Paradigmatic Structures in Morphological Processing

Computational and Cross-Linguistic Experimental Studies

ISBN: 90-76203-2
Cover illustration: “Man catching a tree in his net” by Elena Cobos, 2003
Printed and bound by Ponsen & Looijen bv, Wageningen

Paradigmatic Structures in Morphological Processing

Computational and Cross-Linguistic Experimental Studies

een wetenschappelijke proeve
op het gebied van de Letteren

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
op gezag van de Rector Magnificus Prof. dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen
op woensdag 3 december 2003
des morgens om 10.30 uur precies

door
Fermín Moscoso del Prado Martín
geboren op 24 april 1974 te Ferrol (Spanje)

Promotor:	Prof. dr. R. Schreuder
Co-promotor:	Dr. R. H. Baayen
Manuscriptcommissie:	Prof. dr. W. Vonk
	Prof. dr. J. Nerbonne (Rijksuniversiteit Groningen)
	Prof. dr. R. Shillcock (University of Edinburgh, Schotland)

The research reported in this dissertation was supported by a PIONIER grant from the Dutch National Research Council NWO, the Faculty of Arts of the University of Nijmegen (The Netherlands) and the Max Planck Institute for Psycholinguistics, Nijmegen (The Netherlands).

A mis padres, por su apoyo,
y a mi ahijada Matilde,
que nació al tiempo que esta tesis.

Acknowledgements – Dankwoord – Agradecimientos

The book is ready. Looking back on the time it took to write this dissertation, despite the nice discussions with many people, the hours writing code and analyzing results, the brain-crushing arguments at Harald's office, the nice travelling to conferences, and the fearsome periods of stress, it appears to me that it was the raw passing of time that finally had it written, without much intervetion by myself. Of course this is not completely true, the task has involved large amounts of work by me, and even greater amounts of scientific, technical, and personal support from many people.

Rob en Harald zijn de best mogelijke promotor en co-promotor voor me geweest. Zonder hun hulp, enthousiasme, en advies, zou ik nooit dit proefschrift kunnen hebben schrijven. Harald heeft veel van zijn tijd aan mij besteed; wij hebben samen data geanalyseerd, en zelfs ook samen geschreven! Harald werkt met zo ongelofelijk veel energie en enthousiasme, dat het belangrijk is dat Rob er ook is. Met zijn rust, ervaring, en wijsheid, is Rob de perfecte aanvulling voor de kennis van Harald. Ik heb dus veel geleerd van allebei.

Ik moet ook de andere 'pioniers' danken: Andrea, Nivja, en Rachèl waren mijn kamergenoten, en ik heb met veel plezier met hun samen geleerd, en Mirjam was er ook altijd om me te helpen. /voːrɑl mut ɪk mɪrjɑm dɑŋkə voːr haːr hʏlp mɛt neːdərlɑndsə œytspraːk ən fonoloɣi/. Van Nivja heb ik geleerd waarom bomen niet symmetrisch zijn. Zij is, ook buiten het MPI, een echt goede vriendin geweest, en samen hebben wij de 'Ph.Beer' gecreëerd. Ook de student-assistenten van de groep hebben veel goed werk voor ons gedaan.

There are also many people without whose collaboration different parts of this work would not have been possible. Avital Deutsch and Ram Frost were wonderful hosts in Jerusalem. They managed to create the best possible cocktail of work and pleasure. Without their collaboration, the Hebrew experiment in Chapter 2 of

this dissertation wouldn't have been possible. From Turku, Raymond Bertram and Tuomo Häikiö run the Finnish experiment that is reported in Chapter 3. Chapter 4 emerged from a wonderful, and still ongoing collaboration with Ton Dijkstra and Béryl Schulpen, who made their datasets available to me. Chapter 5 was born out of an intense month of hard working, loud arguing, and compulsive smoking during Saša Kostić's visit to Nijmegen. The models reported in Chapter 7 are the result of my collaboration with Mirjam Ernestus, who made her datasets available to me and designed the feature representations of Dutch phonemes that I also used in Chapter 6. The ideas presented in Chapter 9 are the result of beer conversations in Helsinki with Magnus Sahlgren, with whom I also spent a wonderful time in Las Palmas. Liz Bates, Walter Daelemans, Jeff Elman, Laurie Feldman, Scott McDonald, and Dominiek Sandra gave useful comments, advice, and suggestions on different parts from this dissertation. Wietske Vonk, John Nerbonne, and Richard Shillcock were the manuscript committee for this thesis. Their helpful comments and suggestions turned my original manuscript into a proper dissertation. Many of the ideas contained here, are partly the consequence of the year I stayed at the University of Hertfordshire, where Neil Davey, Dave Messer, and Pam Smith first introduced me to connectionism. Also in Britain, I am indebted to Kim Plunkett, who first put me in contact with Harald. This work would not have been possible without the work of Christa Hausman-Jamin, Jethro Zevenbergen, and Bjorn Baayen, who kept my computer running smoothly for three years. I also want to thank the people at administration of the MPI, they managed to solve every paperwork problem in an amazingly friendly, flexible, and effective way.

My flatmates Anita, Ambra, and Michel were extremely friendly and supportive, epecially during the last months of writing this dissertation, were my stress levels rose far beyond what they should have had to put up with. The whole group of students and post-docs from the MPI and the IWTS were lovely colleagues and friends without whom I wouldn't have survived in Nijmegen. From this group I must give very special thanks to Asifa, Joana, Mandana, Petra, and Stuart who provided invaluable help in organizing things, moving furniture, and preparing talks. Additionally, Heidi helped me change my perspective on The Netherlands. Also Stasinos and his amazing Groningen party band and Nivja's Leiden connection made The Netherlands a fun place to live.

Markus and Anita are my 'paranimphs'. After I left the Netherlands, they helped a lot in arranging stuff for the printing and defense of this dissertation. I will always remember the biertjes at Frowijn with Anita and jazz-mondays at Odessa

with Markus.

A Jesús Cardeñosa y Luis Iraola les agradezco que me introdujeran en el mundo de la investigación durante mi etapa en la Universidad Politécnica de Madrid. En cierto modo, esta tesis es la continuación del trabajo que comencé haciendo en aquel grupo, donde recibí mi primera introducción a la lingüística computacional.

Gracias a Caroline y David, y su maravilloso eje Leiden–Lieja por algunas fiestas inolvidables. Clara y Nadi, junto con unos cuantos que pasaron temporadas, fueron mi pequeña España en Nimega. Desde España, Enrique, Calo, Perico, Cubi, Sandra, Javi, y Elena proporcionaron su apoyo moral, junto con Germán desde Ginebra y Gildo desde Nairobi. A Elena le debo también la preciosa composición que aparece en la portada de esta tesis. Le debo un agradecimiento muy especial a Gaby, que pasó muchos de estos momentos conmigo.

Marián ha sido un apoyo imprescindible durante este último año. Ella soportó toda la tensión del final de mi tesis, la búsqueda de trabajo y mis agobios varios. Desde el principio de esta tesis, mis padres me han apoyado. A ellos más que a nadie, les debo el apoyo moral y económico constante durante la duración completa de esta tesis, que está co-dedicada a ellos. En los días en que esta tesis se terminaba de escribir, nació mi ahijada Matilde, hija de mi hermana Lucía y de Perico, a los que dedico un abrazo muy fuerte.

Nijmegen, September 2003.

Contents

Introduction

CHAPTER 1

Morphological Paradigms

Most words belong to morphological paradigms. Inflectional paradigms have been studied intensively across a wide range of languages. The following Spanish forms illustrate part of the inflectional paradigm of the verb *cantar*, namely, the forms of the indicative present tense.

canto	I sing
cantas	you [sg.] sing
canta	he/she/it sings
cantamos	we sing
cantáis	you [pl.] sing
cantan	they sing

All these forms share the stem (*cant*), followed by suffixes marking different combinations of person and number. Note that the corresponding paradigm in English contains only two forms where Spanish has six.

Paradigms can be found not only in inflection, but also in word formation (derivation and compounding). The difference between inflectional paradigms and what we will call derivational paradigms is that the latter have little structure compared to the former. Nevertheless, derivational paradigms often have some hierarchical structure, as shown by the morphological family of the Dutch verb *werken*:

werk	work
werkloos	unemployed
werkloze	unemployed person
werkloosheid	unemployment
werkloosheidsuitkering	unemployment benefit
werkloosheidsverzekering	unemployment insurance

huiswerk	homework
huiswerkvrij	without homework
werkbaar	feasible
onwerkbaar	unfeasible
onwerkbaarheid	unfeasibility

Within the derivational paradigm of *werk*, we find several nested subparadigms ranging from the semantic domain of unemployment to the semantic domain of feasibility. Each of these subparadigms may contain further subparadigms, as shown in some detail for the subfamily of unemployment. Note that Dutch lexicalizes a range of meanings using the same stem *werk*, while English makes use of a range of different stems, such as *work*, *employ*, and *feasible*.

Previous experimental work (Schreuder & Baayen, 1997; Bertram, Schreuder, & Baayen, 2000; De Jong, 2002) has shown that the size of a word's derivational paradigm, its morphological family size, is an independent predictor of processing latencies in lexical decision side by side with other lexical variables such as word frequency, lexical density, and word length. These studies, however, have been primarily concerned with either documenting the validity of the morphological family size as an independent factor in word recognition, or with establishing that the effect arises at the semantic level of lexical processing.

This thesis investigates the consequences of the *structure* in the derivational paradigms for lexical processing in the mental lexicon, and it also presents some first results on inflectional paradigms. The thesis combines a cross-linguistic experimental perspective with an information-theoretic, computational perspective. Chapters 2, 3 and 4 report a series of visual lexical decision experiments addressing the family size effect in Hebrew (Hamo-Semitic) and in Finnish (Finno-Ugric), as well as in the Dutch-English bilingual lexicon. The results of computational and information-theoretical modelling are described in Chapters 5 to 10.

Experimental Research

Previous research has documented the effect of morphological family size in visual lexical decision in several languages of the Indo-European (Germanic) family, including Dutch (e.g., Schreuder & Baayen, 1997), English (e.g., Baayen, Lieber, & Schreuder, 1997), and German (Lüdeling & De Jong, 2002). These studies showed that the morphological family size effect arises at the semantic level of processing. Schreuder and Baayen (1997) and Bertram et al. (2000) reported that the

exclusion of semantically opaque members of the paradigm from the family size counts improved the correlations with response latencies. De Jong, Schreuder, and Baayen (2000) showed that for an irregular verb stem such as *vocht* in the past participle *gevochten* ("fought"), it is the count of family members of the verb *vechten* that is the relevant predictor, and that the count of family members of the homograph *vocht*, "moisture", is irrelevant. Other evidence supporting the semantic locus of the effect is the finding that homonymic words show differential family size effects for each of their readings as a function of the context in which they appear (De Jong, 2002).

In this thesis, the empirical foundation of the morphological family size effect is broadened by moving from Germanic to two other, non-Indo-European, languages: Hebrew and Finnish. These languages are especially interesting because their morphological systems lead to rather different mappings between meanings and forms. Hebrew morphology is well-known for its non-concatenative structure (Berman, 1978; McCarthy, 1981), combining consonantal roots with 'word patterns' composed of vowels and consonants. Due to the limitations on the combinatorics of non-concatenative word formation and the near absence of compounding, derivational paradigms in Hebrew tend to be small, with at most some 30 family members. Moreover, the structure of Hebrew paradigms tends to be flat, often with just a single hierarchical level of complexity. Few words enter into clear derivational relations, most words are all 'sisters' derived in a single step from their common root.

Finnish contrasts with Hebrew in several ways. Its morphology is often characterized as agglutinative, as in this language inflectional and derivational suffixes combine to form long highly complex forms (e.g., *talo-i-ssa-ni-kin-ko*, house-plural-inessive-my-too-question, "in my houses, too?"). As compounding is also highly productive in Finnish, morphological families may have up to 7000 family members, although on average a words will have 'only' one or two hundred family members. Compared to Dutch (with maximally 550 family members), English (maximally 200 family members), and Hebrew (maximally 30 family members), Finnish distinguishes itself as a language with an extremely rich morphology. Due to the many successive layers of affixation and compounding, Finnish derivational paradigms are often characterized by a structure with multiple hierarchical levels.

Considered jointly, Finnish, Dutch, and Hebrew form a continuum in terms of their morphological complexity, with the morphology of Finnish being the most productive, the morphology of Hebrew being the least productive, and the morphology of Dutch having an intermediate productivity. This typological difference raises two

questions. First, it is not clear whether one should expect to find family size effects in Hebrew and Finnish. The families in Hebrew might be too small to reveal a measurable effect in behavioral measures such as response latencies in visual lexical decision. Conversely, the huge families in Finnish might lead to a ceiling effect, with the differences between ‘few’ and ‘some more’ observed for Dutch being swamped by the differences between ’very many’ and ’even more’ in Finnish. Second, the structure in the derivational paradigms is rather different across the three languages, and this might also affect lexical processing.

In spite of the small size of derivational paradigms in Hebrew, Chapter 2 nevertheless reports evidence for a family size effect in Hebrew. The key finding for Hebrew is that the family size effect is modulated by the semantic structure in the derivational paradigm. Some Hebrew roots give rise to words that belong to only one semantic field. Other roots, however, are homonymic in the sense that they structure words that belong to different semantic fields. For instance, the root {*r,g,l*} participates in two semantic fields, one pertaining to a body part (*regel*–“foot”/“leg”), the other pertaining to espionage (*meraggel*–“spy”). Chapter 2 documents evidence that the count of family members in the same semantic field as the target word is facilitatory, and that the count of family members belonging to the other semantic field(s) is inhibitory. This result supports the conclusions reached by earlier studies, namely, that the family size effect is a semantic effect.

Recall that derivational paradigms in Finnish may have up to 7000 members, which raises the possibility of obtaining a ceiling effect in this language. Chapter 3 nevertheless reports a morphological family size effect for Finnish, but with a twist. For simplex words, the total family size count is the crucial predictor, as in Dutch or English. But for complex words, the family size effect is more restricted in nature, with only the count of derived words and compounds containing the complex word as a constituent, the ‘dominated family size’, as the crucial predictor. Family members that are not derived from the complex word itself (descendants of parallel branches in the tree), do not contribute to the effect.

