracted by the Animyser program from the two texts [$t$ (32) = 1.70, $p$ = .049]. Cohen's effect size value ($d$ = 0.66) suggests that this finding has moderate (0.5) to large (0.8) significance (Cohen, 1988).

Requests by this researcher to gain access to the Precursors Study data have been unsuccessful. A more informed comparison of methods has therefore not been possible.

### 6.16 Caveats and Conclusions

The results of two separate studies in the USA and Australia (presented at the 2014 Alzheimer's Association International Conference, but not at that point published in a peer-reviewed journal) suggest that a "simple" eye-test might identify biomarkers for Alzheimer's disease. Both studies claim an accuracy of up to 85%. Hye *et al* (2014) have reported the development of a blood test, based on analysis of ten proteins, that they claim could predict the onset of AD in patients who already have a diagnosis of mild cognitive impairment. Their test has a reported accuracy of 87%. That seems impressive, but in 13 out of every 100 cases, it will give either a false positive diagnosis or a false negative, both of which have serious consequences for patients.

A minimal conclusion from the studies presented in this chapter, albeit one based on limited datasets, is that computational linguistic analysis is a potentially useful diagnostic tool, but there is as yet no consensus on a model of linguistic diagnosis of AD that would justify the cost of large-scale testing, whilst meeting the need highlighted by Ahmed *et al* (2013: 3728), for linguistic markers of AD to be "simple and specific".

A linguistic diagnostic test is no more definitive than any of the potential biomarkers. There is of course a possibility that a genetic biomarker with 100 % accuracy will be identified and implemented, but on present evidence that seems to be a distant prospect. It is, though, even less likely that clinicians would accept a linguistic diagnostic measure as an independent predictor of AD, even one with a level of accuracy that claims to match a pathological test. Acceptance by the medical establishment of either test, genetic or linguistic, would of course depend on the findings of the original research being replicated across much larger test populations.

Statistical power is constrained by small sample size (Ioannidis, 2005). A major barrier to replication is the difficulty of gaining access to sufficient clinical data, hence the very limited sample sizes of the machine learning studies discussed in 6.7. The Nun Study (Snowdon *et al*, 1996) and the Western Collaborative Group Study (WCGS; Jarrold *et al*, 2010)  are both remarkably valuable data sources. In the UK, OPTIMA (the Oxford Project to Investigate Memory and Ageing) has since 1988 been following a cohort of over 1,100 people, collecting data that included brain scans and samples of blood and cerebrospinal fluid, as well as cognitive tests. All three were longitudinal studies, extending over decades. These data sources provide multi-faceted data – clinical, cognitive, environmental and textual – though the crucial independent variable is a clinically verified cause of death for every participant.

The ongoing collection and cooperative sharing of longitudinal data as an international scientific resource must surely be a priority for dementia research. Dementia might be described as a 'longitudinal disease'. A problem with all longitudinal studies is to identify at the outset the data that will be useful to researchers several decades thence, an almost impossible task. In the Nun Study, the Precursors Study and the WCGS, the opportunity to conduct a linguistic analysis was fortuitous rather than planned. The consequence in all three studies was an attrition of potential participants by the necessarily strict application of selection criteria. Because the WCGS was a study of cardiovascular disease, the number of participants with AD as their verified cause of death was limited, and this is reflected in the sample size available to Jarrold *et al* (2010). My efforts, both directly and through third parties, to gain access to data from the Nun Study, the WCGS and the Precursors Study, were not rejected, they were simply ignored.

The difficulty of obtaining such 'gold standard' data has led researchers to other sources, as reflected in the cited studies of novelists that prompted my own comparative analysis. It is significant that, of those novelists, only Iris Murdoch had a specific clinical diagnosis of AD that was verified post-mortem. In the absence of any clinical data, an AD diagnosis for Agatha Christie can be no more than 'probable', and in fact has been questioned (by Van Velzen, Nanetti and de Deyn, 2014). Even with levels of public awareness as high as they are now, the Alzheimer's Society has estimated that by 2021 there will be half a million cases of undiagnosed dementia in the UK. It is therefore not surprising that there is a dearth of verified AD cases from the past, when the stigma of mental decline was so much greater.

Even if there were a corpus of texts from writers with verified dementia, the data would have to be regarded with great caution. There is, for example, insufficient evidence of how a writer's genre might affect the linguistic analysis.

The genre of fiction for very young children is not well-suited to linguistic analysis, because the language is necessarily simple and concrete. The prolific author of over 700 children's books, Enid Blyton (1897-1968) was reportedly diagnosed with 'pre-senile dementia' (Lancashire, 2014) at the end of her life. Ellis (1996: 491) has concluded, from an analysis of patients' transcripts, that the later stages of AD are characterised by a "lexically driven mode of processing that shares features with early evolutionary stages of language, pidgin speakers, and early developmental stages in children". In other words, the language of the Alzheimer's patient will regress to a state that is "pre-grammatical", to a "more rudimentary form of information staging" (*ibid*), to the vocabulary of the child they once were. The total vocabulary deployed by Enid Blyton in 21 *Famous Five* books, written between 1942 and 1963, was only 11,500 words (Lancashire, 2014).

Linguistic analysis of works of fiction carries another important caveat: writers develop over time, and in different ways. In the Afterword to his *Collected Stories* (2007: 439), the American novelist and Nobel Prize winner Saul Bellow wrote that "It's difficult for me now to read [my] early novels, not because they lack interest but because I find myself editing them, slimming down my sentences". As writers mature, become more self-confident in their craft, they are quite likely to write in simpler, shorter sentences, to be less concerned with demonstrating their own cleverness. Measures of lexical diversity and syntactic complexity feature prominently in the studies cited in this chapter. The decline of these faculties, as a consequence of normal ageing, has been acknowledged, but the conscious simplification of language is a potentially confounding factor that applies particularly to texts written for publication.

Neither works of fiction nor parliamentary speeches are ideal material for analysis. In novels, there is the question of how far the language of the characters reflects the natural language of the author. If indeed "natural language" is an appropriate description, given the extensive re-drafts and revisions made by authors and by their editors. On balance, a text with dialogue excluded is probably more representative of the author's own 'voice'. Provided the text is in a digital format, as is increasingly the case, it should be relatively simple to write a program that identifies and excludes from parsing any text that is enclosed within quotation marks.

Even supposedly spontaneous parliamentary contributions are based on careful preparation, by the speaker and (for a prime minister) a team of aides and advisers. Moreover, the responses are framed in the formality of parliamentary language, with elaborate and archaic forms of address. The pragmatics of parliament are not the pragmatics of everyday discourse. The language deployed is also influenced by the subject matter. The language of a parliamentary debate about democracy will tend to be more abstract than if the debate were about (say) road repairs or cheese manufacture.

There are three other, very obvious limitations of the research presented and reviewed in this chapter. The first is that it is limited to texts written in English. Only the evidence of linguistic analyses in other languages will determine if similar results might be achieved. An advantage of the GR method is that it can be applied to other languages, via English as a proxy (see Chapter 5).

The second limitation is that all of the studies are 'within subject', comparing the language of the same participants, both patients and controls, at different points in time. Should a patient present for 'linguistic diagnosis', it would therefore be necessary to have access to comparable samples of that patient's language from at least several years previously. The alternative would be to establish norms of linguistic ability, perhaps typical of different age-groups and levels of education, that would provide a baseline for an initial diagnosis.

The third limitation is the requirement, to varying degrees, for a pre-processing 'pipeline' that delivers tagged and filtered text to the central process. The GR method is amenable to full automation – although so are other linguistic measures discussed above. The development of systems such as CLAS and CPIDR is important because many prior studies have relied upon a level of manual pre-processing that would not support the operational use of computational linguistic analysis as a cost-effective diagnostic tool.

The analysis in this chapter highlights both the potential benefits and the evident limitations of genitive ratio analysis. If it were to be applied in a clinical context, as one element in the diagnosis or monitoring of AD patients, then the 'register' of that application becomes significant – personal, spontaneous, informal and honest. An online diary or blog might meet those criteria and facilitate computational analysis.

Perhaps computational linguistic analysis is best regarded as one component of a clinical risk assessment model, to be given a relative weighting alongside other factors such as family history, existing medical conditions that predispose to AD (e.g. diabetes), cognitive tests, and (in future) tests for genetic biomarkers. All these are low-cost and minimally-invasive data sources. Such a risk assessment model might have significant benefits.

CHAPTER 7

# The Abstract Language of Depression

*Every man has his secret sorrows which the world knows not; and often times we call a man cold when he is only sad.*
Henry Wadsworth Longfellow

## 7.0 Overview

Alzheimer's and depression are differentiated by more than just the appellation of 'disease' to the former. Alzheimer's disease (AD) is caused by physiological changes to parts of the brain, with the consequence of a cognitive decline that is irreversible. Depression has been characterised as a cognitive bias, and its treatment as cognitive bias modification (see section 7.7).

That cognitive bias is manifest in language production. Just as the language of AD regresses to the concrete, so the language of depression is biased to the abstract. Just as with AD in the previous chapter, it should therefore be possible to use genitive ratio analysis to track a progression, but in this case from concrete to abstract and (with effective treatment) back to concrete.

A substantial body of psychological research supports the association of depression with abstract thinking and language, and specifically with "reduced concreteness thinking" (Watkins, Moberly and Moulds, 2008). Based on their empirical findings, Watkins and colleagues have developed a therapy of "concreteness training" that addresses the "abstract-overgeneral cognitive bias" that is typical of depression.

The LIWC (Linguistic Inquiry and Word Count) software developed by Pennebaker and colleagues has led the way in the application of computational linguistic analysis to psychological disorders. More recent applications have relied on machine learning classifiers. A common factor in these studies is the difficulty of gaining access to current clinical data. A comparison of a concreteness rating (CR) analysis with LIWC suggests that the CR benefits from a simpler and more objective process.

A more recent development of computational linguistic analysis is sentiment analysis, which has been deployed mainly in commercial applications to 'mine' opinions and attitudes from social media and other texts (see section 7.15). Whilst the two methods are very different, the cognitive dimension of genitive ratio analysis might complement the linguistic dimensions of sentiment analysis.

The previous chapter used a case-study paradigm that facilitated comparisons both within subjects (books written by the same author) and between subjects (books written by different authors). This chapter will again utilise case studies for within subjects comparisons, but in addition a between subjects analysis will compare a set of negative ("my depression") blog postings with a set of positive ("life is good") postings.

The results of the analyses in this chapter will represent a proof of concept, a *prima facie* case for the genitive ratio as a metric of textual analysis that might feasibly be applied to monitoring an individual's depressive state.

## 7.1 The Diagnosis of Depression

"It would be nice if we had a biological gold standard, but that doesn't exist, because we don't understand the neurobiology of depression."
(Robert Spitzer, Chair of DSM-3, quoted in Carlat, 2010: 54).

DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, 5th edition, American Psychiatric Association, 2013: 160-161) lists nine symptoms that are indicative of a "major depressive disorder" (MDD). There has been no significant change in the symptomatology of depression since DSM-3 (1989). Diagnosis still relies on the identification by a clinician of five or more of those nine symptoms occurring during a period of two weeks, and they must include at least one of either "depressed mood" or "loss of interest or pleasure". As Davies (2014) has demonstrated, the specification of five symptoms in two weeks is arbitrary, not supported by any empirical evidence. It is probable that the current diagnostic criteria would not survive the discovery of a reliable biological test.

Progress in the development of a possible gold standard, in the form of "a panel of blood biomarkers for early-onset MDD", has been reported by Pajer, Andrus, Gardner, Lourie, Strange, *et al* (2012: 2). Early-onset MDD has been linked to the later development of both Alzheimer's disease and Parkinson's disease. Based on initial experiments with rats, Pajer *et al* identified and tested 26 "candidate blood transcriptomic markers" on two (*N*=14) groups of 15-19 year-old participants matched on age, gender and race: one group with a diagnosis of MDD and a control group. The two groups were found to be differentiated by eleven separate biomarkers. Whilst this research is a positive step towards that "biological gold standard", the authors acknowledge that much larger-scale trials will be necessary before there is any prospect of their test entering clinical practice.

## 7.2 A Mirror of the Mind?

"Languages are the best mirror of the human mind", wrote Gottfried Wilhelm Leibniz (1981: 330) in 1765. Recognition of formal linguistic analysis as a potential diagnostic of psychological disorders is a rather more recent phenomenon: "It may prove valuable for the clinician to analyze the speech content of a particular patient when diagnosis is difficult or ambiguous. This would permit identification of key words to listen for in the patient's discourse that might signal the presence of a certain disorder" (Oxman, Rosenberg, Schnurr and Tucker, 1988: 468).

A growing number of researchers are addressing the challenge of applying computational linguistic analysis and 'big data' to the field of cognitive neuroscience (see Garrard and Elvevåg, 2014, for an overview). Studies have, for example, applied latent semantic analysis (LSA) to the analysis of schizophrenic patients' discourse (e.g. Holshausen *et al*, 2014, and Tagamets *et al*, 2014).

### 7.3 Word-Count Tools: LIWC and Wmatrix

Word-count is a basic component of most tools for computational linguistic analysis. A computer program labelled Linguistic Inquiry and Word Count (LIWC, pronounced *Luke*) has been progressively developed over more than two decades by James Pennebaker and colleagues at the University of Texas in Austin. LIWC is based on a dictionary of 2,300 words, spanning 74 "grammatical and psychological dimensions" (Mehl and Gill, 2010: 113).

LIWC is currently (in 2014) the best-known and most widely-used psychological tool based on linguistic analysis, with increasing adaptation of the system for languages other than English. For example, Fornaciari and Poesio (2013) have applied computational stylometric analysis to the identification of false and deceptive testimony in a corpus of Italian court transcripts. Using the Italian-language version of the LIWC software (Alparone *et al*, 2004), Fornaciari and Poesio tested a number of models, all of which depended on surface features accessible to computational analysis. All performed above chance in identifying deceptive statements.

Pennebaker has applied his linguistic-analytical toolset in a wide range of contexts: from the identification of untruths, to the prediction of academic attainment, to the attribution of authorship, and many more. He contends that: "wherever there is a word trail – no matter how long – computer text analysis methods can help interpret the psychology of the authors" (Pennebaker, 2011: 282).

Using LIWC, Pennebaker and colleagues have accumulated a substantial body of research into the assessment of psychological states. Their principal finding has been that some of the most revealing lexical factors have been not the selection of words with high semantic content, but rather the everyday "function words" that are deployed almost subconsciously in our discourse: pronouns (particularly personal pronouns), determiners, prepositions, auxiliary verbs, conjunctions, negations and quantifiers (Pennebaker, 2011: 22). The relative usage of such words has been shown to correlate with the "age, sex, social class, personality, and social connections" of their authors (*ibid*: 269). As a specific example, article usage is correlated with the concrete/abstract distinction: "Participants … who use a high percentage of articles in their speech by definition are referring to concrete and impersonal objects or events" (Pennebaker and King, 1999: 1309).

With its reliance on pre-defined dictionaries and "dimensions", LIWC functions as a top-down system. By contrast, Wmatrix, developed by corpus linguists at the University of Lancaster, is bottom-up, or data-driven. Wmatrix (Rayson, 2008) is based on corpus comparison, and incorporates pre-processing of a text by part-of-speech (POS) and semantic taggers. Whilst the output of Wmatrix is more comprehensive (more fine-grained) than that of LIWC, because it is not constrained by the content of a custom dictionary, it also requires a greater degree of interpretation by the researcher, based primarily on differences of frequency between two corpora, to identify the relative under- or over-use of particular words or syntactic categories.

## 7.4 LIWC: Sex and Death and Suicidal Poets

Empirical studies have repeatedly found that people with depression view the world in a way that is not only negative but also self-focussed. The latter trait has been linked in a number of studies to a relatively high incidence of first person singular pronouns in their discourse (see Rude, Gortner and Pennebaker, 2004: 1121-1122 for references). Rude *et al* have replicated these findings in their own analysis that relies on the LIWC software. Their comparison of currently-depressed and never-depressed college students shows that the currently-depressed participants were significantly more likely to use the first person singular pronoun "I" (though not "me" or "my"). Their comparatively high usage of negative emotion words was also significant.

Text analysis using the LIWC program has been applied to a 'between subjects' comparison of the poetic language of poets who committed suicide, with that of other poets (matched by era and nationality) who did not. Stirman and Pennebaker (2001: 517) claim that their "findings suggest that linguistic predictors of suicide can be discerned through text analysis". Based on what they acknowledge to be a small sample (of nine suicides and nine controls), Stirman and Pennebaker found that the poetry of suicidal poets was distinguished by a greater use of first-person singular pronouns, which was statistically significant (*p* = .02) compared with the poetry of the control group.

However, their findings of a higher usage of "negative emotion and death words" (*ibid*: 520) only approached significance (*p* = .08), whilst post-hoc analysis revealed a significantly higher usage (*p* = .05) by the suicidal poets of "sexual words" throughout all phases of their careers. The authors simply note that "stronger evidence was found for a pre-occupation with sexual matters than with matters pertaining to death" (*ibid*: 520). It is difficult to rationalise this finding, unless it is due to chance (and sample size). It is also difficult to support the selection of a control group of poets matched simply by era and nationality.

Stirman and Pennebaker argue that their findings support Durkheim's (1952) theory of suicide as social disengagement. This is characterised in their analysis by an unusually high degree of self-reference; by fewer (though not statistically significantly fewer) "communication words" such as *talk* and *listen*; and by more "death words" (though even more "sexual words"). Their argument is not convincing, and it is difficult to see how their analysis contributes to a greater understanding of suicide, or to the anticipation (and thereby treatment) of potential suicides.

A comparison of LIWC and genitive ratio analysis will be presented later in this chapter (7.14).

### 7.5 Machine Learning: Three Studies

Three very different applications of computational linguistic analysis are here reviewed. The first (Pestian *et al*, 2008) provides a methodological comparison with the LIWC/suicidal poets study (Stirman and Pennebaker, 2001), and delivers a more convincing analysis of 'suicidal language', based on a more valid and reliable dataset. The second study (Jarrold *et al*, 2010) challenges the LIWC findings (by Rude, Gortner and Pennebaker, 2004) of self-focussed function words (particularly first-person pronouns) as indicators of depression. The third study (Neuman *et al*, 2012), based on "metaphorical analysis", supports the LIWC findings in one respect: the addition of a first-person pronoun count does improve the accuracy of their process. These three studies together illustrate the diversity of methods and variables in current models of computational linguistic diagnosis, and the potential of machine learning classifiers.

#### Pestian, Matykiewicz, Grupp-Phelan, Lavanier, Combs and Kowatch (2008)

Machine learning algorithms have been applied to the classification of suicide notes by Pestian *et al* (2008). Their sample consisted of 33 genuine suicide notes from "completers" (people who had actually committed suicide), matched with 33 contrived suicide notes written by "simulators", who were matched with the completers on age, gender and socio-economic category.