Parallel to the Hebrew and Finnish experiments, we ran a Dutch lexical decision experiment. In fact, all three experiments (Finnish, Hebrew, and Dutch) used words that were translation equivalents across the three languages. This allowed us to investigate the cross-linguistic predictive power of morphological family size counts. Interestingly, Dutch family size counts turned out to predict the Hebrew visual lexical decision latencies to their translation equivalents. Conversely, Hebrew family size counts predicted Dutch response latencies. We observed the same cross-

language predictivity between Dutch and Finnish. This cross-language predictivity remains even after first partialling out the within-language variables (frequency, length, and family size). No such predictivity, however, was visible between Hebrew and Finnish. These results suggest that there is a large degree of overlap in the conceptual organization of the mental lexicon across speakers of languages with not too dissimilar morphological productivity, such as Hebrew and Dutch, or Dutch and Finnish. The absence of predictivity between Finnish and Hebrew indicates that the extreme differences in their morphological productivity affect the organization of the conceptual systems subserved by these languages.

Chapter 4 addresses the processing of interlingual homographs in the mental lexicon of Dutch-English bilinguals. Interlingual homographs are words with identical spellings in two languages, but with different meanings and to a smaller or greater extent differences in pronunciation (e.g., *angel*, ‘hook’, ‘sting’ in Dutch, and a heavenly creature in English). A re-analysis of two experiments reported by Schulpen (2003) and an additional visual lexical decision experiment revealed unambiguous family size effects of both languages in the bilingual mental lexicon. For participants performing Dutch lexical decision, the Dutch family size count is facilitatory, and the English family size count inhibitory. For participants performing English lexical decision, the effects reverse with the English family size count being facilitatory and the Dutch count inhibitory. This points to parallel activation of the readings of the homograph and their derivational paradigms in both languages. Crucially, these results hold irrespective of whether the participants performed the experiment in their first language, and of whether they were aware of the presence of words from the other language. Note that these results are consistent with the inhibitory and facilitatory effects of the family size counts for the different semantic fields of homonymic Hebrew roots.

Modeling Research

For the modelling of paradigmatic effects in word recognition, we have taken two complementary approaches. The first approach is reported in Chapter 5, which introduces an information-theoretical characterization of derivational and inflectional paradigms. We define three measures quantifying the cognitive cost of recognizing a word: derivational entropy, inflectional entropy, and the information residual of a word. These measures, calculated over pre-defined tree-like symbolic structures (derived from the morphological parses provided by CELEX), estimate the sup-

port that morphological paradigms provide to the recognition of their members. By means of reanalyses of previously published data, we show that our information-theoretical measures outperform in terms of explained variance any combination of traditional measures of morphological complexity such as base frequency, morphological family size, and cumulative root frequency. The advantage of using this concise mathematical formulation is that it allows us to analyze in detail the consequences of the hierarchical structure within the paradigms for lexical processing. An additional advantage is that inflectional and derivational paradigms can be treated within a unified framework.

Despite its explanatory power, this information theoretical approach also has its drawbacks. First, it does not account for the emergence and representation of the tree-like structures over which the measure is calculated. Second, it offers no insight into how the required probabilities and entropies might be estimated in the mental lexicon. Third, and most importantly, the morphological parsing technology on which the calculations are based is too rigid to capture the graded morphological effects that have been reported in the literature. For instance, Bergen (2003) reports morphological priming (controlled for formal and semantic similarity) between certain cluster of letters that exhibit coherent semantic properties, such as the English word initial 'gl-' present in *glow*, *glitter*, *glisten*, and many other words all relating to the concept of light. These effects cannot be accounted for using a decompositional technology that depends on analyzing words into sequences of discrete morphemes.

We therefore also explored the possibilities of a distributed connectionist approach. Is it possible to express tree-like paradigmatic structures in terms of distributed patterns of activation in multi-dimensional similarity space? Might distances in such a multi-dimensional space correspond with human processing load as measured by lexical decision latencies? This approach has the advantage of allowing us to exploit highly sensitive corpus-based measures of lexical similarity. The challenge of using this technique is to explore to what limits such a statistical approach can be driven.

A broad-coverage distributed connectionist model of lexical processing requires the development of realistic representations of the forms and meanings of words, ideally covering all words in a language. Previously published analogy-based models of lexical processing such as AML (Skousen, 1993), as well as previously implemented connectionist models (e.g., Harm & Seidenberg, 1999; Plaut & Booth, 2000; Plaut & Gonnerman, 2000; Shillcock, Ellison, & Monaghan, 2000), make use of lo-

cal, sequential coding schemes. These schemes present problems for positional similarity. For instance, consider the segment overlap in the Dutch words *sap*, *stap*, and *tap*. Any alignment scheme for a slot-based template will miss part of the similarities between them, and will require arbitrary decisions on the placement of 'gaps'. Other authors (e.g., Mozer, 1987, Seidenberg & McClelland, 1989) have used variants of the 'wickelgraph' (cf., Wickelgren, 1969) to solve this problem. However, Prince & Pinker (1988) have shown that 'wickelgraphs' do not distinguish all words in a language.

In Chapter 6, we address the problem of creating realistic form representations. We develop the Accumulation of Expectations (AoE) technique for representing the orthographic and phonetic forms of all words in a language, using a completely distributed corpus-based approach. These orthographic vectors are put to the test on a realistic problem in Chapter 7. In this chapter, we address the problem of past-tense formation and phoneme to grapheme mapping for the complete Dutch verbal system. We show that a distributed connectionist model developed using AoE representations is able to capture the graded paradigmatic effects in Dutch phonological perception as described by Ernestus and Baayen (2001; 2003).

In chapters 8 and 9, we address the problem of creating realistic semantic representations. Currently available resources for semantic representation are not adequate for our goals. Co-occurrence-based approaches such as HAL (Lund & Burgess, 1996) or LSA (Landauer & Dumais, 1997) create their representations on the basis of lemmatized corpora, and therefore completely lack information about inflectional variants. Hence, they cannot be used to study inflectional paradigms. In addition, these techniques make the simplifying assumption of considering texts as 'bags' of words, and therefore lack the sensitivity to sequential order characteristic of human language processing (e.g., Tabor, Juliano, & Tanenhaus, 1997; see also De Jong, 2002 and Kostić, 2003 for morphological paradigms in context). Existing symbolic resources such as WordNet (Miller, 1990) cannot be used either, because they lack inflectional and contextual information altogether.

In the absence of adequate realistic semantic resources, connectionist models of language processing have mostly relied on unrealistic, even artificial, representations of the meanings of words (e.g., Plaut & Booth, 2000). Instead of avoiding this problem, we address it by creating a corpus-based co-occurrence technique sensitive to word order, that provides us with different representations for each inflectional variant of a word. Chapter 8 presents a first attempt to create such semantic vectors. Although this technique did not succeed in capturing sufficient

semantic detail, it produced useful representations of the morpho-syntactic properties of words. Chapter 9 describes a better technique for creating semantic vectors. It builds on the morpho-syntactic vectors developed in Chapter 8, combined with a variant of the Accumulation of Expectations technique (see Chapter 6). A range of tests suggests that the amount ofsemantic detail contained in the vector representations is sensitive enough for our purposes.

Finally, in Chapter 10, we combine the form and meaning representations in a simple three-layered backpropagation network. We trained this network to produce at its output the semantic vector corresponding to an orthographic vector presented at its input, over a large vocabulary of around 50,000 different words. After training, the cosine distance between the network's output and the corresponding semantic prototype emerged as a good predictor of response latencies in visual lexical decision. Our model shows effects similar to those observed for participants in visual lexical decision. Crucially, the effects of the inflectional and derivational entropy measures calculated using symbolic structures (see Chapter 5) emerge as well in this distributed system, indicating that it has captured important aspects of the information structure carried by morphological paradigms.

Our simulation studies have resulted in a model of visual lexical decision with a coverage at least one order of magnitude above previously implemented connectionist models of lexical processing. In this respect, our distributed model constitutes the first real-scale connectionist approximation of this task. Although the model's architecture does not incorporate any representations and operations specifically designed to account for morphology, it does show considerable sensitivity to morphological structure and regularity. What we find attractive in this model is that its representations are corpus-based, using only minimal assumptions about the cognitive system such as the projection of expectations (e.g., Tabor et al., 1996) and sensitivity to word co-occurrence (McDonald & Ramscar, 2001; Boroditsky & Ramscar, 2003).

The price we pay with this approach is that the direct, intuitive, interpretation of the processes driving word recognition gets lost in hyperspace. Therefore, the symbolic information-theoretic approach developed in Chapter 5 remains an important complement to this distributed approach. While the information-theoretical approach provides useful tools for the concise specification and interpretation of the probabilistic factors that influence lexical processing, the distributed connectionist account provides a simple mechanism for accounting for the statistical synergy between form and meaning known as morphology.

References

Baayen, R. H., Lieber, R. and Schreuder, R.: 1997, The morphological complexity of simplex nouns, *Linguistics* **35**, 861–877.

Bergen, B. K.: 2003, The psychological reality of phonaesthemes, *Manuscript submitted for publication, University of Hawai'i at Manoa*.

Bertram, R., Schreuder, R. and Baayen, R. H.: 2000, The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26**, 419–511.

Boroditsky, L. and Ramscar, M.: 2003, Guilt by association: Gleaning meaning from contextual co-occurrence, *Manuscript, Massachusetts Institute of Technology*.

De Jong, N. H.: 2002, *Morphological Families in the Mental Lexicon*, MPI Series in Psycholinguistics, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: 2000, The morphological family size effect and morphology, *Language and Cognitive Processes* **15**, 329–365.

Ernestus, M. and Baayen, H.: 2001, Choosing between the Dutch past-tense suffixes *-te* and *-de*, *in* T. van der Wouden and H. de Hoop (eds), *Linguistics in the Netherlands 2001*, Benjamins, Amsterdam, pp. 81–93.

Ernestus, M. and Baayen, R. H.: 2003, Predicting the unpredictable: Interpreting neutralized segments in Dutch, *Language* **79(1)**, 5–38.

Harm, M. W. and Seidenberg, M. S.: 1999, Phonology, reading acquisition, and dyslexia: Insights from connectionist models, *Psychological Review* **106**, 491–528.

Kostić, A.: 2003, The effects of the amount of information on processing of inflected morphology, *Manuscript submitted for publication, University of Belgrade*.

Landauer, T. and Dumais, S.: 1997, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge, *Psychological Review* **104**(2), 211–240.

Lüdeling, A. and De Jong, N. H.: 2002, German particle verbs and word-formation, *in* N. Dehé, R. Jackendoff, A. McIntyre and S. Urban (eds), *Verb-particle explorations*, Mouton de Gruyter, Berlin, pp. 315–333.

Lund, K. and Burgess, C.: 1996, Producing high-dimensional semantic spaces from lexical co-occurrence, *Behaviour Research Methods, Instruments, and*

Computers **28**(2), 203–208.

McCarthy, J. J.: 1981, A prosodic theory of non-concatenative morphology, *Linguistic Inquiry* **12**, 373–418.

McDonald, S. and Ramscar, M.: 2001, Testing the distributional hypothesis: The influence of context judgements of semantic similarity, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

Miller, G. A.: 1990, Wordnet: An on-line lexical database, *International Journal of Lexicography* **3**, 235–312.

Plaut, D. C. and Booth, J. R.: 2000, Individual and developmental differences in semantic priming: Empirical and computational support for a single mechanism account of lexical processing, *Psychological Review* **107**, 786–823.

Plaut, D. C. and Gonnerman, L. M.: 2000, Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing?, *Language and Cognitive Processes* **15**(4/5), 445–485.

Prince, A. and Pinker, S.: 1988, Wickelphone ambiguity, *Cognition* **30**, 189–190.

Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **37**, 118–139.

Schulpen, B. J. H.: 2003, *Explorations in bilingual word recognition: Cross-modal, cross-sectional, and cross-language effects*, PhD thesis, University of Nijmegen, Nijmegen, The Netherlands.

Seidenberg, M. S. and McClelland, J. L.: 1989, A distributed, developmental model of word recognition and naming, *Psychological Review* **96**, 523–568.

Shillcock, R., Ellison, T. M. and Monaghan, P.: 2000, Eye-fixation behaviour, lexical storage and visual word recognition in a split processing model, *Psychological Review* **107**, 824–851.

Skousen, R.: 1993, *Analogy and Structure*, Kluwer, Dordrecht.

Tabor, W., Juliano, C. and Tanenhaus, M. K.: 1997, Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing, *Language and Cognitive Processes* **12**(2/3), 211–271.

Morphological Family Size across Hebrew and Dutch

CHAPTER 2

This chapter has been submitted as Fermín Moscoso del Prado Martín, Avital Deutsch, Ram Frost, Robert Schreuder, Nivja H. de Jong, and R. Harald Baayen: Changing places: A cross-language perspective on frequency and family size in Hebrew and Dutch.

Abstract

This study uses the morphological family size effect as a tool for exploring the degree of isomorphism in the networks of morphologically related words in the Hebrew and Dutch mental lexicon. Hebrew and Dutch are genetically unrelated, and they structure their morphologically complex words in very different ways. Two visual lexical decision experiments document substantial cross-language predictivity for the family size measure after partialing out the effect of word frequency and word length. Our data show that the morphological family size effect is not restricted to Indo-European languages but extends to languages with non-concatenative morphology. In Hebrew, a new inhibitory component of the family size effect emerged that arises when a Hebrew root participates in different semantic fields.