The researchers constructed an ontology of "emotion words" associated with suicide, for example:

| | | | | |
|---|---|---|---|---|
| *Class*: | Affection | Anger | Depression | Worthlessness |
| *Concepts*: | love | | | |
| | concern for others | | | |
| | gratitude | | | |

A feature space that incorporated specific words (e.g. *love*, *life*, *no*), parts of speech, emotional concepts, and a readability index was tested on a range of machine learning tools. The algorithms' classification of the genuine and contrived suicide notes was then compared with the judgments of mental health professionals, whose accuracy was 71%. The comparable accuracy of the machine learning algorithms ranged from 60% to 79%, with the SMO (sequential minimal optimization) algorithm achieving the best result. An intriguing finding was that the mental health professionals had based their judgments primarily on the emotion words in the notes, whereas the algorithm's performance was actually improved by the exclusion of that vector.

### Jarrold, Peintner, Yeh, Krasnow, Javitz and Swan (2010)

Jarrold *et al* (2010) have drawn on data collected for the Western Collaborative Group Study (WCGS) of cardiovascular disease, to test diagnostic models of both depression and dementia. They obtained speech samples by transcribing audio recordings made in 1988 of semi-structured interviews with WCGS participants. The same participants had also completed the Center for Epidemiologic Studies – Depression Scale (CES-D). The CES-D (Radloff, 1977) is a long-established tool that is widely used as an initial diagnostic of possible depression. Jarrold *et al*

(2010: 304) identified two distinct participant groups: depressed (CES-D score >25) and non-depressed (CES-D score <20).

Jarrold *et al* (2010) addressed three research questions:

1. Would their findings replicate the relatively high frequency of "self-focussed words" (particularly first-person pronouns) that had been found in LIWC studies by Pennebaker and colleagues (e.g. Rude, Gortner and Pennebaker, 2004)?
2. How accurately might cases of depression (those with a CES-D score of >25) be diagnosed by a machine learning model based on lexical features derived from the same participants' speech transcripts?
3. How might the accuracy of a diagnosis be affected by the context (the semi-structured interview) in which the speech-data were obtained?

The analysis by Jarrold *et al* (2010: 304) of the full interview transcripts did not support the LIWC analysis. They found "no association between depression and self-focussed language. Secondly, the diagnostic accuracy of models based on the lexical feature set (including self-focussed word frequency) was only slightly better than chance." However, analysis of the participants' responses to one specific question did show a higher frequency of self-focussed , first-person words in the depressed participants group. This was that question, from the interview:

> Question 24-b: In your work or career, have you accomplished most of the things that you wanted to accomplish? (If No) Why not? What's gotten in the way? Are you doing anything about this? (*ibid*: 304)

Significant results were obtained from the analysis of that single question, with a "nearly complete separation" between the depressed and non-depressed groups in terms of the percentage of first-person words in their discourse: 12.9% (sd 4.2) for the depressed group versus 6.3% (sd 2.03) for the non-depressed group, albeit on a "small sample size" (*ibid*:305).

The primary conclusion from these findings is that the context of the source of speech data is potentially significant. Question 24-b focuses on the respondent with repeated use of the second-person pronoun, demanding the exercise of autobiographical memory, introspection, self-analysis, and forward planning. These cognitive functions will be implicated in the discussion of "reduced concreteness thinking", later in this chapter.

### Neuman, Cohen, Assaf and Kedma (2012)

Neuman *et al* (2012) have developed a system that uses metaphorical analysis to screen texts for evidence of depression. They envisage that this could be operationalised, by participants allowing access to their online postings in social media or blogs. A positive indication of depression from this initial screening would prompt the participant to then complete a diagnostic questionnaire (also online). If the results of the questionnaire were to support the linguistic analysis, then the participant would be urged to seek a clinical diagnosis.

The deployment of metaphorical analysis by Neuman *et al* (2012) involved the construction of a lexicon that contains the metaphors that sufferers from depression apply to the description of their condition, their feelings, their emotions. The 'Pedesis' system developed by Neuman *et al* used Microsoft's Bing search engine to mine the web, using a combination of phrase-search and wild-card (*) to locate "depression is like *" expressions. The macro-economic sense of *depression* was excluded by searching for any "econo*" string on either side of the target phrase.

The Pedesis program automatically extracted the metaphors from their context, together with any elaboration. Manual analysis (of some 20,000 metaphors) then identified the words and phrases that typically depict depression, e.g. *scared*, *lonely*. First and second order synonyms (i.e. synonyms, and synonyms of synonyms) of those words were then extracted from the Corpus of Contemporary American English (Davies, 2009). The resulting lexicon contains 1,723 depression-related phrases, of which the most frequent are (Neuman *et al*, 2012: 22):

[7.1]

dark
black hole
life
disease
cancer
death
pain
box
black
quicksand
emotional
cloud

Neuman *et al* (2012) tested their Pedesis system on two corpora: questions posted to a self-help website for depressives (mentalhelp.net), and weblog entries extracted from the Blog Authorship Corpus (Schler *et al*, 2006). The blog analysis will illustrate their method.

The Blog Authorship Corpus contains 681,288 postings, harvested from 19,320 bloggers on www.blogger.com in August 2004. Neuman *et al* selected 83 postings, the authors of which had self-identified as depressed (D), and 100 postings by authors with no evidence of depression, as controls (¬D). From each posting, Pedesis derived a 'DepScore', i.e. a count of the phrases contained in the depression lexicon. With D or ¬D as the dependent variable and DepScore as the independent variable, a binary logistic regression found that the system correctly classified 84.2% of the postings (*p* < .001). An additional dependent variable was added: the occurrences of the first person pronoun "I" as a percentage of the word-count (cf. Rude *et al*, 2004). The system then correctly classified 90.7% of the postings (*p* < .001).

### 7.6 Key Findings from the Three Studies

The different approaches exemplified by these three studies illustrate the range and diversity of computational linguistic analysis. They also offer three specific pointers to the CR analysis that will be presented in this chapter.

First, the human experts' intuitive assumption that a specific vocabulary (of 'emotion words' in Pestian *et al*, 2008) is symptomatic of a suicidal mindset is not supported empirically. This suggests that an effective linguistic analysis needs to extend beyond the constraints of a pre-defined dictionary.

Second, the context or 'register' of the language sampled is potentially material to the analysis: Jarrold *et al* (2010) obtained significant results from one specific question that focused participants on a particular range of cognitive functions.

Third, the study by Neuman *et al* (2012) provides a contextual precedent for utilising a corpus of blog postings as a data source that is in the public domain (see Hunt, 2013: 92-97 for a full discussion of the ethical issues, which remain rather ill-defined).

## 7.7 Overgeneralization and Depressive Rumination

"One cognitive bias strongly implicated in the onset and maintenance of depression is the tendency to process self-relevant information in an overgeneralized and abstract manner".
(Watkins, Baeyens and Read, 2009: 55)

There is converging evidence that people with depression struggle to retrieve specific autobiographical memories, and are more likely to recall "overgeneral" memories (see Watkins, Baeyens and Read, 2009: 55 for references). These overgeneral memories are categorical. Instead of recalling specific events, depressives group events together within a category, for example of mistakes they have made, or of occasions when they have failed. They therefore interpret a single failure as a sign of total inadequacy.

This "overgeneralization" is a distinguishing characteristic of depression. Overgeneral memories appear to be specific to cases of depression, and are not found in the symptomatology of apparently similar mental health conditions, such as post-traumatic stress disorder (see Watkins, Teasdale and Williams, 2000: 911 for references). It has also been observed that a high propensity for overgeneral memory in depressed patients correlates with both an impairment of their ability to deal with interpersonal problems, and a prolonged prognosis for their condition (*ibid*: 912). Such an "abstract-overgeneral cognitive bias" in depression is "characterized by the abstract construal of self-relevant (particularly negative) actions and events" (Watkins, Baeyens and Read, 2009: 56), and can become self-perpetuating.

Abstract thinking and language are hallmarks of "depressive rumination", a condition that has been defined by Nolen-Hoeksema (1991: 569) in terms of "behaviour and thoughts that focus one's attention on one's depressive symptoms and on the implications of those symptoms". Characteristics of depressive rumination are an inward focus of constant, negative self-evaluation, repeatedly analysing "causes, meanings, consequences, and implications of symptoms of depression, negative social comparisons, and 'Why?' type questions" (Watkins, Moberly and Moulds, 2008: 364). It is also characterised by "reduced concreteness of thinking" (*ibid*: 365).

There is some degree of commonality between rumination, worry, and depression. They are all characterized by habitual negative thinking and by reduced concreteness thinking that limits the scope of an emotional response. The distinction between worry and rumination is that worry fears the future, whereas rumination dwells morbidly on the negatives of the past. Rumination is "like the cow chewing the cud … repetition of a theme in thoughts, without progression toward choice of a solution and a commitment to that solution" (Nolen-Hoeksema, 1996: 136-137).

Watkins and Moulds (2007: 1387) draw a parallel between the elevation of abstract thinking and "overgeneral memory recall" in depressed patients, with experimental evidence that rumination reinforces overgeneral memory (e.g. Watkins and Teasdale, 2001). Crucially, Watkins and colleagues argue that it is possible to counter this abstract-overgeneral bias with a process of "concreteness training".

### 7.8 Theories of Worry and Reduced Concreteness

The avoidance theory of worry (Borkovec, Ray and Stöber, 1998) posits that we worry in order to "escape aversive imagery" (Stöber, 1998: 753). The main component of worry is thought, so the experience of worry is mainly verbal rather than imaged, abstract rather than concrete. Worry is a "flight to abstractness" (Borkovec, Ray and Stöber, 1998: 573). Our physiological response to a difficult or frightening problem is mitigated by conceptualising the problem in abstract terms, and so suppressing thoughts that are more vivid and more concrete, more 'real'. Abstract concepts are less immediate, easier to dissociate from.

Generalized Anxiety Disorder (GAD) is a diagnostic condition, the central component of which is chronic worry (*Diagnostic and Statistical Manual of Mental Disorders* (4th edition), abbreviated *DSM-4*, American Psychiatric Association, 1994). Based upon a sample of actual patients with a clinical diagnosis of GAD, the "reduced concreteness theory of worry" (Stöber and Borkovec, 2002) hypothesises that a reduction in concrete thinking accompanies a reduction in mental imagery, so that an individual's thought processes become less distinct and more generalised. Such images as are formulated are likely to be negative images. Problem-solving – in the sense of being able to visualise a way forward, through one's immediate problems – is impeded by the difficulty of addressing specifics. The result is that the level of anxiety is maintained or worsens.

Watkins and Moulds (2005) place rumination at the core of depression. They argue that the reduced concreteness theory of Stöber and Borkovec (2002) should apply to rumination as well as to the Generalized Anxiety Disorder (or simply 'worry') that was its original application, since a symptom of both disorders is "recurrent self-related negative thinking" (Watkins and Moulds, 2005: 320).

Proponents of the reduced concreteness theory of worry regard it as central to the maintenance of "depressive rumination". Watkins and Moulds (2007: 1387) propose a two-part explanation. First, reduced concreteness thinking limits the patient's ability to envisage a clear plan of action for resolving a problem event. Second, there is empirical evidence (*ibid* for review) that reduced concreteness thinking constrains the "physiological and emotional responses" that would help the patient to deal with difficult situations.

Reduced concreteness thinking (RCT) is characterised by thought processes and autobiographical memories that are abnormally abstract (it is self-evident that reduced concrete thinking and increased abstract thinking are two sides of the same coin). Suppose that an individual is prompted to talk about his or her participation in recent social gatherings. The concrete account might be characterised by a clear recollection of a specific event – where it was, who was there, what they did. Whereas the abstract account might focus negatively on the individual's generalised experience of such events, perhaps on their self-perceived inability to make social connections.

Advocates of RCT hypothesise that other issues flow from a dominance of abstract thinking: that reduced imagery impedes both "emotional processing" and problem-solving (Watkins and Moulds, 2005: 320). In two experiments, based on problem elaboration charts and "catastrophizing interviews", Stöber, Tepperwien and Staak (2000: 224) found that "the more worrisome the topic was, the less concrete were the participants' problem elaborations". Watkins and Moulds (2007:

1392) tested the concreteness of problem descriptions in three groups of participants: currently depressed, recovered depressed, and never depressed. They found no significant difference between the latter two groups, suggesting that reduced concreteness thinking is transiently symptomatic of the state of depression, rather than a dominant trait of people who are prone to depression.

### 7.9 Therapy: Concreteness Training

The potential value of these research findings resides in the possibility of a new type of therapy for cases of mild to moderate depression: "concreteness training" (CNT), a subtype of cognitive behavioural therapy (CBT).

The challenge for the CNT therapist is to restore a capacity for more concrete thinking that will enable patients to address their emotional and social problems. Watkins, Moberly and Moulds (2008) presented empirical (but not clinical) evidence that it is possible to induce a more concrete mindset that is a counter to depressive rumination. However, those findings were derived from a student population of participants with "depressive symptoms well below clinical levels" (Watkins, Moberly and Moulds, 2008: 377).

Watkins and colleagues have subsequently developed a course of treatment for depression, based on "cognitive bias modification" (CBM). They report positive results from clinical trials (Watkins *et al*, 2012). The treatment consists of a self-guided course of concreteness training. CNT seeks to influence two of the cognitive processes that play a role in maintaining depression: abstract rumination and over-generalisation. Through a process of repeated practice, CNT aims to overcome abstract thinking and to reinforce concrete thinking.

CNT directs patients to review a recent emotional or social problem, by concentrating on concrete images: specific details of the problem, how it evolved, how it might be addressed by advancing specific actions and behaviours. CNT (see Watkins, Baeyens and Read, 2009: 57 for a detailed account) involves encouraging and enabling participants to focus on:

- Giving sensory descriptions of events (sight, hearing, touch)
- Recalling specific contextual details
- Narrating the sequence of events ("imagine a movie")
- Setting out a way forward, with each step in the process defined

The training applies these approaches to six standard scenarios (three positive and three negative) and to three autobiographical scenarios that are provided by the participant. Following a session of guided practice, participants are given an instruction booklet, a diary to record their progress, and access to a website for approximately 30 minutes of computer-based exercises per day.

On measured reductions of rumination, self-criticism and depressive symptoms generally, Watkins, Baeyens and Read (2009) found significant improvements in a CNT-trained group, relative to a group that received a "bogus" course of training and a "waiting list" group of untreated controls. This is an encouraging result, though not without reservations. The duration of the training (only one week, perhaps due to the ethics of bogus and untreated groups) was inadequate to determine the longer-term benefits, or the drop-out rate that might occur in a longer course. There is also always the possibility that the commitment and enthusiasm of the researchers might not be replicated by the operational implementation of web-based CNT, leading to higher than anticipated drop-out rates.

Clinical endorsement of CNT as an option for the treatment of depression must of course be subject to further testing and replication of the positive results from the early trials, which reported benefits to be still in evidence at six months after the completion of the course. For the treatment of mild and moderate depression, CNT offers potential benefits of a scalable, readily accessible and relatively low-cost intervention that might replace or complement prescribed medication, and that could be delivered to patients online or as a smartphone app.

## 7.10 The Genitive Ratio as a Possible Metric of Depression

The preceding review has established that abstract vs. concrete thinking and language are key elements of mainstream theories of depression. The development of concreteness training, based on research funded in the UK by the Medical Research Council, offers to over-stretched providers of mental health care (such as the UK's National Health Service) the prospect of an effective self-help therapy for cases of mild and moderate (or even self-diagnosed) depression, that could be rolled out on a large scale without the need for engaged support from hard-pressed clinicians.

The process of concreteness training (CNT) acquires structured language samples from participants at regular intervals. It is feasible that genitive ratio analysis might provide a metric of participants' progress, alerting clinicians when positive results are not being achieved. A monitoring system based on GR analysis would not impinge on participants' experience of CNT. It would run in the background, at low cost and with low maintenance.

The derivation of a reliable norm, a population benchmark of concreteness, is not a goal of the current research. There are significant confounding variables, for example age and level of education. A controlled environment and context would be crucial, for example responses to standard questions within a structured interview, as suggested in chapter 6.14. The current proposal limits the potential application of the concreteness rating to a method of tracking through language an individual's cognitive progress, as a prognostic measure. Diagnosis of depression based on the GR would require representative samples of a subject's language prior to the onset of depression, and whilst not impossible that presents obvious practical difficulties.

### 7.11 A Proof of Concept

The hypothesis derived from the preceding review of psychological research is that the direction of language change, with the successful treatment of depression, should be from an abstract bias to a more concrete bias. Two analytical data sets will be presented. Whilst situated within a paradigm of depressive symptoms, these data have no clinical validity. In the absence of any access to a clinical population, the results of the analyses must simply represent a proof of concept.

The first analysis is 'between subjects', based on two sets of language samples derived from a corpus of blogs. The second analysis is 'within subjects', based on the published (but very personal) writings of three authors who have committed or attempted suicide. Both analyses rely on the Animyser computer program (see chapter 4). Since both sets of data have been derived from texts in the public domain, the principal ethical concerns are of fair representation of, and due respect for, the originators. Both of these concerns have been observed.

### 7.12 Between Subjects: 'My depression' vs. 'Life is good'

"What we see in a blog … is a personal stream of consciousness with no intervention from editors or proofreaders. Language specialists have not made it consistent, articulate or polished, and so such language represents a kind of natural, idiosyncratic public writing not seen in English since the Middle Ages." (David Crystal: *Evolving English*, 2010)

Sampling of web-based language resources has great potential as a source of data, although with some limitations (Lyons, Mehl and Pennebaker, 2006: 256). Contextual information about participants is restricted, so that the researcher is often reliant upon inference and unverifiable self-report. In the specific context of depressive disorders, there might be a significant difference between an individual's self-diagnosis and a clinical diagnosis based on DSM criteria (*Diagnostic and Statistical Manual of Mental Disorders*, 5th Edition: DSM-5, 2013). Although these factors must be taken into account in the design of an analysis, they are arguably outweighed by the advantages of a huge and diverse dataset in digital form and with public domain accessibility.

The Birmingham Blog Corpus (Kehoe and Gee, 2012) contains UK and US texts, postings extracted from popular blogging sites such as WordPress.com. Since the corpus is one component of WebCorp, the WebCorp Linguist's Search Engine enables the user to extract a concordance of items containing a specific phrase. Each item can then be viewed in full context, either as a cached image of the original, or as plain text.

This analysis compares a set of blog postings from contributors who exhibit a negative (assumed to be depressed) mindset, with an equivalent set from contributors with a positive (assumed to be not depressed) mindset.

**'My depression' (MD).** After a series of trials, it was found that the phrase "my depression" was most likely to yield relevant postings that were personal to the blogger and likely to exclude the pervasive macro-economic sense of 'depression' (the postings were dated between 2008 and 2011). The search engine produced a concordance of 132 extracts containing that phrase. By reading through these items, those in any of the following categories were excluded:

- Comments on a posting, since these were usually very brief
- Duplicate postings
- Instances where 'my depression' referred to a past state, e.g. "I let my depression get the best of me for a few weeks"
- Adjectival usages of 'depression', e.g. "I am editing my depression book for Kindle"
- Instances where the content of the posting suggested that the use of "depression" might be an exaggeration or simply made for effect

Nine postings, with a total of 7,864 words, were selected from the remainder, as most likely to reflect genuine depression, albeit in most cases self-diagnosed. These were copied into a text file for analysis.

**'Life is good' (LIG).** Another series of trial searches attempted to identify a phrase typical of a non-depressed mindset. Variations on "I am very happy" were ineffective, since 'happy' is so often used in the sense of 'satisfied' or 'pleased', as in "I am very happy that he finally has a good home". It was the positive assertion "life is good" that yielded the most relevant results, with 410 extracts containing that phrase. A similar process of exclusion and selection produced a text file of 17 postings, with a total of 9,088 words.

Both files were analysed using the Animyser program's POS tagger. The mean CR scores are reported in Table 7.1. Statistical analysis provides strong support for the utility of the genitive ratio as a measure of the abstract cognitive bias that is symptomatic of depression [*t* (2879) = 7.57, *p* < .001].

**Table 7.1: Birmingham Blog Corpus: Genitive Ratio Analysis**

**Prediction**: The mean CR will be significantly higher for "life is good" than for "my depression", indicating a higher incidence of concrete language in the former.

| **Source** | **Blogs** | **Nouns** | **Mean CR** |
|---|---|---|---|
| "Life is good" | 17 | 1582 | +2.311 |
| "My depression" | 9 | 1299 | +1.116 |

However, this analysis compares two sets of aggregated texts. Table 7.2 shows the mean CR scores for the individual blogs within each set.

**Table 7.2:** Mean concreteness rating (CR) scores of individual blog postings

| | "Life is good" (LIG) | | "My depression" (MD) | |
|---|---|---|---|---|
| Rank | Blog # | Mean CR | Blog # | Mean CR |
| 1 | 9 | 3.147 | | |
| 2 | 3 | 2.974 | | |
| 3 | 6 | 2.943 | | |
| 4 | 7 | 2.840 | | |
| 5 | 10 | 2.668 | | |
| 6 | 8 | 2.596 | | |
| 7 | 15 | 2.579 | | |
| 8 | 13 | 2.546 | | |
| 9 | 2 | 2.526 | | |
| 10 | 4 | 2.306 | | |
| 11 | 14 | 2.305 | | |
| 12 | 1 | 2.178 | | |
| 13 | 5 | 2.153 | | |
| 14 | 12 | 2.070 | | |
| 15 | 11 | 1.923 | | |
| **16** | | | **3** | **1.776** |
| **17** | **16** | **1.761** | | |
| **18** | | | **6** | **1.747** |
| **19** | **17** | **1.735** | | |
| 20 | | | 5 | 1.614 |
| 21 | | | 4 | 1.471 |
| 22 | | | 8 | 1.447 |
| 23 | | | 7 | 1.300 |
| 24 | | | 2 | 1.207 |
| 25 | | | 9 | 0.781 |
| 26 | | | 1 | -0.026 |

The sets of the mean CR scores for the two conditions (LIG and MD) are not discrete, though the analysis ranks 15 of the 17 LIG texts above MD, and seven of the nine MD texts below LIG. The four scores in the overlap zone (ranked 16-19) are similar (standard deviation 0.015, variance 0.0002), and perhaps illustrate the limitations of the materials analysed. MD blog #3 (ranked 16) is quite short (only 48 nouns) and refers repeatedly to the writer's *father* (CR +4.184) and *therapist* (+3.406) or *psychologist* (+3.385). In LIG blog #16 (ranked 17), life is good for the writer personally, but "life is anything but good for so many of my fellow Americans".

### 7.13 Within Subjects: Three Case Studies

"Depression is related to suicide, though not all suicidal individuals are clinically depressed".
(Fernández-Cabana *et al*, 2013: 129)

These three case studies of depression are quite different from the case studies of dementia that were presented in the previous chapter, because the clinical basis of depression is quite different. Alzheimer's disease (AD) is a physiological disorder, progressive, degenerative, irreversible, and associated with the atrophy of brain function. Depression is classed primarily as a psychological disorder, though associated with specific areas of the brain (the limbic system), and with symptomatic changes in 'brain chemistry' (low levels of serotonin) that can be medicated by a class of anti-depressants known as selective serotonin reuptake inhibitors, or SSRIs (Andrews and Jenkins, 1999: 130-131; but see Bentall, 2009: 208-212 for a critical review). Because depression is generally not a progressive disorder, it does not have the inevitability of AD. Depression is not a constant – contrary perhaps to the assumptions of Stirman and Pennebaker (2001) in their study of suicidal poets (see section 7.4). It is therefore difficult to associate a specific language sample with the author's contemporaneous 'level' of depression.

Two of the subjects in these case studies are probably (along with Ernest Hemingway) the most famous 'literary suicides' of the 20th century – Virginia Woolf and Sylvia Plath. The third is Frances Medley, author of a blog that she entitled *Victorious Endeavours*, who took her own life in 2013 after a long illness. The tentative assumption, though one possibly challenged by the case of Frances Medley, is that suicide is "a proxy of depression" (Protopescu *et al*, 2012: 1080; Lundin and Hansson, 2014: 666). The standard inquest verdict of "suicide while the balance of the mind was disturbed" is given additional credence by a clinical guide published by the University of Oxford Centre for Suicide Research (*Assessment of suicide risk in people with depression*, 2012), which asserts that depression is a significant factor (though, it should be noted, not necessarily the primary motivator) in at least 60% of suicides.

There is a balancing view, that many 'successful' suicides are not spontaneous acts of despair. On the contrary, they are carefully planned and premeditated acts that give their perpetrators a sense of being in control of their own destiny (Courage *et al*, 1993). Frances Medley's rational decision to end her own life fits this paradigm.

There is also a distinction drawn between high and low intentionality, with stronger evidence of depression in cases of high intentionality (Gorenc, Kleff and Welz, 1983) such as Woolf in 1941 and Plath in 1953, than in cases of low intentionality or of impulsive suicides (Spokas, Wenzel, Brown and Beck, 2012).

The methodology adopted for these case studies samples two items of language from each subject. These are all personal communications, in the form of letters, a journal and a blog posting. Four of the six samples – and perhaps ironically they are those of the professional authors – were not written for publication. In each case, one sample is from a period when the suicidal mindset was apparently absent, as inferred from the content and biographical context. The other language sample is the closest extant communication to the intended suicide.

### Virginia Woolf

Virginia Woolf (1882-1941) was a leading member of the Bloomsbury Set of writers, artists and thinkers, who were very influential in the early decades of the twentieth century. She published novels, short stories, plays, and essays of literary criticism. Her collected letters occupy six volumes. Aged 59, she took her own life by drowning. She walked to the nearby River Ouse, filled her coat pockets with stones, and waded into the cold water. Her body was recovered three weeks later. The coroner's verdict was the standard one, that she did so "while the balance of her mind was disturbed", having endured recurring episodes of a mental illness which might now be diagnosed as bipolar disorder, characterised by recurring episodes of depression (Lee, 1997: 172).

Earlier in the month of her death a friend had described her as "desperate – depressed to the lowest depths". In his introduction to her collected letters (1980), Nigel Nicholson assesses her "motives for suicide. She believed that she was about to go mad again, and would not recover. She was hearing voices. She wanted to spare Leonard [her husband] the anxiety and terrible responsibility of caring for her". Her much-quoted final letter to Leonard is copied below in its entirety: 193 words, with just 15 nouns.

28 March 1941

> Dearest,
>
> I feel certain that I am going mad again. I feel we can’t go through another of those terrible times. And I shan’t recover this time. I begin to hear voices, and I can’t concentrate. So I am doing what seems the best thing to do. You have given me the greatest possible happiness. You have been in every way all that anyone could be. I don’t think two people could have been happier ‘til this terrible disease came. I can’t fight any longer. I know that I am spoiling your life, that without me you could work. And you will I know. You see I can’t even write this properly. I can’t read. What I want to say is I owe all the happiness of my life to you. You have been entirely patient with me and incredibly good. I want to say that – everybody knows it. If anybody could have saved me it would have been you. Everything has gone from me but the certainty of your goodness. I can’t go on spoiling your life any longer. I don’t think two people could have been happier than we have been.
>
> V.

(Woolf, 1980: 481).

The comparator letter was written in 1928 to Leonard, from France. The extract copied below conveys the tone of the whole (much longer) letter.

28 September 1928 (extracts)

> Vita was a perfect old hen, always running about with hot water bottles, and an amazingly competent traveller, as she talks apparently perfect French. I don't think we shall quarrel – indeed, I feel more established, now that we pay little

> attention to the other's moods; not that she has many. The truth is she is [*sic*] an extremely nice, kind nature; but what I like, as a companion, is her memories of the past. She tells me stories of the departed world—Mrs. Keppel, King Edward, how she stayed with the Rothschilds at Chantilly and they ran over a big dog in a motor car and wouldn't stop because they were late for their polo. Then I tell her the life story of Saxon. Then I cross-examine her scientifically; and ask her what she thinks happens if a motor car in which one is travelling at 50 miles an hour is struck by lightning. She has been told that owing to its rubber tyres it is a perfect non-conductor. Then we discuss her lectures on modern English poetry – which by the way she is ready to let us have for a pamphlet if we like…
>
> Lord! how I adore you! and you only think of me as a bagfull of itching monkeys, and ship me to the Indies with indifference!
>
> I think we shall have a very happy and exciting autumn, in spite of the complete failure of Orlando [her latest novel]. It is clearing slightly – we may visit the museum.

(Woolf, 1977: 538-539).

**Table 7.3:** Virginia Woolf: Genitive Ratio Analysis

**Prediction**: The mean CR will be significantly lower in the later sample, indicating an increase in the use of abstract language

| **Source** | **Date** | **Nouns** | **Mean CR** |
|---|---|---|---|
| Letter | 1928 | 66 | +2.549 |
| Letter | 1941 | 15 | -1.455 |

The results of the CR analysis are reported in Table 7.3, and support the prediction of an increase in abstract language in Virginia Woolf's suicide note, compared with the earlier letter [*t* (79) = 2.52, *p* = .007].

**Sylvia Plath**

Sylvia Plath (1932-1963) was born in the USA and came to England to study at Cambridge. Author of a novel, short stories and poetry, she married the British poet Ted Hughes in 1956. Like Virginia Woolf, she had suffered periodic bouts of depression. She died by inhaling coal gas from a domestic oven. Her final letter was written to her mother, one week before her death. This is an edited but representative extract:

4 February 1963

> I appreciate your desire to see Frieda [Sylvia's daughter], but if you can imagine the emotional upset she has been through in losing her father and moving, you will see what an incredible idea it is to take her away by jet to America. I am her one security and to uproot her would be thoughtless and cruel, however sweetly you treated her at the other end ... The children need me most right now, and so I shall try to go on for the next few years writing mornings, being with them afternoons and seeing friends or studying and reading evenings.

Shea (2011: 10), in a practitioner's account of the risk factors that might predict a suicide, can find no such factors in this letter. He finds her letter "particularly puzzling. Every time I read it, I must remind myself that its author killed herself just seven days later".

The poet and author Al Alvarez, who was a friend of Plath and Hughes, offers a possible explanation for the absence of evident risk factors in that final letter: "I am convinced by what I know of the facts that this time she did not intend to die" (Alvarez, 1971: 49). Though his conclusion has been challenged, Alvarez presents persuasive circumstantial evidence in support of his belief. Note his phrase "this time". Perhaps Plath had thought that she would be saved this time, as she had been ten years previously.

Plath's mindset might or might not have been genuinely suicidal at the time of her death, but it most definitely was when she had previously attempted suicide, at the age of 20 in 1953. Alvarez continues: "Her suicide attempt ten years before had been, in every sense, deadly serious" (*ibid*). She had carefully concealed herself in the crawl space underneath her family home and swallowed 50 sleeping pills (Stevenson, 1989: 43-47). "She was found late and by accident, and survived only by a miracle" (Alvarez, 1971: 49). For the next six months she was confined to a psychiatric institution, where the treatment included electro-convulsive therapy (ECT). Her novel *The Bell Jar* (1963) gives an accurate account of this episode.

This case study, then, is based on Plath's last journal entry before her attempted suicide on 24 August. In this extract, she is addressing herself:

14 July 1953

> You looked around and saw everybody either married or busy and happy and thinking and being creative, and you felt scared, sick, lethargic, worst of all, not wanting to cope. You saw visions of yourself in a straight jacket [*sic*], and a drain on the family, murdering your mother in actuality, killing the edifice of love and respect.

(Plath, 2000: 186-187).

The comparator is a letter written to her mother (to whom she was very close) on 27 September 1950, when Plath was just beginning life as a student at Smith College, Massachusetts. The letter begins:

Dearest Mummy,

> Well, only five minutes till midnight, so I thought I'd spend them writing my first letter to my favorite person. If my printing's crooked, it's only because I drank too much apple cider tonight.

The CR analysis of the two letters is reported in Table 7.4. Statistical analysis finds no significant difference between the 1950 letter and the pre-suicide 1963 letter [*t* (106) = 0.97, *p* = .167]. There is a significant difference between the 1950 letter and the 1953 journal entry written before the suicide attempt [*t* (144) = 1.80, *p* = .037].

**Table 7.4:** Sylvia Plath: Genitive Ratio Analysis

**Prediction**: The mean CR will be significantly lower in the sample closest to the suicide attempt (1953), indicating an increase in the use of abstract language, but not in the 1963 'final letter'.

| **Source** | **Date** | **Nouns** | **Mean CR** |
|---|---|---|---|
| Letter | 1950 | 38 | +2.944 |
| Journal | 1953 | 108 | +1.755 |
| Letter | 1963 | 70 | +2.103 |

**Frances Medley**

Frances Medley (1969-2013) was diagnosed with multiple sclerosis (MS) in 2005, forcing her to abandon a very promising career in arts administration (she was Chief Executive of the Arts Council of Wales). From January 2011 she wrote a blog, the title of which – *Victorious Endeavours* – reflected her positive outlook on life. Over the course of the next two years her health (she referred to her MS as 'Cruella') and her standard of living gradually declined.

Her final blog post, dated 23 September 2013 and entitled *A Sophisticated Sign Off*, was posted by friends after her death. In it she wrote that she had resolved to end her life "in a manner and at a time of my choosing; I am very clear that, whilst the law might say otherwise, I AM NOT COMMITTING SUICIDE." Though the formal verdict of the inquest had to be one of suicide, the coroner respected her wishes, by concluding that she had indeed "ended her life at a time and in a manner of her own choosing". She was 44 years old. This is the final section of that final posting. 'The Spinster' is how she habitually referred to herself in the blog:

> The Spinster fortunately peaked early on life [*sic*] and so I don’t leave with rafts of regrets or things I wish I’d done. Happy with my lot is perhaps an exaggeration but had the Spinster persisted my ability to do things would have been daily reduced; my potential it seems has been fulfilled. The values by which the Spinster has conducted her life are: clarity, integrity and wisdom with curiosity and creativity added in for Victorious Endeavours. These principles have served the Spinster well as I am leaving this mortal coil with a clear conscience albeit with a limited bank balance! Integrity is not a road paved with gold!!

> So live life as though it could be snatched away from you in a heartbeat; take managed risks avoiding recklessness; and treat your fellow travellers with tenderness and care. Hold your tongue at times when you risk blurting out judgemental potentially hurtful comments; we seldom know the full back story.
>
> Good bye and good luck ladies (and fellow male travellers too).
>
> The Spinster signs off with sophistication.

The blog posting for comparison is that first one, dated 2 January 2011. It is a long (2,007 words) and affectionate account of a family Christmas. This short extract is typical of the tone:

> I often receive books I've already read, I did this year. But I realized that I have reached the status of maiden aunt in training given some of the gifts received from extended family this year. I thought I might have a few more years but no I am considered suitable for a talcum dusting powder set with the word 'eau d' in the title. I do so hate Lily of the Valley and can barely tolerate lavender so I am utterly doomed!

http://victoriousendeavours.wordpress.com

The results of the CR analysis are reported in Table 7.5. As measured by the nouns' CR scores, although there is a reduction in concreteness, the difference between the language of the *Sophisticated Sign Off* and the earlier posting does not achieve significance, so that the null hypothesis cannot be rejected [*t* (473) = 1.41, *p* = .080]. Frances Medley would probably have considered this a fitting result.

**Table 7.5:** Frances Medley: Genitive Ratio Analysis

**Prediction**: The mean CR will be lower in the blog entry written just before Frances Medley's death, indicating an increase in the use of abstract language.

| **Source** | **Date** | **Nouns** | **Mean CR** |
|---|---|---|---|
| Blog | 2011 | 393 | +1.540 |
| Blog | 2013 | 82 | +0.679 |

**Summary**

Suicide is a debatable proxy for depression, particularly if isolated from its context and precedent behaviours. The American poets Hart Crane, Anne Sexton and John Berryman all committed suicide, but all have been reported to have endured a lifelong struggle with periods of depression, with insufficient biographical information confidently to locate representative depressed and non-depressed writings. The imminently suicidal writings of both Virginia Woolf and Sylvia Plath reflect despair, negativity, hopelessness, and perhaps even guilt about the effects of their illness on others. Woolf "can't fight any longer. I know that I am spoiling your life". Plath sees herself as "a drain on the family". The final communication of Frances Medley is qualitatively different. There is anger, resignation, even some humour, and a capitalised rejection of the word "suicide". She is in control.

If that psychological analysis of the three cases is accepted, then the genitive ratio analysis might have some validity: in cases of depression, it might measure the progression from concrete to more abstract language that is predicted by the theory of 'reduced concreteness thinking'.

### 7.14 LIWC and the Concreteness Rating Compared

LIWC (see section 7.3 above) is the most widely used linguistic toolkit for the analysis of psychological states. It is therefore appropriate to compare it with the CR analysis. Since several case studies have examined the language of suicide using LIWC analysis, the comparison will be based on the three within-subject case studies in this chapter (Woolf, Plath and Medley). Since all three case studies relate to women, gender differences are not a factor. This is significant, since gender differences in LIWC analyses of suicidal language have been reported by Lester, Haines and Williams (2010), Dogra *et al* (2007), and Newman, Groom, Handelman and Pennebaker (2008).

The LIWC study of suicidal poets by Stirman and Pennebaker (2001) has already been cited (see section 7.4). There have been several other studies of suicidal language using LIWC analysis. These include the writings of the explorer Henry Hellyer (Baddeley, Daniel and Pennebaker, 2011); the diaries of the Italian poet and novelist Cesare Pavese (Lester, 2009); and Marilyn Monroe's letters, notes and poems (Fernández-Cabana *et al*, 2013).