Introduction

The morphological family size of a word is the type count of all the complex words in which this word appears as a constituent (Schreuder & Baayen, 1997, Baayen, Lieber, & Schreuder, 1997). Words with large morphological families elicit faster response latencies in visual lexical decision than words with small morphological families. Although morphological family size is highly correlated with word frequency, i.e., the more frequent a word is, the larger its morphological family size tends to be, the effect of family size is present even after having controlled for the effects of word frequency (Schreuder & Baayen, 1997). The growing body of experimental evidence on the morphological family size effect shows that the locus of this effect is at the semantic level. Schreuder and Baayen (1997) and Bertram, Schreuder, and Baayen (2000) pointed out that the exclusion of opaque family members from the family counts improves the correlations with response latencies. De Jong, Schreuder, and Baayen (2000) provided further evidence that the effect of family size is not mediated by surface form. For instance, the Dutch irregular past participle "ge**vocht**en" (i.e., *fought*) activates only the family of "**vecht**en" (i.e., *to fight*) and not that of "**vocht**" (*moisture*). In addition, homonyms show differential family size effects when presented in a disambiguating context (De Jong, 2002). The effect of family size has been observed not only for Dutch, but also for English (Baayen, Lieber, & Schreuder, 1997, De Jong et al., 2002) and German (Lüdeling & De Jong, 2001). Additionally, Feldman and Siok (1997, 1999) report a similar effect in visual lexical decision in Chinese, where characters with semantic radicals (a pseudo-morphological unit without a phonetic realization) that are present in many other characters are recognised faster.

Although previous research indicates that the locus of the family size effect is at the semantic and not at the formal level, it remains an open question how general this finding is. It might be the case that, in other languages with different morphological structures, in which the formal aspect plays a crucial role in lexical organization and processing, it is this formal aspect of the morpheme rather than its semantic characteristics that is the factor governing the family size effect. An example of such a language is Hebrew, which belongs to the Hamo-Semitic language family, in contrast with the Indoeuropean languages that were previously studied.

All Hebrew verbs and most Hebrew nouns and adjectives are comprised of two basic derivational morphemes: the root and the word pattern. The root usually consists of three consonants, while the word pattern consists of a cluster of vowels, or vowels and consonants. The root generally conveys the core meaning of the

word, while the word-pattern creates variations on its meaning, and determines the grammatical properties of the word. This combination of root and word patterns is an example of a non-concatenative morphology (McCarthy, 1981; Berman, 1978). The morphological units of Hebrew words cannot generally be found by splitting words into sequences of morphemes as beads on a string. More complex interposition processes are required.

Although word patterns shape the meaning of words, the exact meaning of a word cannot be unequivocally predicted from its constituent morphemes, the root and the word pattern. This is because of the linguistic characteristics of the word patterns on the one hand, and the existence of homonymic roots on the other hand. As to the word patterns, although they contribute to the meaning of words, their semantic characteristics, most markedly in the nominal system, can be ambiguous: Many of the nominal word patterns can denote more than one semantic category (e.g., the nominal word pattern *-a-e-et* can denote both profession and disease) on the one hand, and a specific semantic category can be expressed by more than one nominal pattern (e.g., profession can be denoted by the nominal patterns *-a-e-et* and *-a–a-*.[1] There are several reasons for the abundance of homonymic roots in Hebrew, however, the description of these reasons falls out of the scope of this article. An example of such a homonymic root is *G-D-R*[2], which is present in the words *GaDeR* (*fence*) and *haGDaRa* (*definition*). In contrast, as an example of a root with a constant semantic contribution, consider the root *SH-M-N*, which always carries the core meaning *fat*. This root can be embedded in the nominal pattern *-e-e-* to form the noun *SHeMeN* (*oil*), in the adjectival pattern *-a-e-* to form the adjective *SHaMeN* (*fat*), or in the verbal pattern *hi--i-* creating the verb *hiSHMiN* (*to become fat*).

Previous studies in Hebrew using the masked and the cross modal priming paradigms have demonstrated the role of the root morpheme in the lexical access of Hebrew complex words, suggesting that root morphemes constitute lexical units in the Hebrew lexicon (Frost, Forster and Deutsch, 1997; Frost, Deutsch, Gilboa, Tanenbaum, & Marslen-Wilson, 2000; Deutsch, Frost, Pollatsek, & Rayner, 2000). Furthermore, their findings indicate that the lexical status of the root morphemes and their role in mediating lexical access is not conditioned by semantic transparency (Frost et al., 1997), although it can interact with semantic effects (Frost, Deutsch, Gilboa et al., 2000). In addition, further investigation of irregular

[1]The long dash between the two vowels represents a gemination or ‘doubling’ of the corresponding root consonant.

[2]Root consonants are shown in upper case throughout this paper

root morphemes, the so-called defective roots in which one or two consonants may be absent, indicates that the role of the root in mediating lexical access is very sensitive to the abstract structural characteristics of the morpheme (Frost, Deutsch, & Forster, 2000). Violation of the standard morphological form inhibits the priming effect between morphologically related words. Possibly, the exact repetition of the root structure allows the root to acquire a formal representation that is independent of meaning. In other words, the Hebrew root seems to be an autonomous form unit, in the sense of Aronoff (1994).

If the representation of the root is indeed independent of meaning, meaning overlap would not be a requirement for a word to participate in a given morphological family. We will refer to this possibility as the *formal family hypothesis*. On the other hand, considering that Hebrew roots generally carry the semantic fields of words, and that the morphological priming effect in Hebrew interacts with semantic transparency (Frost, Deutsch, Gilboa et al., 2000), it is not impossible that the family size effect in Hebrew might have a conceptual origin, just as in Germanic languages. According to this second hypothesis, that we will call the *conceptual family hypothesis*, Hebrew roots which express more than one semantic field should be split into as many homonymic roots as there are semantic fields, similar to previous findings in Indoeuropean languages.

In Experiment 1, using visual lexical decision, we clarify the nature of the family size effect in Hebrew, i.e., whether it is formal or semantic in nature. We do this by contrasting non-homonymic roots (i.e., roots that appear in words that always share the same semantic field), with homonymic roots (i.e., roots that appear in words belonging to more than one semantic field). If the conceptual family hypothesis is correct, we expect to find differential family size effects for these two kinds of roots. Non-homonymic roots should reveal a family size effect just as observed in visual lexical decision for Dutch, German, and English. Homonymic roots, on the other hand, might reveal separate family size effects for each of its semantic fields (cf., De Jong, 2002). Alternatively, if the formal family hypothesis is correct, it makes no sense to distinguish between homonymic and non-homonymic roots in Hebrew. In that case, the family size effect will be caused by the total number of words derived from the homonymic root (regardless of the different semantic fields) just as do non-homonymic roots.

The second issue that this study addresses is the extent to which Hebrew and Dutch are isomorphic at the conceptual level. The issue of cross language isomorphy at the conceptual level is addressed by Bates et al. (in press). They report that

in picture naming in seven different languages, naming latencies correlate with the (objective or subjective) frequency of a word in a language equally strongly as they correlate with the frequency of its translation into another language. They found, for instance, that Chinese frequencies predict English naming latencies just as well as English frequencies do, and vice versa. However, it is at present unclear to what extent perceptual familiarity with the word form on the one hand (Bradley & Foster, 1987, Morton, 1969) and familiarity with specific objects and concepts on the other hand (Becker, 1979, Stanovich & West, 1981, Borowsky & Besner, 1993, Plaut & Booth, 2001), contribute to this symmetric word frequency effect in picture naming across languages. Although it seems likely that there is a substantial conceptual component to the word frequency effect, the cross-linguistic evidence does not rule out the possibility of form playing an important role as well.

To gain further insight into the extent to which the conceptual systems of unrelated languages might be isomorphic, we use the family size effect as a diagnostic tool. Parallel to Experiment 1, we report a second experiment with the Dutch translations of the Hebrew words. This experiment will allow us to ascertain whether the cross-linguistic predictivity of word frequency reported by Bates and colleagues generalizes to visual lexical decision. Additionally this experiment will also make it possible to investigate the cross-linguistic predictivity of another count not considered by Bates and her colleagues, the family size effect. Given the conceptual nature of the family size effect, if Hebrew family sizes predict Dutch response latencies after having partialled out the effects of Dutch word frequency, word length and family size, and, likewise Dutch counts predict Hebrew response latencies, this would provide clear evidence for considerable overlap in cross-linguistic lexico-semantic organization.

Experiment 1

Method

Participants. Forty undergraduate students at the Hebrew University were paid to take part in this experiment. All were native speakers of Hebrew.

Materials. We compiled a list of ninety-nine Hebrew roots and their morphological families using a Hebrew dictionary (Schweika, 1997). Forty-three of these roots were non-homonymic in the sense that all their family members belong to the

same semantic field. The remaining fifty-six roots were homonymic, i.e., their family members belong to two or more different semantic fields. An example of a non-homonymic root is *B-G-R*, which has as family members the words *BaGRut* (*maturity*), *mBuGaR* (*an adult*), *hitBaGRut* (*maturation*), *mitBaGeR* (*teenager*), *BoGeR* (*mature*), *BaGaR* (*to grow up*), *BiGGeR* (*to grow*) and *hitBaGeR* (*to mature*). By homonymic root we refer to Hebrew families whose members cluster into less similar semantic fields. An example of a homonymic family is the root *X-SH-B*, which appears in two semantic fields, one relating to thinking, e.g., *XaSHaB* (*to think*), *maXSHaBa* (*a thought*), *XaSHiBa* (*thinking*), and one relating to arithmetics and calculations, e.g., *XiSHeB* (*to calculate*), *XeSHBon* (*arithmetics*), *XiSHuB* (*calculation*). It can be observed in this example that, although thinking and calculating are undoubtedly related in meaning, their relationship is clearly more distant and less predictable than the relation between any of the words within each of those subfields.

The assignment of words to semantic fields, carried out by the second author on the basis of dictionary definitions, is supported by an analysis of semantic distances based upon Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). As no operational LSA system is available for Hebrew, we obtained vector-based semantic similarity scores for the English translations of our Hebrew words, using the scripts available at `http://lsa.colorado.edu` with their default settings. The average similarity between the pairs of target words from different semantic fields in Hebrew was 0.09. We compared these similarities with pairs of words from Hebrew non-homonymic roots, choosing the target word and a member of the morphological family with, if possible, a translation equivalent in English that was not morphologically related, to be conservative (e.g., *calculation* and *arithmetic*). The average similarity between these pairs of words was 0.31, which is significantly greater than the average similarity score for the unrelated pairs ($W = 448.5, p < 0.0001$, one-tailed Wilcoxon rank sum test). Thus, our traditional lexicographic assignment of words to semantic fields is supported by current co-occurrence based techniques.

For each non-homonymic root, we selected one family member for presentation in the experiment. For each homonymic root, we selected two words belonging to different semantic fields for inclusion in the experiment. Each of the target words is associated with two counts, the count of the *related family size*, the number of words sharing both root and semantic field, and the count of the *unrelated family size*, the number of words sharing the same root but belonging to different semantic fields. For each word, we also obtained a frequency count using a corpus of 200

million words of Hebrew newspaper texts.

We constructed two experimental master lists. Both master lists contained the 43 words representing the 43 non-homonymic roots. To each master list, we added 56 words with homonymic roots. We assigned the words with homonymic roots to these lists such that two words sharing a given homonymic root did not appear in the same list. Thus, exactly the same 99 roots appeared in both lists. We also made sure that the average related family size was approximately the same across the two lists. Each word was paired with a pseudo-word that did not violate the phonological structure of Hebrew. We constructed three pseudo-randomized versions of each master list, each of which was preceded by a practice session of 10 words and 10 pseudo-words. Both words and pseudo-words were displayed with the vowel signs in order to avoid ambiguity of interpretation of the target words. Table 2.1 provides the means and standard deviations for the frequency, word length, and family size counts for this data set as well as for the response latencies (after application of the data-cleaning procedures described in the Results and Discussion).

Table 2.1: Means and standard derivation for the different counts in the Hebrew data set and the by-item averages of the response latencies, and in the subsets of homonymic and non-homonymic (after removing the 3 outliers and 1 word with unreliable family size counts). The word frequency counts are in occurrences per million.

	Total (151 items)		Non-homonymic roots (43 items)		Homonymic roots (108 items)	
	mean	std.dev.	mean	std.dev.	mean	std.dev.
frequency	32.17	88.44	36.09	111.14	30.63	78.32
word length	4.16	0.89	3.90	0.91	4.26	0.87
total family size	12.06	5.94	8.95	5.44	13.28	5.69
related family size	7.21	4.52	8.95	5.44	6.53	3.93
unrelated family size	4.79	4.79	0.00	0.00	6.75	4.37
family size ratio	0.60	1.25	2.14	0.61	0.00	0.86
response latency	681 ms	56 ms	678 ms	47 ms	682 ms	59 ms

Procedure. Participants performed the experiment in a noise-attenuated experimental room. They were asked to decide as quickly and accurately as possible whether the letter string appearing on the computer screen was a real Hebrew word. Each stimulus was preceded by a fixation mark in the middle of the screen for 500 ms followed after 50 ms blank screen by the stimulus, which appeared at the middle of the screen.

Stimuli were presented on a color monitor in white lowercase 28 point letters on a dark background and they remained on the screen for 1500 ms. The maximum

time span allowed for a response was 2000 ms from stimulus onset.

Results and Discussion

All participants in this experiment performed with an error rate less than 15%. Three words elicited error rates above 30% and were removed from the data set. Table **??** provides the means and standard deviation of the response latencies after removing the three outliers and an additional word whose family size counts were unreliable.

Family size is known to be correlated with frequency (Schreuder & Baayen, 1997). In this data set, the correlation between related log family size and log word frequency is $r = 0.24$ ($p = 0.0031$). The correlations between word frequency, family size, and word length introduce medium collinearity in our data matrix (with a condition number κ of 15.40); consequently, our analyses are to be interpreted with caution when generalizing beyond the range of frequencies and family sizes represented in the data set (Belsley, Kuh, & Welsch, 1980). In order to ascertain whether family size contributes to the reaction times independently of word frequency and word length, we first partial out the effect of word frequency and word length before testing for an effect of family size.

For the words with homonymic roots, there are two family size counts, one count for the family members with meanings related to the meaning of the target word (the related family size), and one count of family members with meanings that are unrelated to that of the target (the unrelated family size). The family size variable that we use is the *family size ratio*,

$$\phi = \log\left(\frac{\text{related family size + 1}}{\text{unrelated family size + 1}}\right), \tag{2.1}$$

where we add 1 to the counts in order to avoid having to take the logarithm of zero. Since:

$$\begin{aligned}\log\left(\frac{\text{related family size + 1}}{\text{unrelated family size + 1}}\right) &= \\ = \log(\text{related family size + 1}) &- \log(\text{unrelated family size + 1}),\end{aligned} \tag{2.2}$$

including the family size ratio in the model is equivalent to including both related and unrelated family size with coefficients equal in absolute value, but opposite in sign.