Based on the findings of these and other case-studies, the LIWC dimensions that are apparently most relevant to a suicidal mindset are self-references, positive and negative emotion words, social and cognitive words, together with direct references to religion and death (Fernández-Cabana *et al*, 2013: 125). Although the directionality of these dimensions is not wholly consistent, the most indicative seem to be increased or relatively high levels of self-reference and emotion words (both positive and negative).

Table 7.6 presents the results of a LIWC analysis of the Woolf (1928 and 1941), Plath (1950 and 1953) and Medley (2011 and 2013) texts, focusing on the dimensions that are most pertinent to suicidal language. The right-hand column is the LIWC benchmark for texts that are classed as "personal" rather than "formal". The four LIWC dimensions in the table are among those provided by the website Testing LIWC Online (liwc.net/liwcresearch07.php). The scores represent percentages of the total number of words in each text.

**Table 7.6**: LIWC analysis of three suicide case-studies

| LIWC Dimension | Woolf | | Plath | | Medley | | LIWC Benchmark |
|---|---|---|---|---|---|---|---|
| | *1928* | *1941* | *1950* | *1953* | *2011* | *2013* | |
| Self-reference | 6.75 | 13.30 | 10.53 | 0.63 | 4.91 | 5.54 | 11.4 |
| Positive emotion | 4.76 | 4.93 | 3.76 | 4.43 | 3.50 | 3.02 | 2.7 |
| Negv emotion | 1.19 | 1.97 | 0.75 | 5.38 | 0.78 | 3.02 | 2.6 |
| Articles | 8.73 | 1.97 | 6.77 | 4.75 | 7.88 | 8.82 | 5.0 |

Consider the four LIWC dimensions in Table 7.6:

**Self-references**. In the Woolf texts, the incidence of self-reference doubles between 1928 and 1941. This is very much in line with expectations from prior LIWC studies, but the Plath scores are completely opposite – minimal use of first-person pronouns in the later text. There is no significant difference in the Medley scores.

**Positive emotion words**. Although all three cases score higher than the LIWC benchmark, there are no significant changes within the cases.

**Negative emotion words**. This and self-reference are the dimensions most associated with depression (Rude, Gartner and Pennebaker, 2004), and all three cases (particularly Plath) show significant increases.

**Articles**. Reference has already been made to the significance of article usage (see section 7.3) in "referring to concrete and impersonal objects or events" (Pennebaker and King, 1999: 1309). If an inhibition of concrete language is indicative of depression, then one would expect to find a reduction in article usage, and this is the pattern observed for Woolf and Plath, though not for Medley.

Both the above analysis and the cited studies illustrate the strengths and weaknesses of LIWC. From the 80 output variables provided by LIWC 2007 (Pennebaker, Francis and Booth, 2007), the software highlights the key dimensions, but the pattern of results is not consistent across all cases. Some studies have found significance in cognitive words (Lester, Haines and Williams, 2010) or in social words (Pennebaker and Stone, 2003; Newman *et al*, 2008). There is a sense that the number of output variables provides ample scope for cherry-picking the significant dimensions on a case-by-case basis.

It is difficult to perceive, from a detailed analysis of each cited case-study, a clear model of suicidal language, although Fernández-Cabana *et al* (2013: 129) conclude, from a clinical psychiatric perspective, that LIWC "could be a useful forensic instrument in analyzing suicide notes as well as samples of texts written by suicidal individuals or by people who have made suicide attempts, in order to reach a deeper understanding of the phenomenon and contribute to its prevention".

Perhaps LIWC, in providing such a wide range of dimensions for analysis and interpretation, is a useful tool for psychiatric clinicians. As a measurement of a depressive or suicidal mindset, it appears to be no more accurate than the single dimension of a CR analysis, and it is a good deal more complex.

### 7.15 The Concreteness Rating and Sentiment Analysis

Sentiment analysis is a method of 'mining' texts in order to identify opinions and attitudes. This brief discussion will consider whether current sentiment analysis methods might predict depression, with a focus on document-level sentiment analysis and on one of its sub-fields, emotion detection. See Feldman (2013) and Taboada (2016) for overviews of the techniques and applications of sentiment analysis.

Sources for texts have typically been social media (e.g. Twitter and Facebook), blogs and online reviews of products, films, hotels, etc – "a gold mine for companies and individuals that want to monitor their reputation and get timely feedback about their products and actions" (Feldman, 2013: 82). However, the applications of sentiment analysis are not entirely commercial. Pestian *et al* (2012) have reported on a shared task document-level analysis of the emotional content of suicide notes, The "task" is primarily one for emotion detection: to determine, from notes left by attempted suicides, which of them are likely to make another attempt. The results suggest that "human-like performance on this task is within the reach of currently available technologies" (*ibid*: 3).

**Table 7.7**: Sentiment analysis and emotion detection of Woolf and Medley case-study texts (source PreCeive API)

| | Woolf (1928) | Woolf (1941) | Medley (2011) | Medley (2013) |
|---|---|---|---|---|
| **Sentiment Analysis** | | | | |
| Negative | 38.6% | 31.3% | 41.6% | 65.1% |
| Neutral | 8.7% | 7.5% | 5.7% | 3.8% |
| Positive | 52.7% | 61.1% | 52.6% | 31.1% |
| **Emotion Detection** | | | | |
| Anger | +1.205 | +0.448 | +0.451 | |
| - Agitated Calm + | +4.380 | +1.299 | +2.045 | +1.066 |
| Fear | | +1.491 | +0.053 | +3.743 |
| - Sad Happy + | +4.787 | +1.456 | +4.624 | +1.347 |
| - Dislike Like + | +3.668 | +1.737 | +3.549 | +1.889 |
| Shame | | | +0.791 | +0.718 |
| - Unsure Sure + | | +0.132 | +0.120 | -0.562 |
| Surprise | +0.493 | | +0.225 | |

Table 7.7 presents a document-level sentiment analysis of the texts that were analysed in the case studies of Virginia Woolf and Frances Medley, with overall ratings of negative, neutral and positive content, together with scores for emotion detection across a range of parameters (anger, fear, etc). The analysis is provided by the PreCeive API Demo of the TheySay sentiment analyser, a state-of-the-art commercial product (URL: apidemo.theysay.io).

Look first at the emotion detection scores in Table 7.7. Although they are (with one exception) positive, it is their direction from the early to the final letter that is most interesting. In both case-studies, the emotional content becomes less calm, more fearful, less happy – as would be expected. Now consider the overall sentiment analysis of each document, which measures the polarity of sentiment with classifications of negative, neutral and positive content. The results accurately reflect a reading of the Frances Medley texts, with her final blog entry (2013) scoring significantly higher on negative sentiment compared to the 2011 entry (from 41.6% negative in 2011 to 65.1% in 2013), and significantly lower on positive sentiment (from 52.6% positive in 2011 to 31.1% in 2013).

By contrast, Virginia Woolf's final letter (1941) gains a positive score of 61.1% compared to the 'happy' 1928 letter at only 52.7%. Analysis of the 1941 letter by another provider of sentiment analysis software, Buzzlogix (www.buzzlogix.com/ text-analysis/demo) asserts an even stronger "positive" rating for the document, and with a "99.96 probability". Does Woolf's suicide letter confound sentiment analysis?

Yes and no. It is the absence of self-pity that makes Woolf's suicide letter so poignant. Her principal purpose in writing the letter is indeed positive, to absolve her husband Leonard of any guilt or responsibility for the action she is about to take. Sentiment analysis accurately measures that purpose, but does not detect the subtext of nihilism and despair. In linguistic analysis, degrees of certainty are never absolute, and subtexts – for example, irony, sarcasm and satire – are a particular challenge for sentiment analysis, because they present a "contrast between a positive sentiment and a negative situation" (Riloff *et al*, 2013).

Sentiment analysis "extracts information from positive and negative words in text" (Taboada, 2016: 8.1). If we go back to the Woolf 1941 text and separate the positive and negative adjectival and adverbial phrases from their contexts, we see this:

| | | | |
|---|---|---|---|
| [7.2] | +certain | +best | -mad |
| | +greatest possible | +in every way | -terrible |
| | +happier | +entirely patient | -terrible |
| | +incredibly good | +happier | |

In this surface-level analysis, positives significantly outweigh negatives, and even *mad* is ambiguous (as in "I'm mad about football").

"Entities, which can comprise anything from mentions of people or organisations to concrete or even abstract objects, condition what a text is ultimately about" (my emphasis, Moilanen and Pulman, 2009: 258). A perhaps simplistic hypothesis is that nouns specify what a speaker wants or needs to talk about; the adjectives, adverbs and verbs that she chooses reflect how she talks about it (Taboada, 2016: 8.4). Emotion detection is typically more reliant upon qualifiers (adjectives and adverbs) than upon nouns. Three of the four polarised vectors in the PreCeive model are pairs of adjectives:

| | | | | |
|---|---|---|---|---|
| [7.3] | - | Agitated | Calm | + |
| | - | Sad | Happy | + |
| | - | Unsure | Sure | + |

It is a basic premise of sentiment analysis that the words that we choose are conditioned by sentiment – by our attitudes, beliefs and emotions. We have seen (in 7.8) that the rationale of reduced concreteness training rests, *inter alia*, on an assumption that the same event can be narrated in different ways, because the narrator's word-choice is conditioned by his or her cognitive state. As stated in the introduction, the gradient of animacy that is defined by genitive ratio analysis is a concept with cognitive as well as linguistic dimensions. A depressed person does not consciously select predominantly abstract nouns; the choices made are a product of their cognitive state.

Whilst the concreteness rating is not a direct competitor to sentiment analysis, there is at least the possibility that they might be complementary, with the genitive ratio analysis adding a cognitive dimension to the linguistic dimensions of sentiment analysis. How and if that might usefully be applied must be the subject of further research.

### 7.16 The Ethics of Implementation

By monitoring users' language on social media, linguistic analysis offers low-cost detection of early indicators of depression, but the ethical issues are complex. Any such applications must establish a framework that combines the informed consent of all participants with a response to alerts that is both effective and discreet, whilst acknowledging that the system will not be definitive.

Burns *et al* (2011) have piloted a smartphone app called Mobilyze!, based on "ecological momentary intervention and assessment", to evaluate a user's level of depression. Sensor data collected by the app include the user's location, social context, and recent activity, with periodic requests for users to report on their current "internal states", such as their mood, fatigue, emotions, senses of pleasure and accomplishment, and capacity for concentration. A machine learning algorithm then builds a predictive model for each user, capable of inferring their level of depression. When a level higher than threshold is reached, the app generates alerts to pre-selected family and friends, as well as directing the user to a web-site that offers CBT/self-help intervention activities, together with the option of telephone contact with a coach or therapist.

Jashinsky *et al* (2014) have used a linguistic analysis of Twitter conversations to identify users who might be suicide risks, by searching for key words and phrases such as "I'm being bullied" or "had thoughts [of] killing myself". The acknowledged limitation of this approach is that it is dependent on a reliable compilation of the vocabulary of suicide, with neither too few nor too many phrases. As the authors note, "There is undoubtedly a balance that must be achieved: a sufficient number of search terms to identify risk, but not too many so as not to falsely determine risk" (Jashinsky *et al*, 2014: 57).

In October 2014, the Samaritans (a suicide prevention charity) launched in the UK a smartphone app called Samaritans Radar. The app monitors a user's Twitter contacts to look for words and phrases that might indicate a suicidal mindset, triggering an email alert to the user whenever such indicators are detected. The advantages of basing the app on Twitter are that it has a very large user-base, with over 15 million users in the UK; most of the data is publicly accessible (thus supposedly skirting privacy issues); and it is currently the preferred social medium for spontaneous personal expression.

The Samaritans' initiative offers two important lessons for the deployment of linguistic analysis in the field of mental health. First, the detection of key phrases, without any reference to their precedents or contexts, is a rather 'blunt instrument'. An obvious phrase-search term might be "kill myself". The Birmingham Blog Corpus (Kehoe and Gee, 2012) contains 324 instances of that phrase (as at October 2014). The vast majority are colloquial or ironic usages, e.g.

"My plan for this race was to give a good, solid effort, but not to kill myself" (posted by a triathlete).

Second, the app did not take account of privacy issues. Within days of its launch, an on-line petition was demanding its withdrawal. Detractors warned that indications of an individual's vulnerability would be manna for cyber-stalkers and on-line bullies ("trolls" in the argot of the internet). A BBC report (4 November 2014) described the "overwhelming response [as] negative", quoting a typical comment: "How dare you interfere in the complicated emotional lives of others without so much as a by-your-leave? This is appalling".

Lessons need to be learned. If the genitive ratio is to have a clinical role, it should be as an adjunct to a clinically-sponsored initiative such as concreteness training, professionally administered and delivered to participants who are both co-operative and informed.

## 7.17 Caveats and Conclusions

The primary focus of this thesis is computational and linguistic. This is not a clinical study. The purpose of the analyses presented in this chapter is to test the feasibility of the genitive ratio as a tool for measuring language, and to suggest that computational linguistic analysis, based on the relative usage of concrete and abstract language, might have a role in monitoring depressive disorders.

The three machine learning studies reviewed in 7.5 relied on archived data (suicide notes and audio recordings), simulators and self-diagnosed depressives. As with this study, none had access to a current clinical population. As far as I am aware, none of these studies has been tested in a clinical context. This is an important issue. On its own, or even when combined with other quantitative measures, the genitive ratio is of little value in this context. Its potential value will only be demonstrated when it is tested in a clinical context, and that will only happen when clinicians, rather than computational linguists, take the initiative. There are potentially significant benefits from a successful implementation. Earlier identification of a patient's depressive decline might facilitate a speedier recovery, and with a reduced reliance on medication.

The concreteness training (CNT) therapy, developed by Watkins and colleagues at the University of Exeter, to treat cases of mild to moderate depression, offers a low-cost alternative to other possible interventions – because it can be automated. It supports and guides the patient through a course of self-help therapy. The genitive ratio might complement CNT by analysing the language of patients' responses, in order to measure their progress (or otherwise) through the course of their therapy.

An issue for both clinical applications is that symptoms of dementia and depression sometimes affect the same patient, they are "co-morbid". Based on a longitudinal analysis of clinical data extracted from the National Alzheimer's Coordinating Center Uniform Data Set, Masters, Morris and Roe (2015) have obtained results suggesting that "depressive symptoms may increase with age regardless of incipient dementia". Given that AD causes damage to the brain that is physiological, permanent and progressive, intuitively dementia would be the dominant condition. Without access to a clinical population it is impossible to separate out the effects on patients' relative usage of concrete and abstract language, but anecdote supports intuition. This is from the obituary of psychiatrist Dr Alice Roughton (in *The Independent*, 29 June, 1995): "[Her] increasing dementia brought one benefit ... She had become quite depressed. Her illness returned her to a simpler, more optimistic phase of her life, and her wide smile resumed its spontaneity".

CHAPTER 8

# Reflections

# and

# Directions

*The endless cycle of idea and action,*

*Endless invention, endless experiment.*

T.S. Eliot: *The Rock*

## 8.0 Themes

Each of chapters 2-7 has ended with a discussion of the conclusions drawn from the evidence presented in that chapter. It is not the intention here simply to reiterate all of those conclusions. The intention is rather to reflect upon three over-arching themes that will place the genitive ratio into a perspective of both its potential and its limitations.

## 8.1 Progress Through Innovation

This was the research goal as stated in chapter 1:
**To devise a computational model, based on the genitive ratio, that will reliably classify nouns as animate, concrete or abstract**.

If we define "reliably" as significantly above chance, then the goal has been attained, with reported success rates of over 80% for both the animateness and the concreteness ratings. That represents a more than adequate proof of concept. It has been achieved through a cross-disciplinary process of experiment and analysis.

**The genitive ratio as concept and method**. Whilst the evidence for the role of animacy within genitive selection is well-established, the genitive ratio (GR) as a proxy for relative animacy is a new concept, within the scope of published research. As a computational method, GR analysis has advantages over alternative applications. GR analysis is relatively simple and works well with quite short texts (Virginia Woolf's final letter quoted in chapter 7 has only 15 nouns).

The principal advantage of the GR algorithms is that their application is bounded only by the limits of their data source. The English language accommodates colloquialisms, variant spellings, domain-specific terminology and an ever-expanding vocabulary. Given a sufficiently large data source, the GR will cope with all of these, beyond the range of dictionaries, ontologies such as WordNet, and ratings exercises. Dictionary-based systems such as LIWC (see chapter 7.3) are constrained by the scope of their lexicons. They are reliant upon their designers to pre-select the most relevant keywords.

**Cross-disciplinary approach**. The scope of the research undertaken should be apparent from the literature reviews and the bibliography. The research goal has been achieved by drawing ideas, inspiration and support from psychology, linguistics and computer science. Van Deemter *et al* (2012) have suggested that psycholinguistics and computational linguistics can be mutually supportive, even though their focus differs. Although psycholinguists are interested in the *process* of natural language, whereas computational linguists are interested in language as a *product*, it is possible for the two disciplines to be complementary. In this thesis, models of language production put forward by linguists and psycholinguists have been adapted and subjected to the computational linguists' more rigorous demands of algorithmic definition.

**Animyser**. The GR algorithms achieve results that are comparable with multi-factor machine-learning models, but with potentially lower costs. The Animyser program, written specifically to support this research, could be extended and adapted to new data sources.

**Experimental design**. The design of the online sentence production experiments in two languages (chapter 5) is, to the best of my knowledge, original. With participant recruitment via social media, the field-tested design offers significant benefits of large-scale data obtained at low cost, and of a quality comparable with laboratory-based experiments.

## 8.2 Data: Challenges and Solutions

**Quality of source data**. Whilst the application-based tests in chapters 5-7 demonstrate the relevance of the three-category (animate-concrete-abstract) differentiation, the prospect of a finer-grained analysis, as suggested by the six-category model in chapter 3, has not been realised. The difference in results obtained from a supervised (Denison *et al*, 2008) versus an unsupervised (Wikipedia) corpus suggests that the quality and breadth of the source data are critical success factors.

As a corpus, Wikipedia delivers benefits of extensive subject coverage and well-structured written language, but those are also its weaknesses. It does not reflect colloquial or conversational English; its primary role as an encyclopaedia skews results-counts (as in phrase-search results for *anger* that derive disproportionately from references to *Kenneth Anger*); and even the vast Wikipedia corpus returns some zero phrase-search counts. The principal internet search engines, Google Search and Microsoft's Bing, offer much larger and more diverse data-sets than Wikipedia's, but they lack the necessary phrase-search facilities and/or results-count accuracy, at least at the level of public access to their databases.

The two genitive ratio algorithms defined in chapter 4 are the product of their  different functions, but they are also the product of their source data, its quality and scope, and particularly the incidence of low and zero counts for abstract nouns, even in Wikipedia. Better source data might have produced a better (single) algorithm. The ideal data source would combine the scope of the Google or Bing databases with the tagged annotation of a supervised corpus.