For the non-homonymic roots, the family size ratio is identical to the logarithm

of related family size + 1, given that the unrelated family size is zero for these words. For the homonymic roots, we use the family size ratio for three reasons. The first reason is technical: Instead of having four correlated variables, we now have only three correlated variables, reducing the collinearity of our data matrix. In this respect, we observed that entering Hebrew related family size and Hebrew unrelated family size as separate predictors in our regression analyses produced very unstable models, in which the significance of the facilitatory effect of related family size, and the inhibitory effect of unrelated family size was dependent of which of them was entered first into the regressions. Instead, the family size ratio allows us to enter both variables into the models at the same time. The second reason is a practical one. It turns out that the family size ratio is a better predictor of response latencies than related, unrelated, or total Hebrew family size taken singly or jointly, and for reasons of space we report only the best models that we have been able to fit to the data. Finally, our third reason was also practical. The family size ratio provides us with a uniform treatment of homonymic and non-homonymic roots. Given that the words from Hebrew non-homonymic roots have an unrelated family size of zero, in order to avoid the violation of the normality assumptions that would be caused to have a variable that is zero for a large set of items, one would need to provide separate analyses of teh homonymic and non-homonymic roots.

For a by-participant regression analysis, Lorch and Myers (1990) provide a technique (used, e.g., by Alegre & Gordon, 1999) that yields the marginal significance for the predictors, but that does not allow a straightforward procedure for partialling out the contribution of word frequency before assessing the contribution of family size. Hence, we have made use of a multi-level extension of the Lorch and Myers approach that overcomes this disadvantage while maintaining the strengths of the original Lorch and Myers technique (Pinheiro & Bates, 2000; Baayen, Tweedie, & Schreuder, 2002). A multi-level linear regression with log response latency as dependent variable and log word frequency, word length, and family size ratio as predictors revealed a significant facilitatory effect of word frequency ($F(1, 3716) = 269.31, p < 0.0001$), but no significant effect of word length after having partialled out the effect of word frequency ($F < 1$). However, the family size ratio emerged as significant ($F(1, 3716) = 21.42, p = 0.0001$), after partialling out the effect of word frequency. We also observed a significant interaction between word frequency and family size ratio ($F(1, 3716) = 20.42, p < 0.0001$).

A by-item multiple regression revealed the same pattern of results: a facilitatory effect of word frequency ($F(1, 145) = 89.94, p < 0.0001$), no effect of word length

($F < 1$), and a significant effect of the family size ratio ($F(1, 145) = 7.55, p = 0.0068$) after partialling out the effect of word frequency. As we did in the by-participant analyses, we observed an interaction between frequency and family size ratio ($F(1, 145) = 5.22, p = 0.0237$).

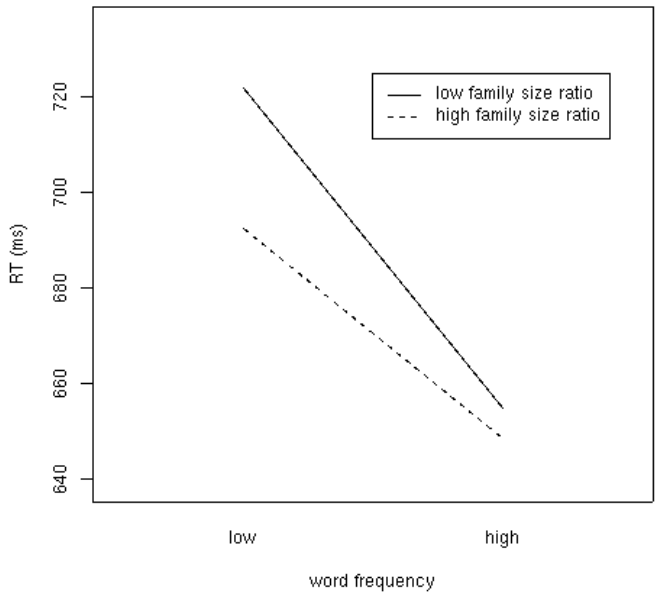

Figure 2.1: Interaction of word frequency and family size ratio in Experiment 1 (factorization by median split).

Figure 2.1 illustrates the nature of the interaction between word frequency and family size ratio by means of a median split factorization. Note that the effect of family size ratio is greater for low-frequency words.

The above analyses showed solid effects of word frequency and morphological family size ratio in both the by-subject and the by-item analyses. None of these analyses revealed an independent contribution of Hebrew word length after having partialled out the effects of word frequency. The effect of written frequency, documented in Hebrew in the present paper for the first time, is not surprising given the strong effects of frequency that have been reported in many other languages, and in cognition in general (e.g., Taft, 1979, Hasher & Zacks, 1984). The lack of reliable effects for word length is probably due to the relatively low variance of Hebrew word length in our data.

More interesting is that, in the three analyses, the family size ratio emerged as a strong independent predictor of response latencies, even after partialling out the effects of word frequency and word length. Our analysis shows a facilitatory effect

for related family size and an inhibitory effect of unrelated family size. In the case of the non-homonymic roots, the unrelated family size count is zero, so only a facilitatory effect is found for these roots. This experiment therefore documents for the first time a Family Size effect for a non Indo-European language. Whereas the Family Size effect in Indo-European languages is anchored in the presence of a shared stem, i.e., an independent morphological unit that, in English, German, or Dutch, can even be a word in the language, the family size effect in Hebrew is anchored in the Semitic root, a non-concatenative morphological unit, which can never exist as a word without a word pattern. Furthermore, the results for the homonymic roots support the conceptual family hypothesis.

The inhibitory effect of the unrelated family size that emerges for the Hebrew words with homonymic roots contrasts with the family size effects documented for homonymic words in Dutch (De Jong, 2002). When Dutch homonymic words are presented in a disambiguating context, such as a function word or a list composed only of nouns or verbs, only the family size of the contextually appropriate meaning has a facilitatory effect. No evidence has been found in Dutch for an inhibitory effect of the contextually inappropriate family members. It is possible that the inhibitory effect of the unrelated family size in Hebrew arises due to the strong role of the Hebrew root. The root may strongly activate words in both families, giving rise to competition between the different meanings of the root.

In order to gain insight into the extent to which Hebrew and Dutch are similar at the conceptual level, even though they are clearly very different at the form level, we ran a second experiment with the Dutch translation equivalents of the Hebrew words. Once we have established the magnitude of the frequency and family size effects for the parallel data sets for Hebrew and Dutch, we will be in the position to address the issues raised by Bates and her colleagues, namely, whether the frequency counts from one language predict the response latencies in the other language equally well as the frequency counts from that other language. Addtionally, this will allow us to investigate whether a similar effect is found for the family size counts, that is, if response latencies in one language can be predicted from family size counts in the other, even after controlling for the effects of the within language variables.

Experiment 2

Method

Participants. Thirty-six undergraduate students at the University of Nijmegen were paid to take part in this experiment. All were native speakers of Dutch.

Materials. The 155 Hebrew words from Experiment 1 were translated into Dutch. The translation was done by hand using a Dutch-Hebrew Dictionary (Bolle & Pimentel, 1984), checked with the English translation of the words provided by the second author, and validated by a Hebrew-Dutch bilingual. When a word had different possible translations into Dutch with different meanings, we included all different translations in the experiment. In this way, after removing those words which appeared twice, we obtained a set of 162 Dutch words. Frequency and family size counts for these words were extracted from the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995) . Each of these words was paired with a pseudo-word whose phonotactics did not violate the phonology of Dutch. Twenty practice trials, ten words and ten pseudo-words were run before the actual experiment. We constructed three different permutations and their corresponding reversed versions of the original word list.

Table 2.2: Means and standard derivation for the different counts after removing the 3 outliers in the Dutch data set.

	mean	standard deviation
frequency	88.26	312.55
word length	7.17	2.47
family size	24.81	30.56
response latency	578 ms	78 ms

Table 2.2 provides the means and standard deviations for word length, logarithmic frequency, and logarithmic family size counts for this data set.

Procedure. Participants performed the experiment in a noise-attenuated experimental room. They were asked to decide as quickly and accurately as possible whether the letter string appearing on the computer screen was a real Dutch word. Following a pause after the test trials, the experiment was run with two further pauses, dividing the experiment in three blocks, each of them containing one third of the stimuli. Each stimulus was preceded by a fixation mark in the middle of the

screen for 500 ms. After 500 ms, the stimulus appeared at the same position. Stimuli were presented on a NEC Multisync color monitor in white lowercase 21 point Arial letters on a dark background and they remained on the screen for 1500 ms. The maximum time span allowed for a response was 2000 ms from stimulus onset.

Results and Discussion

All participants in this experiment performed with an error rate less than 15%. Three words elicited error rates above 30% and were removed from the data set.

A by-participant multilevel linear regression analysis of the data with log reaction time as dependent variable and log frequency, word length, and log family size as independent covariates revealed facilitatory main effects of word frequency ($F(1, 5504) = 783.92, p < 0.0001$), an inhibitory effect of word length ($F(1, 5504) = 251.42, p < 0.0001$, after partialling out the effect of word frequency), and a facilitatory effect of family size ($F(1, 5504) = 38.27, p < 0.0001$, after partialling out the effects of word frequency and word length), with a significant interaction between frequency and word length ($F(1, 5504) = 102.23, p < 0.0001$).

A by-item multiple regression with log reaction times as the dependent variable and logarithmic word frequecy, word length, and logarithmic family size similarly showed a facilitatory main effects of word frequency ($F(1, 155) = 151.05, p < 0.0001$), an inhibitory main effect of word length ($F(1, 155) = 46.72, p < 0.0001$, after the effect of word frequency), and and a facilitatory effect of family size ($F(1, 155) = 7.04, p = 0.0088$, after partialling out the effects of word frequency and word length), and a significant interaction between frequency and word length ($F(1, 155) = 20.13, p < 0.0001$).

The interaction between frequency and word length is similar to the one found in Experiment 1 between frequency and family size ratio in Hebrew, as shown by Figure 2.2 with the effect of word length being more pronounced in low frequency words.

Summing up, just as for Hebrew, the analysis of the Dutch data set revealed strong facilitatory effects of written word frequency and family size, i.e., frequent words and words with large morphological families elicit shorter response latencies. Additionally, in Dutch, we also found an inhibitory effect of word length, i.e., longer words elicit longer response latencies.

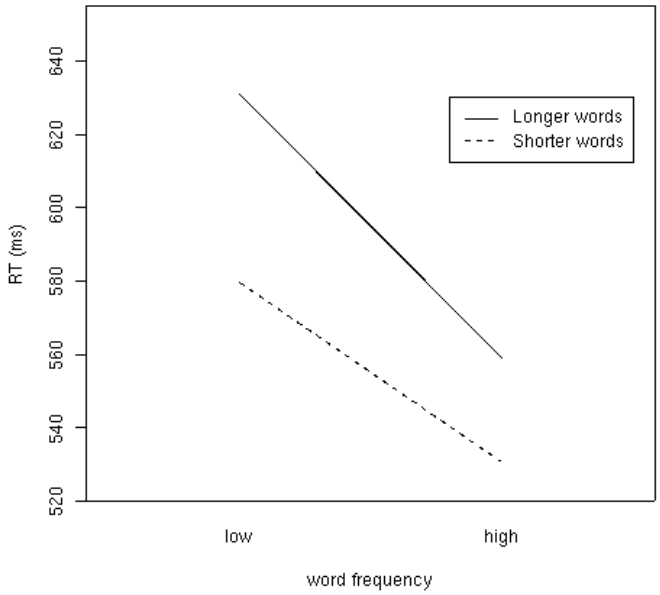

Figure 2.2: Interaction of word frequency and word length in Experiment 2 (factorization by median split).

Cross-language Analyses

We now consider the similarities and differences between family size, word length, and frequency effects in Hebrew and Dutch. Recall that Bates and her colleagues report that picture naming latencies in several unrelated languages can be predicted equally well from frequency counts from the same language as from frequency counts of translation equivalents in another (unrelated) language. The question to be addressed now is whether a similar symmetry in cross-language predictivity is present in our comprehension data for Hebrew and Dutch.

In what follows, we first study the correlations between our independent variables in both experiments. This will allow us to gauge to what extent the frequency, family size and word length counts in the two languages interrelate. We continue by assesing the individual predictive power of each of these variables on the response latencies in both experiments. Finally we present multiple regression analyses in which we assess whether each of these variables has additional predictive power on the response latencies, after having partialled out the effects of the within language counts that were found in the within language analyses.

First, however, recall that the Hebrew and Dutch frequency counts are based on corpora of different size, 200 million words for Hebrew and 42 million words

for Dutch. A preliminary question is to what extent this difference might affect our cross-linguistic comparisons. First consider the frequency counts. Baayen, Moscoso del Prado, Schreuder, and Wurm (2003) show that under the lognormal model (Carroll, 1967, Baayen, 2001) a change in corpus size amounts to a change in the intercept for the regression of frequencies in the one language on the frequencies in the other language. Since our analyses make use of log-transformed frequencies, the change in corpus size has no qualitative effect on our regression models. However, there will be more sampling error in the Dutch counts (which are based on the smaller corpus), but this may be offset by better morphological analyses available for Dutch than for Hebrew, resulting in more accurate family size counts. For Hebrew, the family size counts are based on the knowledge of a native speaker combined with a contrastive analysis of several dictionaries. For Dutch, the counts are based on an exhaustive resource: CELEX. Thus, for both languages, we have used the best available estimates of the family size counts.

Table 2.3: Pairwise Pearson correlation coefficients between the logarithmic frequency, word length, and family size counts. The correlations marked with an asterisk are those that are significant at the 5% level.