There is a possible solution, though it would require the resources, processing power, and above all collaboration of a Microsoft or a Google: what might be termed just-in-time (or just-enough) annotation. Instead of simply augmenting the count of a target phrase, each successful location of the phrase within the database would trigger a tagged annotation of the context (sentence or paragraph) in which the target phrase occurred. Most confounds (such as auxiliary verb contractions and proper nouns) would thus be detected and eliminated. Target phrases that are currently excluded from phrase-search, because an adjective or adjunct intervenes between a determiner and the noun, could be detected (using the wild-card facility) and included in the count. All relevant genitive constructions would thus augment the phrase-search count.

**Sparsity of clinical data**. A recurring theme in chapters 6 and 7, and in other studies cited there, is the difficulty of obtaining data that have clinical credibility. In all three accounts of the longitudinal studies of dementia – the Nun Study, the Western Collaborative Group Study and the Precursors Study – the texts that facilitated linguistic analysis were obtained through serendipity rather than through visionary planning. The fact that studies, including this one, have had to rely on textual comparisons of materials in the public domain tells its own story.

Mihalcea and Liu (2006) have annotated a corpus of blog posts from LiveJournal.com with 'happy' and 'sad' tags, based on a list of key words and phrases. Their analysis yields the "happiness trajectory" of a day and a week. They conclude with a (tongue in cheek) "corpus-inspired liveable recipe for happiness". However, diurnal bouts of sadness are not equivalent to depression.

There is cumulative evidence, from this and other studies, of at least a *prima facie* case for linguistic analysis as a viable tool for prognostic monitoring, but in fields such as dementia and depression the respected research is physiological or bio-chemical rather than linguistic, and understandably so. Clinicians will look for hard evidence in the form of replicated studies before they are convinced that linguistic analysis has a credible role, and linguistic analysis will only be given that role if its implementation is championed by clinicians.

Because the time-course of dementia is typically slow, with sometimes a decade or more between the earliest and the most severe symptoms, longitudinal studies have been the gold standard of dementia research, but such studies are expensive, and they take by definition a long time to deliver results. Above all, it is difficult to specify at their outset the data and samples that might be relevant to the unforeseen technology and methodologies newly available by the time of their conclusion. The obvious example is the sequence of the human genome, published in 2003.

There is a pressing need for more data, and three ways in which that need might be met. First, bona fide researchers should be given open or at least qualified access to the data that are available, for example from the three studies cited above. In practice, requests for access have received no reply. There are of course concerns of ethics and participant anonymity, but they should not be insuperable. Second, computational linguists need access to current clinical data, preferably within formal research partnerships with clinicians. The (so far two) Computational Linguistics and Clinical Psychology Workshops (CLPsych 2014 and 2015) in the USA provide a template for collaboration.

A third way is that linguists seek out their own data sources, as with the Birmingham Blog corpus analysed in chapter 7. The design of the online experiment in chapter 5 offers a template for bringing in participants (either self-diagnosed or self-certified) from internet forums set up for people with depression, such as the mentalhelp.net website utilised by Neuman *et al* (2012).

Coppersmith *et al* (2015) used a "shared task" approach to diagnose depression and post-traumatic stress disorder (PTSD) from linguistic analysis, utilising data (in the public domain but anonymised) from two sets of Twitter users. One set with self-stated diagnoses of depression or PTSD was matched on age and gender with a set of controls.

With appropriate safeguards in place, web-based forums, groups and social media offer discrete populations of participants for future studies, though not without problems. Coppersmith *et al* (2015) had to rely on self-proclaimed rather than clinically confirmed diagnoses, and the age and gender of their participants had to be "estimated ... through analysis of their language" (*ibid*: 32). Whilst such studies do not carry clinical authority, their findings might strengthen the argument for access to clinical data.

**Data for diagnosis**. The analyses and discussions in chapters 6 and 7 have focused on the genitive ratio as a prognostic, rather than diagnostic, measure.

There are two ways in which genitive ratio analysis might form part of a diagnostic model. Both are dependent on the availability of data.

One option would rely upon locating language samples from the individual patient's pre-morbid period, for comparison. This was the premise of Garrard *et al*'s (2005) study of Iris Murdoch (see chapter 6). Diaries and blogs offer possible sources of data, as with Frances Medley (see chapter 7), but are maintained by only a small percentage of the population. The other option would rely upon the availability of sufficient data to establish norms with which individuals' language use might be compared. The analysis of the Birmingham Blog Corpus in chapter 7 suggests that this might be feasible.

Until there are completely accurate and cost-effective tests for conditions such as Alzheimer's disease and depression, clinicians will continue to rely upon multi-factor diagnostic models. It is at least conceivable that the inclusion of genitive ratio analysis in such a model might improve its accuracy.

## 8.3 Developing New Applications

The genitive ratio is a tool with potential applications wherever the relative animacy of language might be significant. The key word is "tool". If researchers have access to a relatively simple, automated method of classifying nouns as animate, concrete or abstract, they are more likely to test for the significance of animacy within their analyses of language samples. The following discussion of possible applications is speculative, based on little research, and merely seeks to suggest where a measure of animateness, concreteness or abstractness might be relevant.

**Cognitive neuroscience**. Abstract thinking "leads to a persistence of negative mood and arousal" (Echiverri et al, 2011: 344), symptoms that are associated with self-harm, addiction and eating disorders. A number of researchers are already addressing the challenge of applying computational linguistics to the field of cognitive neuroscience (see Garrard and Elvevåg, 2014, for a review). Four brief examples must suffice. An impairment of abstract thinking has long been recognised as a characteristic of patients diagnosed with schizophrenia (e.g. Harrow, Adler and Hanf, 1974; Oh *et al*, 2014). Concreteness training has a possible role in combating PTSD (Schaich, Watkins and Ehring, 2013). Concrete thinking is associated with autism and Asperger's syndrome (Hobson, 2012).

Beyond those clinical applications, "one challenge for future studies is surely to find a way to detect some aspect currently neglected of deceptive language" (Fornaciari and Poesio, 2013: 45). An avenue for future research might be to test deceptive language (in witness statements or court testimony, or even optimistic statements made in company reports) for a measurable bias towards either concrete or abstract language.

**Materials selection**. The abstract/concrete distinction has contributed to developing our understanding of how and where language is processed in the brain. Huang and Federmeier (2015) cite numerous studies in support of their own findings, based on event-related potential (ERP) measurements, that "the two hemispheres of the brain make different contributions to concreteness effects" (*ibid*: 507).

Psycholinguistic experiments that feature concrete and abstract words as their materials currently rely on their own classifications if their vocabulary is very specific, or on published ratings such as those of Brysbaert *et al* (2014). An alternative approach would be to select differentiated sets of words by their genitive ratios, or to use GR analysis as an additional check on the rated selections.

**Education**. Students at all levels of education are increasingly submitting written work in digital form. Whilst any proposal to monitor that material would raise ethical issues, there is nevertheless an opportunity to link students' language use, as determined by linguistic analysis, to aspects of their behaviour or academic progress, offering scope for timely interventions and preventative measures. Just as concreteness training (see chapter 7) offers an effective therapy for depression, similar CBT-based treatments might be designed, perhaps embedded in games, to encourage a healthy balance of concrete and abstract thinking in at-risk adolescents.

**Natural language processing (NLP)**. NLP has faced the challenge of progressing beyond systems that are domain-dependent and hand-crafted. There is now a recognition (e.g. Strube, Rapp and Müller, 2002) that co-reference resolution systems for example should be robust (not specific to a single domain) and accessible to automatic or at least semi-automatic annotation.

A limitation of some computational models discussed in chapters 5-7 is that they would be difficult to operationalise. They rely on extensive pre-processing or on a pre-annotated corpus or ontology. By contrast, the goal of the current research has been to devise an operationally-focused algorithm, the components of which are data-driven and relatively simple to implement. A model for this approach, with similar objectives, might be that taken to statistical machine translation, e.g. by Och and Ney (2004).

### 8.4 Coda

*The end is where we start from*.

T.S. Eliot: *Little Gidding*

This thesis began as one thing and has ended as something quite different. It began as an investigation focused on animacy as a factor in salience and co-reference resolution. In applying leverage to that problem, it was the lever, the genitive ratio, that became the focus. If we come to see that the distinction between animate, concrete and abstract referents is a significant factor in unlocking new fields of linguistic analysis, the genitive ratio will hopefully provide a key.

## REFERENCES

Ahmed, S., Haigh, A-M.F., de Jager, C.A. & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, 136, 3727-3737.

Almor, A. & Nair, V.A. (2007). The form of referential expressions in discourse. *Language and Linguistics Compass*, 1(1-2), 84-99.

Alparone, F., Caso, S., Agosti, A. & Rellini, A. (2004). *The Italian LIWC2001 dictionary*. Austin, TX: LIWC.net.

Altarriba, J., Bauer, L.M. & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete and emotion words. *Behavior Research Methods, Instruments and Computers*, 31(4), 578-602.

Altenberg, B. (1982). *The Genitive v. the Of-Construction: A study of syntactic variation in 17th century English*. Lund: CWK.

Alvarez, A. (1971). *The Savage God: A Study of Suicide*. London: Penguin.

Alzheimer's Society (2007). *Dementia UK: the full report*, produced by King's College London and the London School of Economics. London: Alzheimer's Society.

Amanzio, M., Germiniani, G., Leotta, D. & Cappa, S. (2008). Metaphor comprehension in Alzheimer's disease: Novelty matters. *Brain and Language*, 107, 1-10.

American Psychiatric Association (1989). *Diagnostic and Statistical Manual of Mental Disorders* (3rd edition). Arlington, VA: American Psychiatric Press.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th edition). Arlington, VA: American Psychiatric Press.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th edition). Arlington, VA: American Psychiatric Publishing.

Anderson, A.J., Murphy, B. & Poesio, M. (2014). Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *Journal of Cognitive Neuroscience*, 26(3), 658-681.

Andrews, G. & Jenkins, R. (Eds.) (1999). *Management of Mental Disorders* (UK Edition). Sydney: World Health Organization Collaborating Center for Mental Health and Substance Abuse.

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, 24, 65-87.

Arnold, J. (1998). *Reference Form and Discourse Patterns*. PhD Dissertation, Stanford University.

Arnold, J.E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, 31(2), 137-162.

Artiles, J. & Sekine, S. (2009). *Tagged and cleaned Wikipedia (TC Wikipedia)*. Available from http://nlp.cs.nyu.edu/wikipedia-data/.

Baayen, R.H., Piepenbrock, R. & van Rijn, H. (1993). *The CELEX Lexical Database* [CD-ROM]. Philadelphia: Linguistic Data Consortium.

Baddeley, A.D. (1986). *Working Memory*. Oxford: Oxford University Press.

Baddeley, J. L., Daniel, G. R., & Pennebaker, J.W. (2011). How Henry Hellyer's use of language foretold his suicide. *Crisis*, 32, 288–292.

Baldas, V., Lampiris, C., Capsalis, C. & Koutsouris, D. (2011). Early diagnosis of Alzheimer's type dementia using continuous speech recognition. In J. Lin & K.S. Nikita (Eds.). *MobiHealth 2010, LNICST 55*, 105-110.

Barenboym, D.A., Wurm, L.H. & Cano, A. (2010). A comparison of stimulus ratings made online and in person: Gender and method effects. *Behavior Research Methods*, 42(1), 273-285.

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3), 209-226.

Bates, E., Harris, C., Marchman, V., Wulfeck, B. & Kritchevsky, M. (1995). Production of complex syntax in normal ageing and Alzheimer's disease. *Language and Cognitive Processes*, 10(5), 487-539.

Battig, W.F. & Montague, W.E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, 80(3), Monograph, 1-46.

Baudic, S., Barba, G.D., Thibaudet, C.K., Smagghe, A., Remy, P. & Traykov, L. (2006). Executive function deficits in early Alzheimer's disease and their relations with episodic memory. *Archives of Clinical Neuropsychology*, 21, 15-21.

Bayles, K.A. & Tomoeda, C.K. (1983). Confrontation naming impairment in dementia. *Brain and Language*, 19, 98-114.

Bayley, J. (1998). *Iris: A memoir of Iris Murdoch*. London: Duckworth.

Bellow, S. (2007). *Collected Stories*. London: Penguin Classics.

Bentall, R. (2009). *Doctoring the Mind: Why psychiatric treatments fail*. London: Allen Lane.

Bergsma, S. (2005). Automatic acquisition of gender information for anaphora resolution. In *Proceedings of the 18th Conference of the Canadian Society for Computational Studies of Intelligence*.

Berisha, V., Wang, S., LaCross, A. & Liss, J. (2015). Tracking discourse complexity preceding Alzheimer's disease diagnosis: A case study comparing the press conferences of Presidents Ronald Reagan and George Herbert Walker Bush. *Journal of Alzheimer's Disease*, 45: 959-963.

Bertenthal, B.I. (1993). Infants' perception of biomechanical motions: Intrinsic image and knowledge-based constraints. In C. Granrud (Ed.). *Visual Perception and Cognition in Infancy: Carnegie Mellon Symposia on Cognition*. Hillsdale, NJ: Erlbaum.

Biber, D., Johannson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Bock, J.K. & Warren, R.K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21, 47-67.

Börjars, K., Denison, D. & Krajewski, G. (2011). Poss-s vs. poss-of revisited. Workshop on genitive variation in English, *ISLE2*, Boston, 18 June 2011.

Börjars, K., Denison, D., Krajewski, G. & Scott, A. (2013). Expression of possession in English: The significance of the right edge. In K. Börjars, D. Denison & A.K. Scott (Eds.) *Morphosyntactic Categories and the Expression of Possession*. Amsterdam: John Benjamins.

Borkovec, T.D., Ray, W.J. & Stöber, J. (1998). Worry: A cognitive phenomenon intimately linked to affective, physiological, and interpersonal behavioural processes. *Cognitive Therapy and Research*, 22(6), 561-576.

Branigan, H. (1995). *Language Processing and the Mental Representation of Syntactic Structure*. PhD Thesis, University of Edinburgh.

Branigan, H.P., Pickering, M.J. & Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, 118, 172-189.

Brennan, S.E., Friedman, M.W. & Pollard, C.J. (1987). A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the ACL,* Stanford CA.

Bresnan, J., Cueni, A., Nikitina, T. & Baayen, R.H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer & J. Zwarts (Eds.). *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science, 69-94.

Bresnan, J. & Hay, J. (2008). Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua*, 118, 245-259.

Brown, C., Snodgrass, T., Kemper, S.J., Herman, R. & Covington, M.A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2), 540-545.

Brunét, E. (1978). *Le vocabulaire de Jean Giraudoux. Structure et évolution*. Genéve: Slatkine.

Bryman, A. (2008). *Social Research Methods*. Oxford: Oxford University Press.

Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977-990.

Brysbaert, M., Warriner, A.B. & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904-911.

Bucks, R.S., Singh, S., Cuerden, J.M. & Wilcock, G.K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1), 71-91.

Burns, M.N., Begale, M., Duffecy, J., Gergle, D., Karr, C.J., Giangrande, E. & Mohr, D.C. (2011). Harnessing context to develop a mobile intervention for depression. *Journal of Medical Internet Research*, 13(3), e55.

Cantos-Gómez, P. (2010). Analysing linguistic decline in early-stage Alzheimer's disease: A corpus-based approach. In A. Sanchez & M. Almela (Eds.) *A Mosaic of Corpus Linguistics: Selected approaches*. Frankfurt am Main: Peter Lang.

Caramazza, A. & Shelton, J.R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, 10(1), 1-34.

Carlat, D. (2010). *Unhinged: The Trouble with Psychiatry – a doctor's revelations about a profession in crisis*. London: Free Press.

Carter, D. (2000). Discourse focus tracking. In H.C. Bunt & N.J. Black (Eds.). *Abduction, Belief and Context in Dialogue: Studies in computational pragmatics*. Amsterdam: John Benjamins.

Chand, V., Baynes, K., Bonnici, L.M. & Farias, S.T. (2012). A rubric for extracting idea density from oral language samples. *Current Protocols in Neuroscience*, Unit 10.5.doi: 10.1002/0471142301.ns1005s58.

Chertkow, H., Whatmough, C., Saumier, D. & Duong, A. (2008). Cognitive neuroscience studies of semantic memory in Alzheimer's disease. *Progress in Brain Research*, 169, 393-407.

Cios, K.J., Pedrycz, W., Swiniarski, R.W. & Kurgan, L.A. (2007). *Data Mining: A knowledge discovery approach*. New York, NY: Springer.

Clark, H.H. (1965). Some structural properties of simple active and passive sentences. *Journal of Verbal Learning and Verbal Behavior*, 4, 365-370.

Clark, H.H. & Begun, J.S. (1971). The semantics of sentence subjects. *Language and Speech*, 14, 34-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coltheart, M (1981) The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A, 497-505.

Comrie, B. (1981). *Language Universals and Linguistic Typology*. Oxford: Blackwell.

Connell, L. & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, 125, 452-465.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K. & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the Second Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, Colorado, June 5, 2015, 31-39.

Corley, M. & Scheepers, C. (2002). Syntactic priming in English sentence production: Categorical and latency evidence from an internet-based study. *Psychonomic Bulletin and Review*, 9(1), 126-131.

Cote, S. (1998). Ranking forward-looking centers. In M.A. Walker, A.K. Joshi & E.F. Prince (Eds.). *Centering Theory in Discourse*. Oxford: Clarendon Press.

Coulson, M. (1995). Anaphoric reference. In J. Greene & M. Coulson (Eds.) *Language Understanding: Current issues*. Buckingham: Open University Press.

Courage, M.M., Godbey, K.L., Ingram, D.A., Schramm, L.L. & Hale, W.E. (1993). Suicide in the elderly: staying in control. *Journal of Psychosocial Nursing and Mental Health Services*, 31(7), 26-31.

Craik, K. (1943). *The Nature of Explanation*. Cambridge: Cambridge University Press.

Crawley, R.A. & Stevenson, R.J. (1990). Reference in single sentences and in texts. *Journal of Psycholinguistic Research*, 19(3), 191-210.

Croft, W. (1990). *Typology and Universals*. Cambridge: Cambridge University Press.

Cruse, D.A. (1973). Some thoughts on agentivity. *Journal of Linguistics*, 9, 11-23.

Cruse, D.A. (2006). *A Glossary of Semantics and Pragmatics*. Edinburgh: Edinburgh University Press.

Crutch, S.J. & Warrington, E.K. (2006). Partial knowledge of abstract words in patients with cortical degenerative conditions. *Neuropsychology*, 20(4), 482-489.

Crystal, D. (2010). *Evolving English*. London: British Library.

Dabrowska, E. (1998). How metaphor affects grammatical coding: The Saxon genitive in computer manuals. *English Language and Linguistics*, 2(1), 121-127.

Dagan, I. & Itai, A. (1991). A statistical filter for resolving pronoun references. In Y.A. Feldman & A. Bruckstein (Eds.). *Artificial Intelligence and Computer Vision*. Amsterdam: Elsevier.