	Freq. Hebrew	Freq. Dutch	Length Hebrew	Length Dutch	Fam.S. Hebrew	Fam. S. Ratio Hebrew
Fam. S. Dutch	*0.40	*0.63	-0.15	*-0.50	*0.33	*0.17
Fam. S. Ratio Hebrew	0.06	0.07	-0.04	-0.02	0.04	
Rel. Fam. S. Hebrew	*0.24	*0.21	0.11	-0.11		
Length Dutch	*-0.31	*-0.44	*0.30			
Length Hebrew	-0.14	*-0.18				
Freq. Dutch	*0.46					

We begin with an analysis of the correlational structure of the Hebrew and Dutch frequency, word length, and family size measures. Table 2.3 summarizes the correlations between the logarithmic frequency counts, word length, and logarithmic family size in Hebrew and Dutch. We find an asymmetry in the correlations between Dutch family size and the word frequencies in both languages. On the one hand, the correlation of Dutch family size and Dutch frequency ($r = 0.63$) is significantly stronger ($Z = 2.88, p = 0.0020$) than the correlation between Dutch family size and Hebrew frequency ($r = 0.40$). On the other hand, the correlations of Dutch or Hebrew frequency with Hebrew related family size do not differ significantly ($Z = 0.34, p = 0.3657$). In fact, the correlation between Dutch family size and Hebrew frequency ($r = 0.40$) is significantly stronger ($Z = 1.72, p = 0.0428$) than the correlation between Hebrew related family size and Hebrew frequency ($r = 0.24$).

With respect to word length, we observed a clear pattern. Dutch word length correlates better than Hebrew word length both with Hebrew frequency ($Z = 1.74, p =$

0.0409) and with Dutch Frequency ($Z = 2.81, p = 0.0024$), with the correlation between Hebrew length and Hebrew frequency not reaching significance levels ($r = -0.14, p = 0.08$).

Table 2.4: Pairwise Pearson correlation coefficients between the frequency, word length, and family size measures in both languages, with mean logarithmic response latencies from both experiments. The correlations marked with an asterisk are those that are significant at the 5% level.

	RT Hebrew	RT Dutch
Frequency (Hebrew)	*-0.60	*-0.44
Frequency (Dutch)	*-0.39	*-0.65
Related Family Size (Hebrew)	*-0.26	*-0.26
Family Size Ratio (Hebrew)	*-0.21	-0.14
Family Size (Dutch)	*-0.36	*-0.60
Word Length (Hebrew)	0.14	*0.17
Word Length (Dutch)	* 0.30	*0.60

Table 2.4 summarizes the Pearson correlations between the frequency and family size counts and the response latencies in the two languages. First consider the correlations between the frequency measures and the reaction times in both languages. Note that there is an asymmetry between the correlation of frequency and response latencies. The correlations of Hebrew frequency and Hebrew RT ($r = -0.60$) and Dutch frequency and Dutch RT ($r = -0.65$) are not different according to Fisher's Z-transformation for comparing correlation coefficients ($Z = 0.62, p = 0.2593$). By contrast, Dutch frequency shows a correlation with Hebrew RT ($r = -0.39$) that is significantly smaller ($Z = 2.57, p = 0.0051$, according to Downie and Heath (1965)'s method for comparing correlation coefficients from non-independent samples) than the correlation of Hebrew frequency and Hebrew RT ($r = -0.60$). Similarly the correlation between Hebrew frequency and Dutch RT ($r = -0.44$) is significantly reduced ($Z = 2.61, p = 0.0051$) when compared with the correlation between Dutch frequency and Dutch response latencies ($r = -0.65$). In other words, correlations between frequency and reaction times are weaker across languages than within languages. This pattern differs from that reported by Bates et al. (2003) for the picture naming paradigm: They observed equal cross and within-language correlations.

A pattern different from that observed for the frequency-RT correlations emerges in the correlations between family sizes and reaction times, as can be seen in the central sector of Table 2.4. The correlation of Dutch family size with Dutch response

latencies ($r = -0.60$) is significantly stronger ($Z = 4.05, p < 0.0001$) than that of Hebrew family size ratio with Hebrew response latencies ($r = -0.21$), (this is also true for Hebrew related family size and Hebrew reaction times ($r = -0.26, Z = 3.55, p = 0.0002$) or Hebrew total family size ($r = -0.16, Z = 4.42, p < 0.0001$)). At the same time, the correlation between Dutch family size and Hebrew response latencies (-0.36) is slightly greater than the correlation between Hebrew family size ratios and Hebrew response latencies (-0.21). Although this difference is only marginally significant ($Z = 1.55, p = 0.06$), it is interesting since it indicates that Dutch family size is at least as good a predictor of Hebrew response latencies, as any of the Hebrew family size counts.

Word length (see the bottom rows of Table 2.4) shows a similar pattern to that of family size. The correlation between Dutch word length with Dutch reaction times ($r = 0.60$) is significantly greater ($Z = 4.65, p < 0.0001$) than that between Hebrew word length and Hebrew reaction times ($r = 0.14$). At the same time, the correlation between Dutch word length and Hebrew reaction times ($r = 0.30$) is greater ($Z = 1.52, p < 0.05$) than the correlation between Hebrew word length and Hebrew reaction times ($r = 0.14$). This indicates that Dutch word length emerges as a better predictor of response latencies in both languages.

Summing up, Hebrew family size correlates less well with frequencies and reaction times in both Hebrew and Dutch than does Dutch family size. This is probably due to a difference in variance. The range of logarithmic Dutch family size $[0, 4.82]$ is larger than the range of logarithmic Hebrew family size $[0, 3.37]$, with a greater variance in Dutch ($\hat{\sigma} = 1.36$) than in Hebrew ($\hat{\sigma} = 0.57; F(149, 149) = 2.40, p < 0.0001$). This gives the Dutch family size counts an a-priori advantage in predictive power. This is also true for word length. The range of word lengths in Dutch $[3, 18]$ ($\hat{\sigma} = 2.43$), is greater than the range of Hebrew length ($[3, 7]$), with a significantly smaller variance ($\hat{\sigma} = 0.89; F(149, 149) = 2.7303, p < 0.0001$).

We now proceed to ascertain whether frequency, word length and family size are significant independent predictors of the response latencies in the other language. More precisely, the question is whether frequency, word length, and family size counts from one language, explain variance in the response latencies in the other language, after having partialled out the within language variables. For instance, we may ask whether Dutch frequency, family size and word length predict Hebrew response latencies, after having partialled out Hebrew frequency, word length and family size.

For the by-participant analysis, we added log Hebrew frequency, Hebrew word

length, and log Hebrew related family size, one at time, as predictors to the by-participant regression fit to the Dutch data in Experiment 2. Sequential analyses of variance revealed additional significant main effects of Hebrew frequency ($F(1, 5220) = 8.30, p = 0.0040$), and Hebrew related family size ($F(1, 5220) = 13.26, p = 0.0003$), without any significant effect for Hebrew word length ($F < 1$) or family size ratio ($F(1, 5220) = 1.96, p = 0.16$).

For the by-item analysis, we similarly added log Hebrew word frequency, Hebrew word length, and Hebrew family size ratio (one at a time) as predictors to the by-item regression model that was fit to the Dutch data in Experiment 2. Sequential analyses of variance revealed an additional significant facilitatory effect of Hebrew related family size ($F(1, 144) = 4.61, p = 0.0335$), and did not reveal any additional main effects of log Hebrew word frequency ($F(1, 144) = 2.47, p = 0.1182$), Hebrew word length ($F < 1$) or family size ratio ($F < 1$).

What these analyses show are a strong facilitatory effect of Hebrew related family size on Dutch response latencies and an indication of a similar effect of word frequency. Crucially, we do not find an inhibitory effect of the number of unrelated family members in Hebrew on Dutch reaction times. This is in line with our interpretation of the family size effect in Hebrew as a complex effect with an inhibitory and a facilitating component. The facilitating component is not language specific. Due to the translation equivalence between Hebrew and Dutch target words, sizes of the morphological families in the two languages are correlated. This is so because both family size counts provide an estimation of the size of the semantic neighborhood around a word. Consequently, Hebrew related family size predicts Dutch response latencies, even after partialling out the effect of Hebrew family size. The inhibitory component is specific to Hebrew. As a Hebrew speaker activates all words sharing a given root, the unrelated family members are activated along with the related family members. We have seen that the related family members give rise to facilitation in the response latencies, and that the unrelated family members give rise to inhibition. This inhibitory effect does not arise in Dutch because the unrelated family members of the Hebrew target words fall completely outside the morphological families of their Dutch translation equivalents.

In order to study the predictive power of Dutch word frequency, word length and family size on Hebrew response latencies in more detail, we added the Dutch family counts and frequencies from Experiment 2 as predictors for the Hebrew response latencies of Experiment 1.

By-participant multilevel regression analyses revealed additional significant main

effects of Dutch frequency ($F(1, 3715) = 10.63, p = 0.0011$), Dutch word length ($F(1, 3715) = 9.75, p = 0.0018$), and Dutch family size ($F(1, 3715) = 11.07, p = 0.0009$) in sequential analyses of variance, i.e., after having partialled out the Hebrew predictors.

By-item regression analyses revealed additional main effects of log Dutch word frequency ($F(1, 144) = 3.92, p = 0.0497$), and only marginally significant effects of Dutch family size ($F(1, 144) = 2.91, p = 0.0902$), and Dutch word length ($F(1, 144) = 3.62, p = 0.0592$) in sequential analyses of variance, under conservative two-tailed testing; as predicted, the cross-language effect of Dutch family size on Hebrew response latencies is also facilitatory, as it was the case in the within language analyses.

These analyses show that frequency is a significant predictor across the two languages. In addition, related family size in Hebrew and family size in Dutch emerge as cross-language predictors. We interpret these two family size measures as capturing cross-language similarities in semantic space. Note that the morphology of Hebrew introduces an asymmetry for the Hebrew homonymic roots, with cross-language predictivity for the related family size but not for the unrelated family size.

General Discussion

In this study we have addressed two main questions. First, is family size in Hebrew determined only by morphological form or also by meaning? Second, do the conceptual structures that give rise to the family size effect overlap across languages?

In Experiment 1, we used a regression design with frequency, word length, and family size as independent variables, and response latencies and error scores as dependent variables. We observed independent effects of frequency and family size, both for the response latencies and for the error measure. The fact that family size emerged as an independent predictor supports the hypothesis advanced in Feldman, Frost, and Pnini (1995) that the number of words in which a Hebrew root appears might be an important processing variable. In this respect, an interesting new finding is that the effect of family size was different for the non-homonymic roots and the homonymic roots. For non-homonymic roots, the family comprises only the more semantically related words. The count of this number of words is negatively correlated with reaction times. In the case of homonymic roots, the count of all family members, irrespective of whether they are more or less semantically related, does not correlate well with reaction times. However, when we count the

closely semantically related family members and the semantically more distant or unrelated family members separately, both counts correlate with the reaction times and errors, but with opposite signs. The related family size is facilitatory, while the unrelated family size is inhibitory. These different effects of related and unrelated family size counts argue in favor of the conceptual family hypothesis. According to this hypothesis, the family size effect in Hebrew arises at the conceptual level, just as it does in Dutch or English. However, the presence of an inhibitory effect for the unrelated family members suggests that Hebrew speakers are indeed sensitive to the formal aspect of the three-consonantal roots of Hebrew words. This is partially in line with the formal family hypothesis: It suggests that sharing a purely formal root leads to activation of all family members with that root. However, contrary to the formal family hypothesis, the unrelated family members give rise to longer instead of shorter RT's, indicating that Hebrew speakers are also sensitive to the degree of semantic relatedness between the morphologically related words.

At this point, the question arises of whether the inhibitory effect of the unrelated family size arises at the level of orthographic or phonological form, or at the conceptual level. Consider the possibility that it arises at the form level, keeping in mind that the semantic distinction between related and unrelated family size members cannot be operative at the pre-lexical level of form processing by definition. In this case, one would predict that a greater total family size would lead to greater inhibition, as all family members share the same root form. This prediction, however, holds true only for the less related family members, and not for the more related family members for words with homonymic and non-homonymic roots alike. Hence, we must be dealing with a competition effect arising at more central levels of lexical processing. Since the formal aspects of the Hebrew root play a central role in the process of word recognition, each homonymic root being associated with more than one core meaning results in incoherent activation of conflicting semantic fields, leading to competition between them (see, e.g., Gaskell & Marslen-Wilson, 1997, Plaut & Booth, 2000, for more detailed theories of competition at the semantic level).

Experiment 2 investigated the processing in the visual modality of the Dutch translation equivalents of the Hebrew words from Experiment 1. We observed independent effects of written frequency and family size. Interestingly, cross-language comparisons revealed that reaction times for Hebrew words can be predicted from the frequency and family size counts of their Dutch translation equivalents, and vice-versa. Crucially, this is so even after having partialled out the effects of the

within language variables. The predictivity of frequency across languages in visual lexical decision is in line with the results reported by Bates et al. (2003) for picture naming.

Recall that Bates and colleagues report equal predictivity for word frequency within and across languages, whereas we observed asymmetrical cross-language predictivity. We suspect that this is due to the different way in which the stimuli were selected in the different experiments. Bates et al. (2003) used pictures that were selected by their ‘potential cross-cultural validity’. In other words, their pictures represent concepts which are relatively common across the seven languages covered in their study. In contrast, our datasets contain words without any such restriction, covering a wide range of frequencies. Thus, our cross-linguistic comparison is more susceptible to cultural-specific differences in the usage (and perhaps meaning) of translation equivalents.

Whereas the cross-linguistic predictivity of word frequency provides only (inconclusive) evidence for the similarity and salience of concepts across languages, the cross-language predictivity of family size documented in the present study provides unambiguous evidence for a surprising degree of isomorphism in the way concepts are organized in the Hebrew and Dutch mental lexicon. Although these two languages are genetically unrelated and make use of radically different means for creating morphologically complex words, the networks of morphologically related words are similar enough to allow response latencies in the one language to be predicted from the network size in the other language. Thus, the family size effect emerges as an excellent tool for mapping the degree of isomorphism in the conceptual relations in the Hebrew and Dutch mental lexicons.

References

Alegre, M. and Gordon, P.: 1999, Frequency effects and the representational status of regular inflections, *Journal of Memory and Language* **40**, 41–61.

Aronoff, M.: 1994, *Morphology by itself: stems and inflectional classes*, The MIT Press, Cambridge, Mass.

Baayen, R. H.: 2001, *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht.

Baayen, R. H., Dijkstra, T. and Schreuder, R.: 1997, Singulars and plurals in Dutch: Evidence for a parallel dual route model, *Journal of Memory and Language* **37**, 94–117.