Dahl, Ö. (2000). Animacy and the notion of semantic gender. In B. Unterbeck & M. Rissanen (Eds.). *Gender in Grammar and Cognition*. Berlin: de Gruyter.

Dahl, Ö. (2008). Animacy and egophoricity: Grammar, ontology and phylogeny. *Lingua*, 118, 141-150.

Dahl, Ö. & Fraurud, K. (1996). Animacy in grammar and discourse. In T. Fretheim & J.K. Gundel (Eds.). *Reference and Referent Accessibility*. Amsterdam: John Benjamins.

Danner, D.D., Snowdon, D.A. & Friesen, W.V. (2001). Positive emotions in early life and longevity: Findings from the Nun Study. *Journal of Personality and Social Psychology*, 80(5), 804-813.

D'Arcais, G.B.F. (1987). Perceptual factors and word order in event descriptions. In G. Kempen (Ed). *Natural Language Generation*. Dordrecht: Martinus Nijhoff.

Davies, J. (2014). *Cracked: Why psychiatry is doing more harm than good*. London: Icon Books.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990-2008+). *International Journal of Corpus Linguistics*, 14, 159-190.

Denison, D., Scott, A. & Börjars, K. (2008). What's wrong with possessive 's? Presentation at ISLE1, University of Freiburg, 8-11 October 2008 [*http://tinyurl.com/DD-UMan*]

De Swart, P., Lamers, M. & Lestrade, S. (2008). Animacy, argument structure, and argument encoding. *Lingua*, 118, 131-140.

Dixon, R.M.W. (1979). Ergativity. *Language*, 55(1), 59-138.

Dogra, T. D., Leenaars, A. A., Raintji, R., Lalwani, S., Girdhar, S.,Wenckstern, S., & Lester, D. (2007). Menstruation and suicide: An exploratory study. *Psychological Reports*, 101, 430–434.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67, 547-619.

Durkheim, E. (1952). *Suicide: A study in sociology*. Translated by J.A. Spaulding and G. Simpson. London: Routledge and Kegan Paul.

Echiverri, A.M., Jaeger, J.J., Chen, J.A., Moore, S.A. & Zoellner, L.A. (2011). 'Dwelling in the past': the role of rumination in the treatment of post-traumatic stress disorder. *Cognitive and Behavioral Practice*, 18(3), 338-349.

Eisenbeiss, S., Matsuo, A. & Sonnenstuhl, I. (2009). Learning to encode possession. In W. McGregor (Ed.). *The Expression of Possession*. Berlin: De Gruyter.

Ellis, D.G. (1996). Coherence patterns in Alzheimer's discourse. *Communications Research*, 23, 472-495.

Engelman, M., Agree, E.M., Meoni, L.A. & Klag, M.J. (2010). Propositional density and cognitive function in later life: Findings from the Precursors Study. *Journal of Gerontology. Series B, Psychological Sciences and Social Sciences,* 65(6), 706-711.

Eslinger, P.J., Damasio, A.R. & Benton, A.L. (1984). *The Iowa Screening Battery for Mental Decline*. Iowa City, Iowa: Department of Neurology, the University of Iowa College of Medicine.

Evans, R. (2002). Refined salience weighting and error analysis in anaphora resolution. In *Proceedings of the Workshop on Reference Resolution and Natural Language Processing (RRNLP '02)*. Alicante, Spain.

Evans, R. & Orăsan, C. (2000). Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of DAARC 2000,* Lancaster, UK, 154-162.

Fabozzi, F.J. (2013). *Encyclopedia of Financial Models,* Volume I. Hoboken, NJ: John Wiley & Sons.

Farias, S.T., Chand, V., Bonnici, L., Baynes, K., Harvey, D., Mungas, D., Simon, C. & Reed, B. (2012). Idea density measured in late life predicts subsequent cognitive trajectories: Implications for the measurement of cognitive reserve. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences,* 67 (6), 677-686.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56 (4), 82-89.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge MA: MIT Press.

Ferguson, C.J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology*, 40(5), 532-538.

Ferguson, G.A. (1976). *Statistical Analysis in Psychology and Education*. 4th edition. New York: McGraw-Hill.

Fernández-Cabana, M. , García-Caballero, A. , Alves-Pérez, M. , García-García, M. , Mateos, R. (2013). Suicidal traits in Marilyn Monroe's fragments: An LIWC analysis. *Crisis*, *34*(2), 124–130.

Fillmore, C. (1968). The case for case. In E. Bach & R.T. Harms (Eds.). *Universals in Linguistic Theory*. New York: Holt, Rinehart & Winston.

Fillmore, C.J. (2003). *Form and Meaning in Language. Volume 1: Papers on Semantic Roles*. Stanford: CSLI Lecture Notes, Number 121.

First, M.B. & Tasman, A. (2004). *DSM-IV-TR: Mental Disorders: diagnosis, etiology and treatment*. Chichester: Wiley.

Fodor, J.A. (1983). *The Modularity of Mind*. Boston, MA: MIT Press.

Foley, W.A. & Van Valin, R.D. (1985). Information packaging in the clause. In T. Shopen (Ed.). *Language Typology and Syntactic Description, Volume 1: Clause Structure*. Cambridge: Cambridge University Press.

Foot, H. & Sanford, A. (2004). The use and abuse of student participants. *The Psychologist*, 17(5), 256-259.

Fornaciari, T. & Poesio, M. (2013). Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*, 21(3), 303-340.

Fraser, K.C., Hirst, G. Graham, N.L., Meltzer, J.A., Black, S.E. & Rochon, E. (2014). Comparison of different feature sets for identification of variants in progressive aphasia. *Proceedings of the First Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, Baltimore, Maryland.

Fraurud, K. (1996). Cognitive ontology and NP form. In T. Fretheim & J.K. Gundel (Eds.). *Reference and Referent Accessibility*. Amsterdam: John Benjamins.

Fung, T.D., Chertkow, H. & Templeman, F.D. (2000). Pattern of semantic memory impairment in dementia of Alzheimer's type. *Brain and Cognition*, 43, 200-205.

Garnham, A. (1987). *Mental Models as Representations of Discourse and Text*. Chichester: Ellis Horwood.

Garnham, A. (2001). *Mental Models and the Interpretation of Anaphora*. Hove: Psychology Press.

Garrard, P. (2009). Cognitive archaeology: Uses, methods and results. *Journal of Neurolinguistics*, 22, 250-265.

Garrard, P., Maloney, L.M., Hodges, J.R. & Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128, 250-260.

Garrard, P. & Elvevåg, B. (2014). Language, computers and cognitive neuroscience. *Cortex*. *http://dx.doi.org/10.1016/j.cortex.2014.02.012*.

Garretson, G., O'Connor, M., Skarabela, B. & Hogan, M. (2004). Coding practices used in the project Optimal Typology of Determiner Phrases. On-line. Boston University. *http://npcorpus.bu.edu/documentation/index.html*.

Gatto, M. (2014). *Web as Corpus*. London: Bloomsbury.

Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate-inanimate distinction as examples. *Cognitive Science*, 14, 79-106.

Gelman, R. & Spelke, E. (1982). The development of thoughts about animate and inanimate objects: Implications for research on social cognition. In J.H. Flavell & L. Ross (Eds.). *Social Cognition Development: Frontiers and possible futures*. Cambridge: Cambridge University Press.

Gernsbacher, M.A. (1997). Two decades of structure building. *Discourse Processes*, 23, 265-304.

Gilquin, G. (2010). Language production: A window to the mind? In H. Götzsche (Ed.) *Memory, Mind and Language*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Glenberg, A.M. & Langston, W.E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language,* 32, 129-151.

Gorenc, K. D., Kleff, F., & Welz, R. (1983). Intentionality and seriousness of suicide attempts in relation to depression. *Boletín de Estudios Médicos y Biológicos*, 32, 233–247.

Gottschalk, L.A., Uliana, R. & Gilbert, R. (1988). Presidential candidates and cognitive impairment measured from behavior in campaign debates. *Public Administration Review*, 48(2), 613-619.

Grafmiller, J. (2014). Variation in English genitives across modality and genres. *English Language and Linguistics*, 18(3), 471-496.

Grafmiller, J. & Shih, S. (2011). New approaches to end weight. *Variation and Typology: New Trends in Syntactic Research*. Helsinki.

Gray, H.M., Gray, K. & Wegner, D.M. (2007). Dimensions of mind perception. *Science*, 315: 619. Supporting online material at www.sciencemag.org/cgi/content/full/315/5812/619/DC1

Grosz, B.J. (1978). Focusing in dialog. In *Proceedings of the 1978 Workshop on Theoretical Issues in Natural Language Processing*. Urbana-Champaign IL, 96-103.

Grosz, B.J., Joshi, A.K. & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of ACL*, 44-50.

Grosz, B. J., Joshi, A.K. & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21 (2), 203-225.

Grosz, B.J., Pollack, M.E. & Sidner, C.L. (1989). Discourse. In M.I. Posner (Ed.). *Foundations of Cognitive Science*. Cambridge MA: MIT Press.

Grosz, B.J. & Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 175-204.

Grüning, A. & Kibrik, A.A. (2005). Modelling referential choice in discourse: A cognitive calculative approach and a neural network approach. In A. Branco, T. McEnery & R. Mitkov (Eds.) *Anaphora Processing: Linguistic, cognitive and computational modelling*. Amsterdam: John Benjamins.

Haegeman, L. (2006). *Thinking Syntactically: A guide to argumentation and analysis*. Oxford: Blackwell.

Hale, S.C. (1988). Spacetime and the abstract/concrete distinction. *Philosophical Studies*, 53, 85-102.

Harnish, S.M. & Neils-Strunjas, J. (2008). In search of meaning: reading and writing in Alzheimer's disease. *Seminars in Speech and Language*, 29, 44-59.

Harrow, M., Adler, D. & Hanf, E. (1974). Abstract and concrete thinking in schizophrenia during the prechronic phases. *Archives of General Psychiatry*, 31(1), 27-33.

Hays, W.L. (1973). *Statistics for the Social Sciences*. 2nd edition. New York: Holt, Rinehart and Winston.

Hinrichs, L. & Szmrecsanyi, B. (2007). Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics*, 11(3), 437-474.

Hobbs, J.R. (1978). Resolving pronoun references. *Lingua*, 44, 311-338.

Hobson, P. (2012). Autism, literal language and concrete thinking: Some developmental considerations. *Metaphor and Symbol*, 27(1), 4-21.

Hodges, J.R. (2007). Overview of frontotemporal dementia. In J.R. Hodges (Ed.) *Frontotemporal Dementia Syndromes*. Cambridge: Cambridge University Press.

Hoffman, P., Lambon Ralph, M.A. & Rogers, T.T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual use of words. *Behavior Research Methods*, 45(3), 718-730.

Hoffman, P., Meteyard, L. & Patterson, K. (2013). Broadly speaking: vocabulary in semantic dementia shifts towards general, semantically diverse words. *Cortex*, *http://dx.doi.org/10.1016/j.cortex.2012.11.004*

Holshausen, K., Harvey, P.D., Elvevåg, B., Foltz, P.W. & Bowie, C.R. (2014). Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. *Cortex*, 55, 88-96.

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association of Literary and Linguistic Computing Bulletin*, 7, 172-177.

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley.

Huang, H-W. & Federmeier, K.D. (2015). Imaginative language: What event-related potentials have revealed about the nature and source of concreteness effects. *Language and Linguistics*, 16(4), 503-515.

Huddleston, R. and Pullum, G.K. (2002). *Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hundt, M. & Szmrecsanyi, B. (2012). Animacy in early New Zealand English. *English World Wide*, 33 (3), 241-263.

Hunt, D. (2013). *Anorexia nervosa, depression and medicalisation: A corpus-based study of patients and professionals*. PhD Thesis, University of Nottingham.

Hye, A. *et al* (2014). Plasma proteins predict conversion to dementia from prodromal disease. *Alzheimer's & Dementia*, published online 8 July 2014. *DOI:10.1016/j.jalz.2014.05.1749*

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLOS Medicine*, 2 (8), e124.

Itagaki, N. & Prideaux, G.D. (1985). Nominal properties as determinants of subject selection. *Lingua*, 66, 135-149.

Jacques, A. & Jackson, G.A. (2000). *Understanding dementia (third edition).* London: Harcourt.

Jäger, G. & Rosenbach, A. (2006). The winner takes it all – almost: Cumulativity in grammatical variation. *Linguistics*, 44(5), 937-991.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L. & Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. *Proceedings of the First Workshop on Computational Linguistics and Clinical Psychology (CLPsych): From Linguistic Signal to Clinical Reality*, 27-37. Baltimore, Maryland.

Jarrold, W.L., Peintner, B., Yeh, E., Krasnow, R., Javitz, H.S. & Swan, G.E. (2010). Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic Alzheimer's disease. *Brain Informatics: Lecture Notes in Computer Science*, 6334, 299-307.

Jashinsky, J., Burton, S.H., Hanson, C.L., West, J., Giraud-Carrier, C., Barnes, M.D. & Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. *Crisis*, 35(1), 51-59.

Jefferies, E., Patterson, K., Jones, R.W. & Lambon Ralph, M.A. (2009). Comprehension of concrete and abstract words in semantic dementia. *Neuropsychology*, 23(4), 492-499.

Ji, H. & Lin, D. (2009). Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*.

Johnson-Laird, P.N. (1983). *Mental Models: Towards a cognitive science of language, inference and consciousness*. Cambridge: Cambridge University Press.

Kehoe, A. & Gee, M. (2012). Reader comments as an aboutness indicator in online texts: Introducing the Birmingham Blog Corpus. In S. Oksefjell Ebeling, J. Ebeling & H. Hasselgård (Eds.). *Studies in Variation, Contacts and Change in English Volume 12: Aspects of Corpus Linguistics: Compilation, Annotation, Analysis*. University of Helsinki. Available at: *http://www.helsinki.fi/varieng/journal/volumes/12/kehoe_gee/*

Keil, F. (1979). *Semantic and Conceptual Development: An ontological perspective*. Cambridge, MA: Harvard University Press.

Keizer, E. (2007). *The English Noun Phrase*. Cambridge: Cambridge University Press.

Keller, F. (2000). *Gradience of Grammar: Experimental and computational aspects of degrees of grammaticality*. PhD Dissertation: University of Edinburgh.

Kemper, S., Thompson, M. & Marquis, J. (2001). Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging*, 16(4), 600-614.

Kibrik, A.A. (1999). Reference and working memory: Cognitive inferences from discourse observations. In K. van Hoek, A.A. Kibrik & L. Noordman (Eds.). *Discourse Studies in Cognitive Linguistics*. Amsterdam: John Benjamins.

Klaiman, M.H. (1991). *Grammatical Voice*. Cambridge: Cambridge University Press.

Knobe, J. (2008). Can a robot, an insect or God be AWARE? *Scientific American Mind*, December 2008/ January 2009.

Koh, S. & Clifton, C. (2002). Resolution of the antecedent of a plural pronoun: ontological categories and predicate symmetry. *Journal of Memory and Language*, 46, 830-844.

Koptjevskaja-Tamm, M. (2001). "A piece of cake" and "a cup of tea": partitive and pseudo-partitive nominal constructions in the Circum-Baltic languages. In Ö. Dahl and M. Koptjevskaja-Tamm (Eds.). *The Circum-Baltic Languages: Typology and Contact*, Volume 2. Amsterdam: John Benjamins, 523-568.

Kotu, V. & Deshpande, B. (2014). *Predictive Analytics and Data Mining: Concepts and practice with RapidMiner*. Waltham, MA: Morgan Kaufmann.

Kousta, S-T., Vigliocco, G., Vinson, D.P., Andrews, M. & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, 140(1), 14-34.

Kroll, J.F. & Merves, J.S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 12(1), 92-107.

Kučera, H. & Francis, W. (1967). *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

Kwong, O.Y. (2013). The potentials and limitations of modelling concept concreteness in computational semantic lexicons with dictionary definitions. *Language Resources and Evaluation*, 47, 1149-1161.

Lake, J.K., Cardy, S. & Humphreys, K.R. (2010). Brief report: Animacy and word order in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 3 March 2010.

Lancashire, I. (2014). Vocabulary and dementia: Six novelists. Poster presented at the *24th Annual Rotman Research Institute Conference, Memory*, 10-12 March *2014.*

Lancashire, I. & Hirst, G. (2009). Vocabulary changes in Agatha Christie's mysteries as an indication of dementia: A case study. Paper and poster presented at the *19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice*, 8-10 March 2009, Toronto.

Landes, S., Leacock, C. & Tengi, R.I. (1998). Building semantic concordances. In C. Fellbaum (Ed.). *WordNet: An Electronic Lexical Database*. Boston, MA: MIT Press.

Langacker, R.W. (1991). *Foundations of Cognitive Grammar, Volume 2: Descriptive Application*. Stanford: Stanford University Press.

Lappin, S. & Leass, H.J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4), 535-561.

Le, X., Lancashire, I., Hirst, G. & Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Literary and Linguistic Computing*, 26(4), 435-461.

Lee, H. (1997). *Virginia Woolf*. London: Vintage.

Leech, G., Francis, B. & Xu, X. (1994). The use of computer corpora in the textual demonstrability of gradience in linguistic categories. In C. Fuchs & B. Victorri (Eds.). *Continuity in Linguistic Semantics*. Amsterdam: John Benjamins.

Leibniz, G.W. (1981 [1765]). *New Essays on Human Understanding*. P. Remnant & J. Bennett (translators and editors). Cambridge: Cambridge University Press.

Lester, D. (2009). Learning about suicide from the diary of Cesare Pavese. *Crisis*, 30, 222–224.

Lester, D., Haines, J., & Williams, C. L. (2010). Content differences in suicide notes by sex, age, and method: A study of Australian suicide notes. *Psychological Reports*, 106, 475–476.

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge MA: MIT Press.

Lezak, M.D., Howieson, D.P. & Loring, D.W. (2004). *Neuropsychological Assessment* (4th edition). New York: Oxford University Press.

Liberman, M. (2008). A correlate of animacy [Blog post]. Retrieved from *languagelog.ldc.upenn.edu/nll/?p=646*.

Lim, K. (2011). The Nun Study: Past, Present and Future. Presentation to the New York Academy of Medicine. Retrieved from: *www.nyam.org/events/2011/docs/Kelvin-Lim-Presentation.ppt*.

Lundin, A. & Hansson, A. (2014). Unemployment and dispensed prescribed antidepressants in Stockholm County. *European Journal of Public Health*, 24(4), 666-668.

Lyketsos, C.G. (2009). Dementia and milder cognitive symptoms. In D.G. Blazer *& D.C.* Steffens *(*Eds.) *Textbook of Geriatric Psychiatry*. Arlington, VA: American Psychiatric Publishing Inc.

Lyons, C. (1999). *Definiteness*. Cambridge: Cambridge University Press.

Lyons, E.J., Mehl, M.R. & Pennebaker, J.W. (2006). Pro-anorexics and recovering anorexics differ in their linguistic Internet self-presentation. *Journal of Psychosomatic Research*, 60, 253-256.

Manning, C.D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Masters, M.C., Morris, J.C. & Roe, C.M. (2015). "Noncognitive" symptoms of early Alzheimer disease: A longitudinal analysis. *Neurology*, published online before print January 14, 2015.

Mayer-Schönberger, V. & Cukier, K. (2013). *Big Data*. London: John Murray.

McDonald, J.L., Bock, J.K. & Kelly, M.H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, 25(2), 188-230.

Medelyan, O., Milne, D., Legg, C. & Witten, I.H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67 (9), 716-754.

Mehl, M.R. & Gill, A.J. (2010). Automatic text analysis. In S.D. Gosling & J.A. Johnson (Eds.). *Advanced Methods for Conducting Online Behavioural Research*. Washington DC: American Psychological Association.

Mihalcea, R. & Liu, H. (2006). A corpus-based approach to finding happiness. In N. Nicolov, F. Salvetti, M. Liberman & J.H. Martin (Eds.). *Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium*. AAAI Press, Menlo Park CA, 145-152. Tech. rep. SS-06-03.

Miller, G.A. (1998). Nouns in WordNet. In C. Fellbaum (Ed.). *WordNet: An electronic lexical database*. Cambridge MA: MIT Press.

Mitkov, R. (2000). Pronoun resolution: The practical alternative. In S. Botley & A.M. McEnery (Eds.). *Corpus-based and Computational Approaches to Discourse Anaphora*. Amsterdam: John Benjamins.

Mitkov, R. (2002). *Anaphora Resolution*. London: Longman.

Mitkov, R. (2003). Anaphora resolution. In R. Mitkov (Ed.) *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Mitkov, R., Evans, R. & Orăsan, C. (2002). A new, fully-automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of CICLing 2002*, 168-186.

Modjeska, N.N., Markert, K. & Nissim, M. (2003). Using the Web in machine learning for other-anaphora resolution. In R. Dale, K. van Deemter & R. Mitkov (Eds.). *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*. Budapest.

Morgan, J. (1984). *Agatha Christie: A biography*. London: HarperCollins.

Moss, H.E. & Gaskell, M.G. (1999). Lexical semantic processing during speech comprehension. In S. Garrod & M.J. Pickering (Eds.). *Language Processing*. Hove: Psychology Press.

Murdoch, B.E. (1990). *Acquired Speech and Language Disorders: A neuroanatomical and functional neurological approach*. London: Chapman and Hall.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys,* 41(2), 1-69.

Neuman, Y., Cohen, Y., Assaf, D. & Kedma, G. (2012). Proactive screening for depression through metaphorical and automatic text analysis. *Artificial Intelligence in Medicine,* 56, 19-25.

Newman, M.L., Groom, C. J., Handelman, L. D., & Pennebaker, J.W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes,* 45, 211–236.

Ng, V. (2002). Machine learning for coreference resolution: Recent successes and future challenges. *Technical Report CUL CIS/TR 2003-1918*. Cornell University.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala Sweden, 11-16 July 2010, 1396-1411.

Nøklestad, A. (2009). *A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection*. PhD Thesis, University of Oslo.

Nolen-Hoeksema, S. (1991). Responses to depression and their effects on the duration of depressive episodes. *Journal of Abnormal Psychology*, 100(4), 569-582.

Nolen-Hoeksema, S. (1996). Chewing the cud and other ruminations. In R.S. Wyrer Jr (Ed.). *Ruminative Thoughts*. Mahwah, NJ: Erlbaum.

Nordquist, D. (2004). Comparing elicited data and corpora. In M. Achard & S. Kemmer (Eds.) *Language, Culture, and Mind*. Stanford: CSLI Publications.

Oakes, M.P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Oakhill, J., Garnham, A. & Vonk, W. (1989). The on-line construction of discourse models. *Language and Cognitive Processes*, 4(3/4), 263-286.

Och, F.J. & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics,* 30(4), 417-449.

O'Connor, C., Maling, J. & Skarabela, B. (2013). Nominal categories and the expression of possession. In K. Börjars, D. Denison & A. Scott (Eds.). *Morphosyntactic Categories and the Expression of Possession*. Amsterdam: John Benjamins.

Oh, J., Chun, J.W., Lee, J.S. & Kim, J-J. (2014). Relationship between abstract thinking and eye gaze pattern in patients with schizophrenia. *Behavioral and Brain Functions*, 10(13).

Ohm, T.G., Müller, H., Braak, H. & Bohl, J. (1995). Close-meshed prevalence rates of different stages as a tool to uncover the rate of Alzheimer's disease-related neurofibrillary changes. *Neuroscience,* 64(1), 209-217.

Orăsan, C. & Evans, R. (2001). Learning to identify animate references. In *Proceedings of the Workshop on Computational Natural Language Learning (CONLL-2001,* 129-136. Toulouse, France.

Orăsan, C. & Evans, R. (2007). NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research,* 29, 79-103.

Osgood, C.E. (1971). Where do sentences come from? In D. Steinberg & L. Jacobovits (Eds.) *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*. Cambridge: Cambridge University Press.

Osgood, C.E. & Bock, J.K. (1977). Salience and sentencing: Some production principles. In S. Rosenberg (Ed.). *Sentence Production: Developments in research and theory*. Hillsdale NJ: Erlbaum.

Oxman, T.E., Rosenberg, S.D., Schnurr, P.P. & Tucker, G.J. (1988). Diagnostic classification through content analysis of patients' speech. *American Journal of Psychiatry*, 145 (4), 464-468.

Paivio, A. (1971). *Imagery and Verbal Processes*. New York: Holt, Rinehart and Winston.

Paivio, A. (1986). *Mental Representations: A dual coding approach*. New York: Oxford University Press.

Paivio, A. (2007). *Mind and its Evolution: A dual-coding theoretical approach*. Mahwah, NJ: Erlbaum.

Paivio, A., Yuille, J.C. & Madigan, S.A. (1968). Concreteness, imagery and meaningfulness of 925 nouns. *Journal of Experimental Psychology* – Monograph Supplement, 76 (No.1, Part 2), January 1968, 1-25.

Pajer, K., Andrus, B.M., Gardner, W., Lourie, A., Strange, B. et al (2012). Discovery of blood transcriptomic markers for depression in animal models and pilot validation in subjects with early-onset major depression. *Translational Psychiatry*, 2, e101, doi:10.1038/tp.2012.26.

Pakhomov, S., Chacon, D., Wicklund, M. & Gundel, J. (2011). Computerized assessment of syntactic complexity in Alzheimer's disease: A case study of Iris Murdoch's writing. *Behavior Research*, 43, 136-144.

Pakhomov, S.V.S. & Hemmy, L.S. (2013). A computational linguistic measure of clustering behaviour on semantic verbal fluency task predicts risk of future dementia in the Nun Study. *Cortex*, *http://dx.doi.org/10.1016/j.cortex.2013.05.009*

Patton, P. (2008). One world, many minds: Intelligence in the animal kingdom. *Scientific American Mind*, December 2008.

Payne, J. & Huddleston, R. (2002). Nouns and noun phrases. In R. Huddleston & G.K. Pullum (eds.) *The Cambridge Grammar of the English Language*, 323-523. Cambridge: Cambridge University Press.

Pearson, J., Poesio, M. & Stevenson, R. (2001). The effects of animacy, thematic role, and surface position on the focusing of entities in discourse. Unpublished draft paper prepared for the *First Workshop on Cognitively Plausible Models of Semantic Processing*, Edinburgh, July 2001.

Pennebaker, J.W. (2011). *The Secret Life of Pronouns: What our words say about us*. New York: Bloomsbury Press.

Pennebaker, J.W., Francis, M.E., & Booth, R. J. (2007). *Linguistic Inquiry and Word Count (LIWC): LIWC2007.* Mahwah, NJ: Erlbaum.

Pennebaker, J.W. & King, L.A. (1999). Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6), 1296-1312.

Pennebaker, J.W., & Stone, L. D. (2003).Words of wisdom: Language use over the lifespan. *Journal of Personality and Social Psychology*, 85, 291–301.

Pestian, J.P., Matykiewicz, P., Grupp-Phelan, J., Lavanier, S.A., Combs, J. & Kowatch, R. (2008). Using natural language processing to classify suicide notes. *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, 96-97. Columbus, Ohio, June 2008.

Pestian, J.P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., … Brew, C. (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights,* 5 (Suppl. 1), 3-11.

Pickering, M.J. & Branigan, H. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, 39, 633-651.

Pinker, S. (2012). *The Better Angels of our Nature: A history of violence and humanity*. London: Penguin Books.

Plath, S. (1963). *The Bell Jar*. London: Faber and Faber.

Plath, S. (1975) *Letters Home: Correspondence, 1950-1963*. Selected and edited with commentary by Aurelia Schober Plath. London : Faber and Faber.

Plath, S. (2000). *The Journals of Sylvia Plath*. K.V. Kukil (Ed.). London: Faber and Faber.

Platt, J.C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *Microsoft Research: Technical Report MSR-TR-98-14*.

Plaut, D. C. and Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, 10, 377-500

Poesio, M., Cheng, H., Henschel, R., Hitzeman, J., Kibble, R. & Stevenson, R. (2000). Specifying the parameters of Centering Theory: A corpus-based evaluation using text from application-oriented domains. In *Proceedings of the 38th Annual Meeting of the ACL*.

Poesio, M., Mehta, R., Maroudas, A. & Hitzeman, J. (2004). Learning to resolve bridging references. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, 143-150.

Poesio, M., Stevenson, R., Di Eugenio, B. & Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30 (3), 309-363.

Prat-Sala, M. & Branigan, H.P. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language*, 42, 168-182.

Preacher, K.J. (2001, April). Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence [Computer software]. Available from *http://quantpsy.org*

Protopescu, C., Raffi, F., Brunet-François, C., Salmon, D., Verdon, R. and others (2012). Incidence, medical and socio-behavioural predictors of psychiatric events in an 11-year follow-up of HIV-infected patients on antiretroviral therapy. *Antiviral Therapy*, 17, 1079-1083.

Pulvermüller, F., Pye, E., Dine, C., Hauk, O., Nestor, P. & Patterson, K. (2008). Word category deficits in semantic dementia. *Paper presented at the Cognitive Neuroscience Society 2008 Annual Meeting*.

Quinlan, R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Radford, A. (2004). *Minimalist Syntax: Exploring the structure of English*. Cambridge: Cambridge University Press.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3), 385-401.

Rasinger, S.M. (2008). *Quantitative Research in Linguistics*. London: Bloomsbury Press.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519-549.

Reilly, J., Troche, J. & Grossman, M. (2014). Language processing in dementia. In A.E. Budson & N.W. Kowall (Eds.) *Handbook of Alzheimer's Disease and Other Dementias*. Chichester: Blackwell.

Reinhart, T. (1981). Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, 27(1), 53-94.

Reips, U-D. (2002). Standards for internet-based experimenting. *Experimental Psychology,* 49(4), 243-256.

Reips, U-D. (2007). The methodology of internet-based experiments. In A.N. Joinson, K.Y.A. McKenna, T. Postmes & U-D. Reips (Eds.). *Oxford Handbook of Internet Psychology*. Oxford: Oxford University Press.

Riley, K.P., Snowdon, D.A., Desrosiers, M.F., & Markesbery, W.R. (2005). Early life linguistic ability, late life cognitive function, and neuropathology: findings from the Nun Study. *Neurobiology of Aging*, 26(3), 341-347.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N. & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013).*

Rissenberg, M. & Glanzer, M. (1987). Free recall and word finding ability in normal aging and senile dementia of the Alzheimer's type: The effect of item concreteness. *Journal of Gerontology*, 42(3), 318-322.

Rosen, G. (2014). Abstract objects. In E.N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2014 Edition).
URL *http://plato.stanford.edu/fall2014/entries/abstract-objects*

Rosenbach, A. (2002). *Genitive Variation in English: Conceptual factors in synchronic and diachronic studies*. Berlin: Mouton de Gruyter.

Rosenbach, A. (2003). Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In G. Rohdenburg & B. Mondorf (Eds.). *Determinants of Grammatical Variation in English*. Berlin: Mouton de Gruyter.

Rosenbach, A. (2006). Descriptive genitives in English: A case study on constructional gradience. *English Language and Linguistics*, 10(1), 77-118.

Rosenbach, A. (2008). Animacy and grammatical variation – findings from English genitive variation. *Lingua*, 118, 151-171.

Rosenbach, A. (2014). English genitive variation – the state of the art. *English Language and Linguistics*, 18(2), 215-262.

Rude, S.S., Gortner, E-M. & Pennebaker, J.W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8), 1121-1133.

Rutkowski, P. (2007). The syntactic structure of grammaticalized partitives (pseudo-partitives). *University of Pennsylvania Working Papers in Linguistics*, 13(1), 337-350.

Sanford, A.J. & Garrod, S.C. (1981). *Understanding Written Language: Explorations in comprehension beyond the sentence*. Chichester: Wiley.

Sanford, A.J. & Garrod, S.C. (1998). The role of scenario mapping in text comprehension. *Discourse Processes*, 26, 159-190.

Sanford, A.J., Moar, K. & Garrod, S.C. (1988). Proper names as controllers of discourse focus. *Language and Speech*, 31(1), 43-56.

Sapsford, R. & Jupp, V. (1996). *Data Collection and Analysis*. London: Sage.

Schaich, A., Watkins, E. & Ehring, T. (2011). Can concreteness training buffer against the negative effects of rumination on PTSD? An experimental analogue study. *Journal of Behavior Therapy and Experimental Psychiatry*, 44(4), 396-403.

Schler, J., Koppel, M., Argamon, S. & Pennebaker, J. (2006). Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Schwanenflugel, P. (1991). Why are abstract concepts hard to understand? In P. Schwanenflugel (Ed.) *The Psychology of Word Meanings*. Hillsdale, NJ: Erlbaum.

Schwanenflugel, P. J., Akin, C., & Luh, W. (1992). Context availability and the recall of abstract and concrete words. *Memory & Cognition*, 20(1), 96-104.

Seoane Posse, E. (1999). Inherent topicality and object foregrounding in Early Modern English. *ICAME Journal*, 23, 117-140.

Shea, S.C. (2011). *The Practical Art of Suicide Assessment*. Stoddard, NH: Mental Health Presses.

Sidner, C.L. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. PhD Thesis. Cambridge MA: Massachusetts Institute of Technology.

Sidner, C.L. (1983). Focusing in the comprehension of definite anaphora. In M. Brady & R. Berwick (Eds.). *Computational Models of Discourse*. Cambridge MA: MIT Press.

Siewierska, A. (2004). *Person*. Cambridge: Cambridge University Press.

Simpson, P. & Mayr, A. (2010). *Language and Power*. Abingdon: Routledge.

Smith, S.W., Jucker, A.H., & Müller, S. (2000). "Some artist guy": the role of salience and common ground in the formulation of referring expressions in conversational narratives. In *Papers from the 7th International Pragmatics Conference,* 528-542.

Snowdon, D.A. (2001). *Aging with Grace: The Nun Study and the science of old age*. London: Fourth Estate.

Snowdon, D.A., Greiner, L.H., Kemper, S.J., Nanakakkara, N. & Mortimer, J.A. (1999). Linguistic ability in early life and longevity: Findings from the Nun Study. In J-M. Robine, B. Forette, C. Francheschi & M. Allard (Eds.). *The Paradoxes of Longevity*. Berlin: Springer Verlag.

Snowdon, D.A., Greiner, L.H. & Markesbery, W.R. (2000). Linguistic ability in early life and the neuropathology of Alzheimer's disease and cerebrovascular disease: Findings from the Nun Study. *Annals of the New York Academy of Sciences*, 903, 34-38.

Snowdon, D.A., Kemper, S.J., Mortimer, J.A., Greiner, L.H., Wekstein, D.R. & Markesbery, W.R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Journal of the American Medical Association*, 275(7), 528-532.

Soon, W.M., Ng, H.T. & Lim, D.C.Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4), 521-544.

Spokas, M., Wenzel, A., Brown, G. K., & Beck, A.T. (2012). Characteristics of individuals who make impulsive suicide attempts. *Journal of Affective Disorders,* 136, 1121–1125.

Sridhar, S.N. (1988). *Cognition and Sentence Production*. New York: Springer-Verlag.

Sridharan, S. & Murphy, B. (2012). Modeling word meaning: Distributional semantics and the corpus quality-quantity trade-off. In *Proceedings of the 3$^{rd}$ Workshop on Cognitive Aspects of the Lexicon (Cog-Alex-III),* COLING 2012, Mumbai, December 2012, 53-68.

Stevenson, A. (1989). *Bitter Fame: A Life of Sylvia Plath*. London: Viking.

Stevenson, R.J. (1996). Mental models, propositions, and the comprehension of pronouns. In J. Oakhill & A. Garnham (Eds.) *Mental Models in Cognitive Science: Essays in honour of Phil Johnson-Laird*. Hove: Psychology Press.

Stevenson, R.J. (2002). The role of salience in the production of referring expressions: A psycholinguistic perspective. In K. van Deemter & R. Kibble (Eds.). *Information Sharing*. Stanford: CSLI.

Stevenson, R.J., Crawley, R.A. & Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4), 519-548.

Stirman, S.W. & Pennebaker, J.W. (2001). Word use in the poetry of suicidal and non-suicidal poets. *Psychosomatic Medicine*, 63, 517-522.

Stöber, J. (1998). Worry, problem elaboration and suppression of imagery: The role of concreteness. *Behaviour Research and Therapy*, 36, 751-756.

Stöber, J. & Borkovec, T.D. (2002). Reduced concreteness of worry in generalized anxiety disorder: Findings from a therapy study. *Cognitive Therapy and Research*, 26(1), 89-96.

Stöber, J., Tepperwien, S. & Staak, M. (2000). Worrying leads to reduced concreteness of problem elaborations: Evidence for the avoidance theory of worry. *Anxiety, Stress and Coping*, 13, 217-227.

Strain, E., Patterson, K., Graham, N. & Hodges, J.R. (1998). Word reading in Alzheimer's disease: cross-sectional and longitudinal analyses of response time and accuracy data. *Neuropsychologia*, 36, 155-171.

Strang, K.D. (2015). Articulating a research design ideology. In K.D. Strang (Ed.). *The Palgrave Handbook of Research Design in Business and Management*. New York: Palgrave Macmillan.

Strube, M. (1998). Never look back: An alternative to centering. *Paper cmp-lg/9806018 in the Computing Research Repository (CRR)*.

Strube, M. & Hahn, U. (1996). Functional centering. In *Proceedings of the 34th annual meeting of the Association for Computational Linguistics*.

Strube, M. & Ponzetto, S.P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21$^{st}$ National Conference on Artificial Intelligence,* AAI Press, 1419-1420.

Strube, M., Rapp, S. & Müller, C. (2002). The influence of minimum edit distance on reference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Philadelphia, July 2002, 312-319.

Surowiecki, J. (2005). *The Wisdom of Crowds*. New York: Anchor Books.