Baayen, R. H., Lieber, R. and Schreuder, R.: 1997, The morphological complexity of simplex nouns, *Linguistics* **35**, 861–877.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX lexical database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baayen, R. H., Tweedie, F. J. and Schreuder, R.: 2002, The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon, *Brain and Language* **81**, 55–65.

Bates, E., D'Amico, S., Jacobsen, T., Szekely, A., Andonova, E., Devescovi, A., Herron, D., Ching Lu, C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutiérrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., Tzeng, A. and Tzeng, O.: in press, Timed picture naming in seven languages, *Psychonomic Bulletin and Review*.

Becker, C. A.: 1979, Semantic context and word frequency effects in visual word recognition, *Journal of Experimental Psychology: Human Perception and Performance* **5**, 252–259.

Belsley, D. A., Kuh, E. and Welsch, R. E.: 1980, *Regression Diagnostics. Identifying Influential Data and sources of Collinearity*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Berman, R.: 1978, *Modern Hebrew Structure*, University Publishing Projects, Tel Aviv.

Bertram, R., Schreuder, R. and Baayen, R. H.: 2000, The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity, *Journal of Experimental Psychology: Learn-*

ing, Memory, and Cognition **26**, 419–511.

Bolle, M. and Pimentel, J.: 1984, *Woordenboek Nederlands Hebreeuws*, Strengholt's Boeken, Naarden, The Netherlands.

Booij, G. E.: 2002, *The morphology of Dutch*, Oxford University Press, Oxford.

Borowsky, R. and Besner, D.: 1993, Visual word recognition: A multistage activation model, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**, 813–840.

Bradley, D. C. and Forster, K. I.: 1987, A reader's view of listening, *Cognition* **25**, 103–134.

Carroll, J. B.: 1967, On sampling from a lognormal model of word frequency distribution, *in* H. Kučera and W. N. Francis (eds), *Computational Analysis of Present-Day American English*, Brown University Press, Providence, pp. 406–424.

De Jong, N. H.: 2002, *Morphological Families in the Mental Lexicon*, MPI Series in Psycholinguistics, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M. and Baayen, R. H.: 2002, The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects, *Brain and Language* **81**, 555–567.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: 2000, The morphological family size effect and morphology, *Language and Cognitive Processes* **15**, 329–365.

Deutsch, A., Frost, R. and Forster, K. I.: 1998, Verbs and nouns are organized and accessed differently in the mental lexicon: Evidence from Hebrew, *Journal of Experimental Psychology: Learning, Memory and Cognition* **24**, 1238–1255.

Deutsch, A., Frost, R., Pollatsek, A. and Rayner, K.: 2000, Early morphological effects in word recognition in hebrew: Evidence from parafoveal preview benefit, *Cross-linguistic perspectives on morphological processing* **15**(4-5), 487–506.

Downie, N. M. and Heath, R. W. (eds): 1965, *Basic Statistical Methods*, Harper and Row, New York.

Feldman, L. B., Frost, R. and Pnini, T.: 1995, Decomposing words into their constituent morphemes: Evidence from English and Hebrew, *Speech research* **21**, 235–254.

Feldman, L. B. and Siok, W. W. T.: 1997, The role of component function in vi-

sual recognition of Chinese characters, *Journal of Experimental Psychology: Learning, Memory and Cognition* **23**, 778–781.

Feldman, L. B. and Siok, W. W. T.: 1999, Semantic radicals contribute to the visual identification of Chinese characters, *Journal of Memory and Language* **40**, 559–576.

Feldman, L. B. and Soltano, E. G.: 1999, Morphological priming: The role of prime duration, semantic transparency and affix position, *Brain and Language* **68**(1-2), 33–39.

Frost, R., Deutsch, A. and Forster, K.: 2000, Decomposing morphologically complex words in a nonlinear morphology, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26**(3), 751–765.

Frost, R., Deutsch, A., Gilboa, O., Tannenbaum, M. and Marslen-Wilson, W. D.: 2000, Morphological priming: Dissotiation of phonological, semantic and morphological factors, *Memory and Cognition* **28**(8), 1277–1288.

Frost, R., Forster, K. I. and Deutsch, A.: 1997, What can we learn from the morphology of Hebrew? A masked-priming investigation of morphological representation, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **23**, 829–856.

Hasher, L. and Zacks, R. T.: 1984, Automatic processing of fundamental information. The case of frequency of occurrence, *American Psychologist* **39**, 1372–1388.

Landauer, T. and Dumais, S.: 1997, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge, *Psychological Review* **104**(2), 211–240.

Lorch, R. F. and Myers, J. L.: 1990, Regression analyses of repeated measures data in cognitive research, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**, 149–157.

Lüdeling, A. and De Jong, N. H.: 2002, German particle verbs and word-formation, *in* N. Dehé, R. Jackendoff, A. McIntyre and S. Urban (eds), *Verb-particle explorations*, Mouton de Gruyter, Berlin, pp. 315–333.

Marslen-Wilson, W. D., Tyler, L. K., Waksler, R. and Older, L.: 1994, Morphology and meaning in the English mental lexicon, *Psychological Review* **101**, 3–33.

McCarthy, J. J.: 1981, A prosodic theory of non-concatenative morphology, *Linguistic Inquiry* **12**, 373–418.

Morton, J.: 1969, The interaction of information in word recognition, *Psychological Review* **76**, 165–178.

Pinheiro, J. C. and Bates, D. M.: 2000, *Mixed-effects models in S and S-PLUS*, Statistics and Computing, Springer, New York.

Plaut, D. C. and Booth, J. R.: 2000, Individual and developmental differences in semantic priming: Empirical and computational support for a single mechanism account of lexical processing, *Psychological Review* **107**, 786–823.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C. and Gremmen, F.: 1999, How to deal with 'the language-as-fixed-effect-fallacy': Common misconceptions and alternative solutions, *Journal of Memory and Language* **41**, 416–426.

Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **37**, 118–139.

Schweika, Y.: 1997, *Modern Hebrew Dictionary*, Rav-Milim Matach, Tel-Aviv.

Stanovich, K. E. and West, R. F.: 1981, The effect of sentence context on ongoing word recognition: Test of a two-process theory, *Journal of Experimental Psychology: Human Perception and Performance* **7**, 658–672.

Taft, M.: 1979, Recognition of affixed words and the word frequency effect, *Memory and Cognition* **7**, 263–272.

Family Size in a Morphologically Rich Language: Finnish

CHAPTER 3

This chapter has been submitted as Fermín Moscoso del Prado Martín, Raymond Bertram, Tuomo Häikiö, Robert Schreuder, and R. Harald Baayen: Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew.

Abstract

Finnish is a language with a very rich morphology. The number of different complex words in which a given stem can appear as a constituent, the morphological family size, mounts to several thousands in Finnish, while Dutch or English stems do not appear in more than 500 complex words at most, and a Hebrew root occurs in a maximum of 30 words. This study addresses the processing consequences of the large morphological family sizes in Finnish as compared to Dutch and Hebrew by means of a visual lexical decision experiment. We observed that in Finnish, words with larger morphological families elicited shorter response latencies. However, for complex Finnish words, it is not the complete morphological family of a complex word that codetermines response latencies, but only the set of words containing the complex word itself as a constituent, its dominated family size. Comparisons with parallel experiments using translation equivalents in Dutch and Hebrew show substantial cross-language predictivity of family size between Finnish and Dutch, but not between Finnish and Hebrew. We think that the absence of cross-language predictivity for Finnish and Hebrew is due to the greater distance in the way their morphological systems contribute to the semantic organization of concepts in the mental lexicon.

Introduction

In languages such as English and Dutch, stems differ in their productivity. Some stems give rise to a great many complex words. For instance, in English, the stem *man* appears in nearly 200 complex words. Other stems hardly ever give rise to complex words, e.g., the noun *scythe*, which only has its verbal conversion alternant as morphological relative. Previous research has shown that the morphological family size of a stem, defined as the number of different complex words in which the stem appears as a constituent, is a robust predictor of response latencies in tasks such as visual lexical decision, auditory lexical decision, and subjective familiarity rating. Words with a larger morphological family size elicit shorter response latencies and higher subjective familiarity scores than do words with smaller family sizes matched for frequency (Schreuder & Baayen, 1997).

The effect of morphological family size is well-established for Germanic languages (Dutch: Schreuder & Baayen, 1997, Bertram, Baayen, & Schreuder, 2000, De Jong, Schreuder, & Baayen, 2000; English: De Jong, Feldman, Schreuder, Pastizzo, & Baayen, 2002; German: Lüdeling & De Jong, 2002). Recently, an effect of morphological family size has also been established for a Semitic language, Hebrew (Moscoso del Prado Martín, Deutsch, Frost, Schreuder, De Jong, & Baayen, 2003). In this language, morphological family size is defined in terms of the number of words that share a given consonantal root. The morphological family size in Hebrew ranges between 1 and 30, and is therefore much more restricted than English family sizes (range 1 to 200) and Dutch family sizes (range 1 to 550). Even though morphological families tend to be small in Hebrew, morphological family size emerged as a reliable predictor of response latencies independently of word frequency.

The family size effect is semantic in nature (Bertram et al., 2000; De Jong, 2002). Recent evidence supporting this conclusion has been obtained in Hebrew and in the bilingual lexicon. The Hebrew family size effect has a specific property that is particular to the Hebrew root, namely, that for words with homonymic roots the semantically related family members lead to facilitation while the semantically unrelated family members give rise to inhibition. In Dutch, such an effect has not been observed for homonymic stems (De Jong, 2002). However, a similar effect has been observed for interlingual homographs in the Dutch-English bilingual lexicon (Dijkstra, Moscoso del Prado, Schulpen, Schreuder, & Baayen, 2003). When Dutch bilinguals performed a Dutch visual lexical decision task, the Dutch family members of an interlingual homograph gave rise to facilitation while the English family

members gave rise to inhibition. These results show unambiguously that the effect of the morphological family size arises at semantic levels of lexical processing.

In this study, we report an experiment addressing the possible existence of a family size in Finnish. Finnish belongs to the Finno-Ugric language family, and is well known for its rich and complex morphology. It combines a complex inflectional system with a great many cases with productive derivation involving rampant stem allomorphy and very productive compounding. In Finnish, a stem such as *työ*, 'work', has roughly 7000 family members, including *työntekijä*, 'employee', *työehtosopimus*, 'wage rate treaty', *työstökone*, 'machine tool', *työläs*, 'laborious', and *työväenluokka*, 'working class'. Obviously, most Finnish stems have smaller morphological families, but many are very sizeable anyway with family sizes of some two hundred words or more.

While the Hebrew study established that family size effects generalize from Germanic concatenative morphology to Semitic non-concatenative morphology, the present study investigates whether family size effects also exist in the agglutinative morphology of Finnish. It is far from evident that this would be the case. Just as the word frequency effect, the family size effect is logarithmic in nature. Robust effects are typically observed in the range of 0–40 family members, after which we generally have a floor effect. Given the large families counted for Finnish stems, no effect of family size might be observed due to an overall floor effect. As we will show below, this prediction is partially correct, requiring a more limited family size definition for complex words.

From the study on family size effects in Hebrew (Moscoso del Prado et al., 2003) we know that Hebrew response latencies can be predicted from the Dutch family sizes of the corresponding translation equivalents even after Hebrew frequency and Hebrew family size are partialled out, and vice versa. This result shows that there is substantial similarity in semantic lexical organization in Dutch and Hebrew, even though these languages are typologically fundamentally different. In this study we address the question of whether a similar cross-language predictivity might be observed for Finnish and Dutch translation equivalents, and for Finnish and Hebrew translation equivalents. The patterns of cross-language predictivity have important implications for the degree of isomorphy in semantic organization across languages with typologically different morphological systems.

We performed a new visual lexical decision experiment that addresses the questions raised above. It is designed along the lines of the Hebrew and Dutch experiments reported by Moscoso del Prado et al. (2003), and makes use of translation

equivalents of the Hebrew and Dutch words used in that study.

Experiment

Method

Participants. Twenty-six undergraduate students of the University of Turku participated in the experiment. All were native speakers of Finnish and had normal or corrected-to-normal vision.

Materials. The materials of these studies are the translation equivalents of the Hebrew and Dutch words used in the experiments reported in Moscoso de Prado et al. (2003). As our point of departure, we took the 162 Dutch words from their Experiment 2, and translated them into Finnish. The translations were done using a Dutch-Finnish Dictionary (Suomi-Hollanti-Suomi taskusanakirja, Porvoo: WSOY, 1992), and they were extensively validated by the second and third author. When a word had different possible translations into Finnish with different meanings, we excluded that word from the Experiment. Four of the original Dutch words could only be translated into Finnish using multi-word utterances, and were excluded from the experiment as well. In this way we obtained a set of 143 Finnish words.

Frequency counts for these words are based on the unpublished computerized Turun Sanomat Finnish newspaper corpus of 22.7 million word forms accessed with the help of the WordMill database program of Laine and Virtanen (1999). Morphological family size counts were also based on this database, with each of the potential family members evaluated by the third author, in some occasions aided by a dictionary (Nykysuomen sanakirja, Porvoo: WSOY, 1978). Each of these words was paired with a pseudo-word whose phonotactics did not violate the phonology of Finnish. Twenty practice trials, ten words and ten pseudo-words were run before the actual experiment. We constructed three different permutations and their corresponding reversed versions of the original word list for counterbalancing. Table 3.1 provides a summary of the distributional properties of the data set.

Procedure. Participants were tested in noise-attenuated experimental rooms. They were asked to decide as quickly and accurately as possible whether the letter string appearing on the computer screen was a real Finnish word. Following a pause after the test trials, the experiment was run with two further pauses, dividing the

experiment in three blocks, each containing one third of the materials. Items were preceded by a fixation mark in the middle of the screen for 500 ms. After 500 ms, the stimulus appeared at the same position. Stimuli were presented on NEC Multisync color monitors in white lowercase 21 point Arial letters on a dark background and they remained on the screen for 1500 ms. The maximum time span allowed for a response was 2000 ms. from stimulus onset.

Results and Discussion

All participants in this experiment performed with an error rate of less than 15%. One item elicited errors for more than 30% of the participants, and was excluded from the analyses. Additionally, we excluded from the analyses four items that elicited response latencies of more than two and a half standard deviations above or below the mean.