Szmrecsanyi, B., Ehret, K. & Wolk, C. (2014). Quirky quadratures: on rhythm and weight as constraints on genitive variation in an unconventional data set. *English Language and Linguistics*, 18(2), 263-304.

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2, doi: 10.1146/annurev-linguistics-011415-040518.

Tagamets, M.A., Cortes, C.R., Griego, J.A. & Elvevåg, B. (2014). Neural correlates of the relationship between discourse coherence and sensory monitoring in schizophrenia. *Cortex*, 55, 77-87.

Thomas, C., Kešelj, V., Cercone, N., Rockwood, K. & Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *Mechatronics and Automation,* 2005 IEEE International Conference, 3, 1569-1574. Niagara Falls, Ontario, Canada.

Thompson, L. (2007). *Agatha Christie: An English mystery*. London: Headline Publishing.

Traxler, M.J., Williams, R.S., Blozis, S.A. & Morris, R.K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language*, 53, 204-224.

Tunmer, W.E. (1985). The acquisition of the sentient-nonsentient distinction and its relationship to causal reasoning and social cognition. *Child Development*, 56, 989-1000.

Van Casteren, M. & Davis, M.H. (2007). Match: A program to assist in matching the conditions of factorial experiments. *Behavior Research Methods*, 39(4), 973-978.

Van Deemter, K., Gatt, A., Van Gompel, R.P.G. Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4, 166-183.

Van Dijk, T.A. & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Orlando: Academic Press.

Van Velzen, M. & Garrard, P. (2008). From hindsight to insight – retrospective analysis of language written by a renowned Alzheimer's patient. *Interdisciplinary Science Reviews*, 33(4), 278-286.

Van Velzen, M.H., Nanetti, L. & de Deyn, P.P. (2014). Data modelling in corpus linguistics: How low may we go? *Cortex*, 55, 192-201.

Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X. & Moschitti, A. (2008). BART: A modular toolkit for coreference resolution. In *Proceedings of the ACL-08: HLT Demo Session (Companion Volume)*, 9-12.

Von Studnitz, R.E. & Green, D.W. (2002).  The cost of switching language in a semantic categorization task. *Bilingualism: Language and Cognition*, 5(3), 241-251.

Walker, M.A. (1989). Evaluating discourse processing algorithms. *Proceedings of the 27th Annual Meeting of the ACL*, 26-29 June 1989, Vancouver BC, Canada.

Walker, M.A., Joshi, A.K. & Prince, E.F. (1998). Centering in naturally occurring discourse: An overview. In M.A. Walker, A.K. Joshi & E.F. Prince (Eds.). *Centering Theory in Discourse*. Oxford: Clarendon Press.

Warrington, E.K. & McCarthy, R.A. (1987). Categories of knowledge: Further fractionations and an attempted integration. *Brain*, 110, 1273-1296.

Warrington, E.K. & Shallice, T. (1984). Category-specific semantic impairments. *Brain*, 107, 829-854.

Watkins, E.R., Baeyens, C.B. & Read, R. (2009). Concreteness training reduces dysphoria: Proof-of-principle for repeated cognitive bias modification of depression. *Journal of Abnormal Psychology*, 118(1), 55-64.

Watkins, E., Moberly, N.J. & Moulds, M.L. (2008). Processing mode causally influences emotional reactivity: distinct effects of abstract versus concrete construal on emotional response. *Emotion*, 8(3), 364-378.

Watkins, E. & Moulds, M. (2005). Distinct modes of ruminative self-focus: impact of abstract versus concrete rumination on problem solving in depression. *Emotion*, 5(3), 319-328.

Watkins, E. & Moulds, M. (2007). Reduced concreteness of rumination in depression: A pilot study. *Personality and Individual Differences*, 43, 1386-1395.

Watkins, E., Taylor, R.S., Byng, R., Baeyens, C., Read, R., Pearson, K. & Watson, L. (2012). Guided self-help concreteness training as an intervention for major depression in primary care: a phase II randomized controlled trial. *Psychological Medicine*, 42(7), 1359-1371.

Watkins, E. & Teasdale, J.D. (2001). Rumination and overgeneral memory in depression: Effects of self-focus and analytic thinking. *Journal of Abnormal Psychology*, *110*(2), 353-357.

Watkins, E., Teasdale, J.D. & Williams, R.M. (2000). Decentring and distraction reduce overgeneral autobiographical memory in depression. *Psychological Medicine*, 30, 911-920.

Weckerly, J. & Kutas, M. (1999). An electrophysiological analysis of animacy effects in the processing of object relative sentences. *Psychophysiology,* 36, 559-570.

Williams, K., Holmes, F., Kemper, S. & Marquis, J. (2003). Written language clues to cognitive changes of aging: An analysis of the letters of King James VI/I. *Journal of Gerontology: Psychological Sciences*, 58B(1), 42-44.

Winograd, T. (1972). *Understanding Natural Language*. New York: Academic Press.

Witten, I.H., Frank, E. & Hall, M.A. (2011). *Data Mining: Practical machine learning tools and techniques*. Burlington MA: Elsevier.

Wolk, C., Bresnan, J., Rosenbach, A. & Szmrecsanyi, B. (2013). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*, 30(3), 382-419.

Woolf, V. (1977) *A Change of Perspective: The Letters of Virginia Woolf.* Volume 3, 1923-1928. N. Nicholson (Editor). London: Hogarth Press.

Woolf, V. (1980) *Leave the Letters Till We're Dead: The Letters of Virginia Woolf*. Volume 6, 1936-1941. N. Nicholson (Editor). London: Hogarth Press.

Wundt, W. (1970). The psychology of the sentence. In A.L. Blumenthal (Ed.). *Language and Psychology: Historical aspects of psycholinguistics*. New York: Wiley. (First published 1900).

Yamamoto, M. (1999). *Animacy and Reference*. Amsterdam: John Benjamins.

Yamamoto, M. (2006). *Agency and Impersonality*. Amsterdam: John Benjamins.

Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M.C. & Wasow, T. (2004). Animacy encoding in English: why and how. *Workshop on Discourse Annotation. 42nd Annual Meeting of the ACL*. Barcelona, Spain, 118-125.

Zec, R.F. (1993). Neuropsychological functioning in Alzheimer's disease. In R.W. Parks, R.F. Zec and R.S. Wilson (Eds.) *Neuropsychology of Alzheimer's Disease and Other Dementias*. New York: Oxford University Press.

## APPENDIX 4.1: Animyser - Outline Specification

**Software.** The program has been written in Python 2.7, within a Windows 7 environment.

**Imported objects**. The Animyser program imports the master-package and several optional objects from the Pattern 2.6 data mining software (De Smedt and Daelemans, 2012). The Wikipedia object from pattern.web incorporates Wikipedia's own API and provides the phrase-search and results-counts functions that are core to the program. From pattern.en, two functions are imported: a POS (part-of-speech) tagger and a function that converts plural nouns to their singular form.

**POS Tagger**. If selected, the part of speech (POS) tagger annotates all the words in a text with their grammatical category, thus enabling Animyser to identify all of the nouns, tagged as singular, or plural, or proper nouns. Proper nouns are excluded: only singular common nouns are analysed. Whilst the imported tagger is generally very reliable, it incorrectly classifies indefinite pronouns (termed 'compound determinatives' by Huddleston and Pullum, 2002: 423) as nouns. Examples are *everyone, something* and *nobody*. The Animyser program rectifies this error by a work-around referral to a tuple of indefinite pronouns that changes their tag to "PRN" (pronoun), and thus excludes them from analysis.

**Singularize**. Measurement of the genitive ratio is based on the singular form of the noun. Plural nouns must therefore be 'singularized'. The singularize function in Pattern 2.6 is reliable for regular and common irregular plurals, (e.g. *children → child*, *women → woman*), but not for the classes of 'plural nouns' and 'singular plurals' that have been identified as confounds (see section 4.10): the singularize function would convert *billiards* and *alms* (for example) to *billiard* and *alm*. Nouns in these two classes are contained in two separate tuples, for reference by the program. A dictionary of low-frequency irregular plurals (e.g. *alumni* : *alumnus*) provides additional data to supplement the singularize function in Pattern.

These relatively minor modifications to Pattern's singularize function work well in combination with the POS tagger. However, when an input method does not involve the tagger, the singularize function relies upon a limited set of simplistic rules that are much less reliable, in that they incorrectly infer plural endings. These are the rules in Pattern and their consequences:

[1] Always remove a final 's':

sadness = sadnes

foetus = foetu

diagnosis = diagnosi

[2] Any '-ia' ending becomes '-um' (as in Latin):

amnesia = amnesium

dementia = dementium

[3] Any '-ice' ending becomes '-ouse' (based on *mice* and *lice*):

police = polouse

chalice = chalouse

[4] Any '-our' ending becomes '-my':

flour = flmy

demeanour = demeanmy

The Animyser program incorporates a new set of seven rules to compensate for these 'false plurals', together with a dictionary of exceptions.

**Exceptions**. In addition to the indefinite pronouns and irregular plurals, the program also incorporates reference lists of temporal and exceptional nouns (see section 4.10).

**User options**. The user is informed that:

"This program will configure a list of nouns into a series of phrase searches that are submitted to Wikipedia. The program will return an output file. You must choose:

1. The method of input – file or manual.
2. The name of the output file.
3. The rating required (AR or CR)."

**Input method**. Three input methods are offered to the user:

Option 1. The program will retrieve a designated text (.txt) file as specified by the user, and extract all of the nouns identified within that text.

Option 2. The noun_list will be constructed manually, by the user entering one or more singular nouns, followed by "qq" to signify the end of the list.

Option 3. The program will import a list of nouns from a text file that is specified by the user.

The product of all three options is a list of nouns (noun_list) for processing.

Words in the noun_list are converted to a standard format of lower case, singular form, and with the appropriate indefinite article assigned.

**Phrase-searches**. The program will automatically select all six phrase-searches:

| | |
|---|---|
| OD | "of the cat" |
| DS | "the cat's" |

| | |
|---|---|
| OI | "of a cat" |
| IS | "a cat's" |
| ON | "of cat" |
| NS | "cat's" |

The user is then asked to name their output file, which must be located in the same directory as the program, in a .txt format that can be opened with Windows Notepad.

**Indefinite article**. Two of the six phrase-searches contain an indefinite article. The program must therefore assign the appropriate indefinite article (*a* or *an*) to the target noun. The regular rule in English is simple: the default value of the indefinite article is *a*, with the variant *an* preceding nouns with an initial vowel. The program interprets the regular rule and the exceptions to that rule, allowing for easy adaptation should further exceptions be identified.

There are two classes of irregular usage, both based upon the phonetic rather than the lexical form of the noun. The nouns *honour* and *hour* both carry a silent *h* and therefore take the *an* form of the indefinite article, whereas other words beginning with *ho-* (e.g. *house*, *hound*) follow the basic rule.

The second class of 'irregular indefinites' are nouns beginning with the phonetic form **ju**:. There are three vowel-initial lexical representations of this phonetic form: *uni-* (*uniform*, *university*), *eu-* (*euphemism*, *euphoria*) and *ew-* (*ewe, ewer*), all of which take the default *a* form of the indefinite article.

**Phrase-searches**. By reference to a dictionary (named prior_hits), the program checks if the noun has already been encountered in the current run. If so, the results of the previous phrase searches are added to the output file. Otherwise, the program constructs a search phrase for each noun in the noun_list, in each of the search patterns selected by the user (or by default).

**Post-processing**. No-score and high-score results are dealt with as special cases (see section 4.10).

**Rating calculation**. The rating (AR or CR) requested by the user is calculated by the program, taking account of exceptions and special cases.

**Output**. The results-count of each phrase-search submitted to Wikipedia is written to the output file (if specified): each noun followed by the results-counts and the requested rating.

**Import to Excel**. The output is easily imported as a text file (comma-separated) to an Excel worksheet.

## APPENDIX 5.1: Sentence Production Materials

**Key:**

**P** Pilot experiments

**M** Main experiment

### Names v Names

| | | | |
|---|---|---|---|
| Adam | Amnesty International | P | |
| Billy | Aston Villa | | M |
| Bob | BBC | | M |
| Adam | Cyprus | P | M |
| Billy | London | | M |
| Amnesty International | Cyprus | P | |
| Aston Villa | London | | M |
| BBC | India | | M |

### Non-Names v Names

| | | | |
|---|---|---|---|
| husband | Jack | P | M |
| wife | Michael | | M |
| god | France | P | |
| girl | Egypt | | M |
| baby | York | | M |
| politician | CIA | P | M |
| dog | Co-operative Society | | M |
| hotel | Labour Council | P | |
| town | BBC | | M |
| weapon | Labour Party | | M |
| hair | Bob | P | |
| ball | Adam | | M |
| book | Arthur | | M |
| tree | Africa | P | M |
| palace | Cyprus | | M |

**Non-Names v Non-Names**

| | | | |
|---|---|---|---|
| husband | machine | P | |
| teacher | building | | M |
| boy | car | | M |
| sheep | bridge | P | |
| husband | kitchen | | M |
| girl | book | | M |
| doctor | window | P | M |
| wife | road | | M |
| politician | territory | P | M |
| lord | tree | | M |
| park | beach | P | |
| dog | seat | | M |
| horse | fire | | M |
| horse | floor | P | |
| son | view | | M |
| person | action | | M |
| fish | fire | P | |
| building | village | | M |
| town | wall | | M |
| hotel | value | P | M |
| paper | table | | M |
| palace | border | P | |
| glass | floor | | M |
| room | word | | M |

## APPENDIX 5.2: Sentence Production Experiment - Script

1.1

**THANK YOU FOR YOUR INTEREST IN THIS RESEARCH**

To take part in this experiment, you

Need to be at least 18 (this is for ethical reasons)

and you

Need to have been born in the UK with British English as your first language (this is for consistency).

Please confirm that you meet these criteria and wish to

○ continue (1)

○ Or exit (2)

2.1

**A CHARITY WILL BENEFIT**

There is no payment for taking part, but £1 will be donated to the charity Kids Company, up to a maximum of £200, for every completed response.

Kids Company provides "practical, emotional and educational support to vulnerable inner-city children".

www.kidsco.org.uk/

**QUESTIONS**

If you have any questions or concerns about this study, please contact

Kevin Glover    kjglov@essex.ac.uk

2.2

**PLEASE READ THIS: INFORMATION FOR PARTICIPANTS**

1. This research study will test your response to particular words.
2. Participation is completely anonymous. No personal data will identify you.
3. The experiment has ethical approval. There are no known risks or discomforts.
4. You may withdraw from the experiment at any time.
5. By completing the questions and submitting your response you consent to taking part.

2.3

Please confirm that you have read and understood the Information for Participants, that you agree to take part in this experiment, and have not taken part previously.

○ Yes, I agree to take part and I have not taken part previously (1)

○ No, I do not wish to take part (2)

3.1

This experiment tests your reactions to particular words.

You will see pairs of nouns and/or names.

Your task is to think of a short sentence that contains both words, in any order. You simply have to click on which of the two words comes first in your sentence.

So, presented with the nouns FISH and COST, you might think of this sentence: *The cost of fish is high*.

You would click on COST, because that came first in your sentence.

3.2

Presented with the noun BABY and the name ANNE, you might think of:

*Anne smiled at the baby*

and click on ANNE

Plurals are OK. So, presented with MOUSE and DESK, you might think of:

*The mice ran under the desk*

and click on MOUSE

Try to make each sentence 'natural' - something you might hear or read.

Keep it simple and don't take too long - first thoughts are usually best.

3.3

Here are two examples for practice.

Think of a short sentence that contains both these words. Which word comes first in your sentence?

○ baby (1)

○ Labour Party (2)

3.4

Think of a short sentence that contains both these words. Which word comes first in your sentence?

❍ BBC (1)

❍ victim (2)

3.5

Please type here the sentence you just thought of

3.6

Timing

3.7

**THIS IS VERY IMPORTANT**

Before you answer you must have a sentence clearly in your mind, because you will sometimes be asked to type it out.

3.8

This experiment is in two parts. Each part should take only a few minutes to complete. The first part will begin when you click on **Next**.

4.1

Which of these comes first in your sentence?

○ Jack (1)

○ husband (2)

4.2

Please type here the sentence you thought of

4.3 Timing

5.1

Which of these comes first in your sentence?

○ teacher (1)

○ building (2)

6.1

Which of these comes first in your sentence?

○ boy (1)

○ car (2)

7.1

Which of these comes first in your sentence?

○ husband (2)

○ kitchen (3)

8.1

Which of these comes first in your sentence?

○ Billy (1)

○ Aston Villa (2)

9.1

Which of these comes first in your sentence?

○ Michael (1)

○ wife (2)

10.1

Which of these comes first in your sentence?

○ wife (1)

○ road (2)

10.2

Please type here the sentence you thought of

10.3

Timing

11.1

Which of these comes first in your sentence?

❍ York (1)

❍ baby (2)

12.1

Which of these comes first in your sentence?

❍ girl (1)

❍ book (2)

13.1

Which of these comes first in your sentence?

❍ doctor (1)

❍ window (2)

14.1

Which of these comes first in your sentence?

❍ Bob (1)

❍ BBC (2)

15.1

Which of these comes first in your sentence?

❍ Egypt (1)

❍ girl (2)

16.1

**THANKS**

That's the end of the first part.

Please click on **Next** when you are ready to go to the second part.

17.1

Which of these comes first in your sentence?

○ lord (1)

○ tree (2)

17.2

Please type here the sentence you thought of

17.3

Timing

18.1

Which of these comes first in your sentence?

○ Co-op (1)

○ dog (2)

19.1

Which of these comes first in your sentence?

○ CIA (1)

○ politician (2)

20.1

Which of these comes first in your sentence?

○ politician (1)

○ territory (2)

21.1

Which of these comes first in your sentence?

○ dog (1)

○ seat (2)

22.1

Which of these comes first in your sentence?

○ Adam (1)

○ Cyprus (2)

23.1

Which of these comes first in your sentence?

○ Labour Party (1)

○ weapon (2)

23.2

Please type here the sentence you thought of

23.3

Timing

24.1

Which of these comes first in your sentence?

○ Billy (1)

○ London (2)

25.1

Which of these comes first in your sentence?

○ person (1)

○ action (2)

26.1

Which of these comes first in your sentence?

○ horse (1)

○ fire (2)

27.1

Which of these comes first in your sentence?

○ BBC (1)

○ town (2)

28.1

Which of these comes first in your sentence?

○ son (1)

○ view (2)

29.1

Thanks!

**JUST A FEW MORE QUESTIONS**

Your participation is completely anonymous, but some general information will help with analysing the data. Is that OK?

○ Yes (1)

○ No, I will finish now (2)

30.1

Your gender?

○ Male (1)

○ Female (2)

○ Prefer not to say (3)

30.2

Your age?

○ 18-30 (1)

○ 31-50 (2)

○ over 50 (3)

○ Prefer not to say (4)

30.3

Please type any comments in the box below