Table 3.1: Medians, means, standard deviations, and ranges for the different counts and response latencies in Experiment 1, after excluding the five outliers.

	median	mean	standard deviation	range
frequency	670	3,155	7,097	1–56,193
word length	7	7.2	2.3	3–14
family size	298	620	892	8–6,029
dominated family size	88	273	485	0–3,080
non-dominated family size	29	347	762	0–5,835
response latency	604 ms	617 ms	63 ms	530–808 ms

Table 3.1 provides the medians, means, standard deviations, and ranges for the frequency, family size, and word length counts for this data set, and the average response latencies in the experiment after excluding the four outliers.

A by-participant multilevel regression model (Lorch & Myers, 1990; Alegre & Gordon, 1999; Pinheiro & Bates, 2000) fit to the dataset, with log response latency as dependent variable and log frequency, log family size, and word length as independent variables revealed a facilitatory main effect for word frequency ($F(1, 3625) = 521.86, p < 0.0001$), an inhibitory main effect of word length ($F(1, 3625) = 137.66, p < 0.0001$, after partialling out the effect of frequency), and a facilitatory main effect of family size ($F(1, 3625) = 24.62, p < 0.0001$), after partialling out the effects of frequency and word length. We also observed a significant interaction between word length and word frequency ($F(1, 3625) = 89.21, p < 0.0001$), after partialling out the

main effects.

A by-item linear regression with log reaction time as the dependent variable and log word frequency, word length, and log family size as the independent variables confirmed the effects of word frequency ($F(1, 134) = 85.45, p < 0.0001$), word length ($F(1, 134) = 15.05, p = 0.0002$, after partialling out the effect of word frequency), and family size ($F(1, 134) = 10.88, p = 0.0012$, after partialling out the effects of word frequency and word length). Again, the interaction between word frequency and word length was significant ($F(1, 134) = 4.85, p = 0.0294$).

These results document, for the first time, the presence of a morphological family size effect in Finnish. As in English, German, and Dutch, and as in Hebrew, words with larger families give rise to shorter response latencies than words with smaller families. The presence of a morphological family size effect in three genetically unrelated language families, Indo-European, Hamo-Semitic, and Finno-Ugric, shows that, across typologically very different morphological systems, the paradigmatic organization of morphologically related words is an important factor in lexical processing.

Thus far, it would seem that the possibility we considered in the introduction, namely, that the large family sizes of Finnish compared to English or Dutch would lead to a floor effect, is not borne out. However, consider a selection of the members of the morphological family of *kirja*, ‘book’ in Finnish:

kirja	book
väitöskirja	dissertation
muistikirja	notebook
päiväkirja	diary, notebook
romaanikirjallisuus	novel literature
aikakauskirja	journal
kirjasto	library
lainakirjasto	public library
kirjastonhoitaja	librarian
kirjoitus	writing
kirjoitusjärjestelmä	writing system
kirjepaino	paper weight
kirjailija	author
kirjailijantoiminta	authorship
asiakirja	document
kirjailla	embroider

kirjoittaa	write
kirje	letter (a written communication)
kirjain	letter (the symbol)
kirjeenkantaja	postman
kirjeenvaihtaja	correspondent
kirjeenvaihtotoveri	pen-pal
kirjoittautua	register
kirjoituskone	typewriter

Note that while there is a family member that has a translation in English that contains the stem *book* (notebook), all other family members require translations with quite different stems in English, ranging from *author* to *library* and from *register* to *dissertation*. Note furthermore that some family members form semantically cohesive clusters, such as the words for *library*, *librarian*, and *public library*. This suggests the possibility that the family size effect in Finnish might be carried predominantly or perhaps even exclusively by the semantically more closely related family members.

One way of obtaining an objective and replicable way of defining the notion of being more closely related semantically, is to make a distinction between the family members of a word that are its direct descendants (its dominated family) and the other family members (its non-dominated family). Figure 3.1 illustrates the distinction between the dominated and non-dominated family size for the English family of *interchangeable*. The dominated family size of *interchangeable* consists of the words *interchangeability* and *interchangeably*. Its non-dominated family size consists of words such as *interchange* and *change*, as well as all the other words that are dominated by *change* and not by *interchange*, such as *moneychanger* and *changelessly*. Note that the dominated family members are more closely related in meaning to each other than is the case for the non-dominated family members. This leads to the hypothesis that in Finnish, the morphological family size might be carried predominantly or perhaps exclusively by the dominated family size.

In order to test this hypothesis, we selected the 83 complex Finnish words in our dataset. (We excluded the simplex words from this analysis, as for simplex words the family size as a whole is identical to the dominated family size, the non-dominated family size being the empty set.) For these complex words, we determined the dominated and non-dominated family size. We then carried out a regression analysis, with log word frequency, word length, log dominated family size,

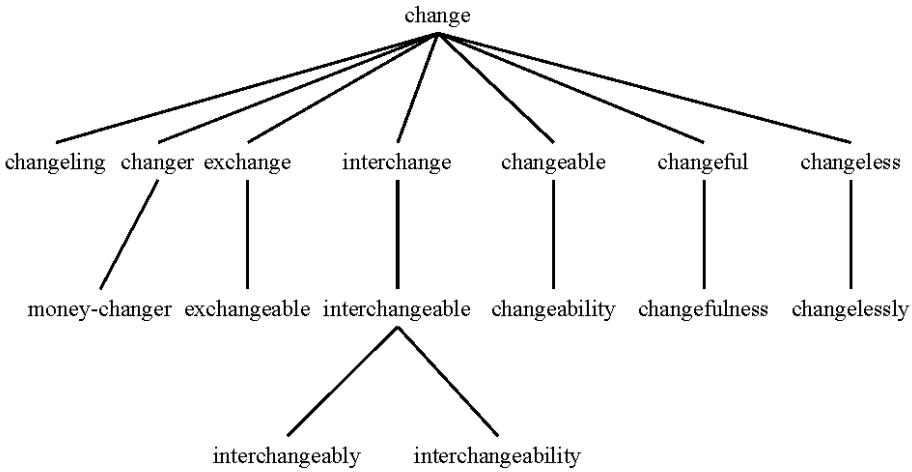

Figure 3.1: The position of *interchangeable* in the family of *change*.

and log non-dominated family size as independent variables, and log response latencies as the dependent variable. A by-participant multilevel regression analysis revealed a highly significant effect for the dominated family size ($F(1, 2127) = 20.25, p < 0.0001$) and no effect whatsoever for the non-dominated family size ($F < 1$). A by-item regression analysis revealed the same pattern of results ($F(1, 78) = 8.42, p = 0.0048$ for the dominated family size, $F < 1$ for the non-dominated family size). In fact, it turns out that the total family size is not a good predictor for the complex words in our data (both F values < 1). This shows that adding the non-dominated family members to the family size count for complex words in Finnish amounts to adding so much noise that the effect of the true predictor, the dominated family size, is completely masked.

The non-existence of a family size effect for the non-dominated family is partly in line with the intuition outlined in the introduction that with the large family sizes of Finnish the family size effect might be reduced due to a floor effect. However, restriction of the effect to the dominated family suggests that the degree of semantic relatedness in the family might be the key determinant rather than size as such. To gain further insight into the weight of these two factors, the magnitude of the family on the one hand, and its semantic cohesion on the other, we reanalysed the Dutch analogue of the present experiment reported in Moscoso del

Prado et al. (2003), in which the translation equivalents of the Finnish words studied in the present paper were analysed. From their Dutch items, we selected the 59 words that were morphologically complex. A multilevel by-subject regression model revealed significant effects of both dominated and non-dominated family size, although the beta weight for the dominated family size ($\hat{\beta} = -0.085$, standard error $= 0.017, t(2018) = -4.875, p < 0.0001$) was more than twice as large as the beta weight of the non-dominated family size ($\hat{\beta} = -0.030$, standard error $= 0.011, t(2018) = -2.790, p = 0.0053$).

This result suggests that the dominated family size is the prime carrier of the family size, but that the non-dominated family size may also have some predictive power, at least in Dutch. This is probably due to the relatively small sizes (at least compared to Finnish) of morphological families in Dutch. Within these small families, there is enough semantic similarity between the non-dominated and the dominated family members to allow a non-dominated family size effect to emerge. In Finnish, by contrast, the range of meanings covered by the non-dominated family is too broad, leading to semantic neighborhoods that are too sparsely populated to give rise to a measurable family size effect in the response latencies.

At this point, it should be made explicit that we do not claim that the distinction between the dominated and the non-dominated family is an absolute distinction for Finnish. To the contrary, we believe that closely related non-dominated family members will also contribute to the family size effect. However, we leave it to further research to establish principled ways in which the contributing non-dominated family members might be ascertained.

Summing up, the crucial contribution of the present experiment to our knowledge of the family size effect in human cognition is that by examing the family size effect in a highly productive agglutinative language such as Finnish, the semantic nature of the effect is clarified in more detail. If the family size effect were just a form effect, the distinction between the dominated and non-dominated family should not have been relevant, contrary to fact. This shows that the family size effect depends on the combination of shared morphological form and shared semantics. When the condition of semantic overlap is not met, as for most non-dominated family members in Finnish, those family members no longer contribute to the effect.

Cross-language Analyses

As mentioned in the introduction, Moscoso del Prado et al. (2003) observed that Hebrew response latencies can be predicted from the Dutch family sizes of the corresponding translation equivalents even after Hebrew frequency and Hebrew family size have been partialled out first, and vice versa. This result is indicative of substantial similarity in semantic lexical organization in Dutch and Hebrew, even though these languages are typologically fundamentally different. We now turn to investigate whether a similar cross-language predictivity might be observed for Finnish and Dutch translation equivalents, and also for Finnish and Hebrew translation equivalents. This will allow us to obtain insight in the extent of cross-language predictivity across typologically unrelated languages and its implications for the degree of isomorphy in semantic organization across radically different morphological systems.

For this cross-language multiple regression analysis, we selected those items that elicited less than 30% errors in the three experiments in Hebrew, Dutch, and Finnish. In this way, we obtained a total of 131 items, each with three response latencies. For each word in each of the three languages, we added as predictors length (in letters), word frequency, and morphological family size in that language. The key question of interest is whether length, frequency, and family size of, e.g., Dutch, predict response latencies in Finnish, even after the effects of Finnish frequency, Finnish word length, and Finnish family size, have been partialled out first.

Table 3.2 summarizes the results obtained for the 6 pairwise comparisions (Hebrew to Dutch, Hebrew to Finnish, Dutch to Hebrew, Dutch to Finnish, Finnish to Hebrew, and Finnish to Dutch). When predicting from language A to language B, we took the best regression model fitted to the data from language B as point of departure. For the details of these models for Hebrew and Dutch, the reader is referred to Moscoso del Prado et al. (2003). For the Finnish data, this model incorporates the effects of word frequency, length in letters, and family size: the full family size for the simplex words, and the dominated family size for the complex words. We then investigated, one at a time, whether frequency, length, and family size from language A had additional predictive value for the response latencies in language B. As can be seen in the third row of Table 3.2, Finnish frequency is an excellent predictor of Dutch response latencies, after having partialled out the effect of Dutch length, frequency, and family size. Finnish family size likewise emerged as a highly significant predictor, and even Finnish length turned out to have some predictive value.

Table 3.2: Cross language predictivity of word frequency, word length, and morphological family size between translation equivalents in Hebrew, Dutch, and Finnish, in sequential analyses of variance in multilevel regression analyses. Significance codes are: $^{+}p < 0.1000$, $^{*}p < 0.0500$, $^{**}p < 0.0050$, and $^{***}p < 0.0005$.

		Hebrew	Dutch	Finnish
Hebrew	frequency	-	$F(1, 4603) = 7.39^{**}$	$F(1, 3263) = 20.57^{***}$
	word length	-	$F(1, 4603) = 2.53$	$F(1, 3263) = 11.51^{**}$
	related family size	-	$F(1, 4603) = 20.22^{***}$	$F(1, 3263) = 1.38$
Dutch	frequency	$F(1, 3184) = 15.28^{***}$	-	$F(1, 3263) = 35.99^{***}$
	word length	$F(1, 3184) = 6.17^{*}$	-	$F(1, 3263) = 56.94^{***}$
	family size	$F(1, 3184) = 15.03^{***}$	-	$F(1, 3263) = 18.62^{***}$
Finnish	frequency	$F(1, 3184) = 2.20$	$F(1, 4603) = 15.98^{***}$	-
	word length	$F < 1$	$F(1, 4603) = 4.90^{*}$	-
	family size	$F(1, 3184) = 1.52$	$F(1, 4603) = 16.75^{***}$	-

What Table 3.2 shows is that frequency is an excellent predictor in five out of six cases. The only instance in which frequency fails to have additional predictivity is when Finnish frequency is used to predict Hebrew reaction times. Note that, in turns of stem productivity, the typological distance is greatest between Hebrew and Finnish, with Dutch taking an intermediate position. Family size emerges alongside with word frequency as a remarkable explanatory variable in four out of six cases. The two cases where family size fails as a cross-language predictor is from Finnish family size to Hebrew reaction times and from Hebrew family size to Finnish reaction times. Again, cross-language predictivity breaks down where the typological difference in morphological structure and stem productivity is greatest. Finally, even word length shows some cross-language predictivity. The only language pair for which word length is predictive in both directions is Finnish and Dutch. The small differences in word length in Hebrew seem not to be predictive for Dutch but predictive for Finnish. Conversely, the big differences in word lengths in Finnish emerge as predictive for Dutch but not for Hebrew.

General Discussion

The questions addressed in this study were, first, whether the family size effect might be observed in Finnish, and second, to what extent Finnish might participate in the cross-language predictivity of family size observed for Hebrew and Dutch. As to the first question, a visual lexical decision experiment revealed that, as it is in Germanic languages such as Dutch, English, and German, and in He-

brew (Semitic), the morphological family size is also relevant for lexical processing in Finnish, a Finno-Ugric language. This finding provides further evidence for the cross-linguistic generality of the family size effect.

Earlier studies (De Jong, 2002; Moscoso del Prado et al., 2003) established that the observed effect of the morphological family size probably arises at the level of semantic processing. These studies also established that semantic similarity shared between the family members is crucial for the effect to emerge. Inspection of morphological families in Finnish, however, suggests that the larger families as a whole are semantically fairly diverse. To obtain further insight into the role of semantic similarity, we introduced the notion of the dominated versus the non-dominated family size for complex words. The counts of dominated family members (the semantically more similar morphological descendants of a complex word) turned out to be the crucial predictors for Finnish. A reanalysis of Dutch data showed both dominated and non-dominated family size to be relevant in this language. Given that morphological families in Dutch are both smaller and semantically more cohesive, we argued that this result supported the hypothesis that the family size effect crucially depends on semantic similarity. The operationalization of semantic similarity in terms of dominated versus non-dominated family size is a first objective and replicable operationalization for differentiating between clusters of semantically related words. We leave it to future research to develop more fine-grained operationalizations of semantic relatedness within morphological families.

Following the line of research developed by Bates et al. (2003) for the cross-linguistic predictivity of frequency in picture naming, and the cross-linguistic predictivity of frequency and family size in Moscoso del Prado at al. (2003), we investigated the cross-language predictivity of frequency and family size across Finnish, Dutch, and Hebrew. We observed substantial cross-language predictivity for frequency across the three languages, and more limited cross-language predictivity for word length. This suggests that there is considerable similarity in concept frequency in these languages, and that Zipf's observation that more frequent words tend to be shorter holds to some extent even across unrelated languages. Following Bates et al. (2003), we interpret these results as suggestive of a substantial semantic component to the word frequency effect.

The most important cross-linguistic finding, however, is that the cross-language predictivity of family size is absent when the distance between the morphological systems, as reflected in the degree of stem productivity, becomes very large. Finnish and Hebrew, the languages with the greatest and the smallest stem pro-

ductivity, showed no additional predictivity for family size once the within-language measures (frequency, length, and family size) have been taken into account. This lack of predictivity contrasts markedly with the significant predictivity of family size from Hebrew to Dutch and vice versa. This suggests to us that there is a higher degree of overlap between the semantic organization in the mental lexicon of morphologically related words in Hebrew and Dutch, and in Finnish and Dutch, than there is for Finnish and Hebrew.

Although the cross-language predictivity of family size shows that there may be considerable overlap in semantic organization, in the sense that words in dense morphological neighborhoods tend to have translation equivalents that also have dense morphological neighborhoods, the absence of such predictivity for Finnish and Hebrew shows that there are limits to this cross-language predictivity. To understand why these limits arise, consider, for instance, the consequences of the different degrees of productivity of compounding in Finnish, Dutch, and Hebrew. In Finnish, compounding is extremely productive, in Dutch, it is productive, and in Hebrew, it is marginally productive at best. Thus, complex concepts expressed by compounds in Finnish will have lexical (instead of phrasal) counterparts in Dutch relatively often, but very seldom in Hebrew. In Hebrew, many Finnish words will require phrasal translations. Consequently, the patterns of lexical co-activation in Finnish will resemble the coactivation patterns of their translation equivalents to a much larger degree in Dutch than in Hebrew. If, as has been argued by De Jong et al. (2003), the co-activation of the morphological family members indeed co-determines the semantic percept of a word, then the present results support the Whorfian view of language, according to which language co-determines thought (see, e.g., Boroditsky, 2001). For languages with similar morphologies, the morphology guides thought along similar paths, thereby giving rise to considerable cross-language predictivity of family size. When morphological systems are very different, as for Hebrew and Finnish, the well-worn paths along which morphology leads thought become notably different, as witnessed by the breakdown of the cross-linguistic predictivity of family size for these languages.

References

Alegre, M. and Gordon, P.: 1999, Frequency effects and the representational status of regular inflections, *Journal of Memory and Language* **40**, 41–61.

Bates, E., D'Amico, S., Jacobsen, T., Szekely, A., Andonova, E., Devescovi, A., Herron, D., Ching Lu, C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutiérrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., Tzeng, A. and Tzeng, O.: in press, Timed picture naming in seven languages, *Psychonomic Bulletin and Review*.

Bertram, R., Baayen, R. H. and Schreuder, R.: 2000, Effects of family size for complex words, *Journal of Memory and Language* **42**, 390–405.

Boroditsky, L.: 2001, Does language shape thought? english and mandarin speakers' conceptions of time, *Cognitive Psychology* **43**(1), 1–22.

De Jong, N. H.: 2002, *Morphological families in the mental lexicon*, PhD thesis, University of Nijmegen, Nijmegen, The Netherlands.

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M. and Baayen, R. H.: 2002, The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects, *Brain and Language* **81**, 555–567.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: 2000, The morphological family size effect and morphology, *Language and Cognitive Processes* **15**, 329–365.

Dijkstra, T., Moscoso del Prado Martín, F., Schulpen, B., Schreuder, R. and Baayen, R.: 2003, Family size effects in bilinguals, *Unpublished manuscript, University of Nijmegen*.

Laine, M. and Virtanen, P.: 1999, *WordMill Lexical Search Program*, Center for Cognitive Neuroscience, University of Turku, Finland.

Lorch, R. F. and Myers, J. L.: 1990, Regression analyses of repeated measures data in cognitive research, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**, 149–157.

Lüdeling, A. and De Jong, N. H.: 2002, German particle verbs and word-formation, *in* N. Dehé, R. Jackendoff, A. McIntyre and S. Urban (eds), *Verb-particle explorations*, Mouton de Gruyter, Berlin, pp. 315–333.

Moscoso del Prado Martín, F., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H. and Baayen, R. H.: 2003, Changing places: A cross-language perspective on frequency and family size in Hebrew and Dutch, *Manuscript submitted for*

publication, Max Planck Institute for Psycholinguistics.

Pinheiro, J. C. and Bates, D. M.: 2000, *Mixed-effects models in S and S-PLUS*, Statistics and Computing, Springer, New York.

Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **37**, 118–139.

Morphological Family Size in Bilinguals: Interlingual Homographs

CHAPTER 4

This chapter has been submitted as Ton Dijkstra, Fermín Moscoso del Prado Martín, Béryl Schulpen, Robert Schreuder and R. Harald Baayen: A roommate in cream: Morphological family size effects in bilinguals.

Abstract

In monolingual studies, target word recognition is affected by the number of words that are morphologically related to the target. Larger morphological families lead to faster recognition. We investigated the role of the morphological family size (MFS) effect in bilingual word recognition. First, re-analysis of available English lexical decision data from Dutch-English bilinguals reported by Schulpen (2003) revealed a facilitatory English MFS effect in purely English words and in Dutch-English interlingual homographs (e.g., ROOM, a word that exists in both English and Dutch, where it means “cream”). For interlingual homographs, the Dutch MFS was found to simultaneously induce inhibitory effects, supporting a language non selective access process. The MFS effect was independent of the relative frequency of the two readings of the homographs. Task-dependence of the MFS effect was demonstrated in generalized Dutch-English lexical decision data, which led to facilitatory effects of both families. Next, the pervasiveness of the MFS effect was demonstrated in a Dutch lexical decision task performed by the same type of bilinguals. Facilitatory effects of Dutch MFS were found for Dutch monolingual words and interlingual homographs, which were also affected by inhibitory effects of English MFS. The results are discussed in relation to the task-sensitive BIA+ model of bilingual word recognition.

Introduction

A large number of reaction time (RT) studies in the last decade have provided evidence that the recognition of visually presented words by bilinguals proceeds in a language non-selective way. Thus, Dutch-English bilinguals reading a book in their second language, English, will be affected by the lexical knowledge of their first language, Dutch, even when they are not aware of it (e.g., Van Heuven, Dijkstra, & Grainger, 1998; Van Hell & Dijkstra, 2002). To the extent that Dutch words are orthographically similar to the English words that the bilinguals are reading, they will be coactivated and may affect item selection. More surprisingly, the bilinguals will even be affected by the knowledge of similar English words when they are reading in Dutch. This indicates that the architecture of the lexical processing system is fundamentally non-selective in nature, although the actually observed effects of course depend on the degree of cross-language similarity of the involved items. A consequence of this theoretical position is that word recognition in L1 and L2 is open to effects of a variety of variables found to affect monolingual word recognition. In this paper, we will argue that this is indeed the case for a recently discovered independent variable in monolingual word recognition, morphological family size (MFS).

So far, the available evidence supporting language non-selective access has been collected in studies basically manipulating the degree of cross-language similarity of items in one of two ways (see Dijkstra & Van Heuven, 2002, for a review of studies). A first type of item manipulation has been to compare the RTs to words existing in one language only to words that share their orthographic, phonological, and/or semantic characteristics across languages. For instance, an item like LIST shares its orthography but not its meaning across Dutch and English. In contrast to such interlingual homographs, cognates like FILM share (most of) their meaning and orthography across languages. Studies have generally found facilitatory effects for cognates relative to one-language control items under various circumstances. The direction and size of RT effects for interlingual homographs appears to be dependent on task demands, stimulus list composition, and the relative frequency of the homograph readings in the two languages (e.g., Dijkstra, Grainger, & Van Heuven, 1999; Dijkstra, Van Jaarsveld, & Ten Brinke, 1998). Note that interlingual homographs are words in two languages.

In a second type of manipulation, the number of orthographically or phonologically similar items to the target word from the same and the other language has been manipulated. As an example, words that are orthographically similar to WORK (called neighbors) are CORK and WORD in English, and VORK and WERK in

Dutch. It has been found that intra-lingual and interlingual neighborhood density effects do indeed affect the RT patterns observed for target words (Van Heuven et al., 1998; Jared & Kroll, 2001). The manipulation of neighborhood density provides convincing evidence in support of language non-selective access, because it is concerned with "on-line" effects that concern "normal" words existing in only one language, in contrast to "special" words like homographs and cognates.

The effects of neighborhood density were first reported in monolingual studies before they were also demonstrated in bilingual studies. Recent monolingual studies have shown that the RTs in various word identification tasks are affected by yet another variable, called a word's "morphological family size" (MFS; see, e.g., Baayen, Lieber, & Schreuder, 1997; Bertram, Baayen, & Schreuder, 2000; de Jong, 2002; Schreuder & Baayen, 1997). For instance, a Dutch word like WERK (meaning "work") is a constituent of many morphologically complex words, among which are HUISWERK ("homework"), WERKBAAR ("workable"), and VERWERKEN ("to process"). Experiments have revealed that words with larger morphological families are processed faster and more accurately. This effect is independent of other lexical effects such as word frequency or length, and it is at least partially semantic in nature (De Jong, Schreuder, & Baayen, 2000). For instance, the effect of MFS appears for both regular and irregular past participles (e.g., GEROEID, "rowed" vs. GEVOCHTEN, "fought") although the irregular past participles do not share the exact orthographic or phonological form across family members (e.g., ROEIER, "rower", vs. VECHTER, "fighter"). Furthermore, only morphologically related words that are also semantically related contribute to the MFS (Bertram, Baayen, & Schreuder, 2000; Schreuder & Baayen, 1997). For instance, GEMEENTE ("municipality") is morphologically but not semantically related to GEMEEN ("nasty") and the correlation between RTs and family size decreases if GEMEENTE is included in the MFS count for GEMEEN.

Moscoso del Prado Martín, Deutsch, Frost, Schreuder, De Jong, and Baayen (2003) report an additional semantic characteristic of the MFS effect in Hebrew. The MFS of Hebrew words for which the morphological root is active in two semantic fields needs to be split into two different sub-families, one for each semantic field. Both subfamilies show effects of a similar magnitude on the RTs to a particular Hebrew word. However, the direction of the effect is reversed for the subfamily that contains the words that are in a semantic field different from that of the target. For instance, the Hebrew root R-G-L can form words whose meaning is related to "foot" (REGEL), and words whose meaning is related to "spy" (MERAGGEL). Response

latencies to REGEL are facilitated by the MFS containing those members of the family of R-G-L that are more related in meaning to “foot” and inhibited by the MFS containing the members of the R-G-L family that are more related in meaning to “spy”.

It is likely that bilinguals acquiring a second language will start to develop the morphological and semantic relations between words from their second language as well. Of course, the morphological family size in the second language (L2) may initially be smaller than in the first language (L1), but it should develop with vocabulary size. Therefore, just like an effect of interlingual neighbors can arise in bilingual word recognition (in spite of a smaller number of known L2 words), the MFS of a word would be expected to start playing a role in L2 as well. In other words, English word recognition in Dutch-English bilinguals should be affected by the English MFS of the target item. Even more interestingly, both the English and the Dutch MFS should play a role in the recognition of interlingual homographs, because these items belong to both English and Dutch families. Additionally, one might expect to find inhibition effects akin to those reported by Moscoso del Prado Martín et al. (2003) for Hebrew. For interlingual homographs, participants performing visual lexical decision in L2 might show a facilitatory effect of the MFS of the target in L2, and an inhibition of the MFS of the target in L1. Inversely, participants performing visual decision in their L1 will show a MFS effect of the MFS of the word in L1, and for interlingual homographs, they might also show an inhibition caused by the MFS of the word in L2.

These predictions follow straightforwardly from the basic assumptions of a recent model for bilingual word recognition, the BIA+ model (Dijkstra & Van Heuven, 2002). According to this language non-selective access model, word recognition entails parallel activation of words from different languages in an integrated word identification system. A task/decision system monitors lexical activity and uses it in accordance with the demands of the task at hand, allowing for context sensitive performance patterns. For instance, in an English lexical decision task, a higher word frequency in the target language may lead to faster RTs for interlingual homographs, while a higher word frequency in the non-target language induces inhibition (cf. Dijkstra et al., 1998). If the MFS behaves like word frequency, a large English MFS of the homograph should also exert a facilitatory effect in English lexical decision, because in this task it would support the selection of the target item and be indirectly linked to the correct response. At the same time, a large Dutch MFS, associated with the competitor item, should exert an inhibitory effect. This predic-

tion can be contrasted to that for generalized lexical decision, in which participants give a ”yes” response to both English and Dutch words. In this task situation, both English and Dutch MFS should have a facilitatory effect on the RTs.

In sum, in the present study we will investigate a number of issues with respect to morphological families in L1 and L2. First, we will search for the expected MFS effects in the L1 and L2 of bilinguals. Second, we will test if in L2 MFS effects of both L1 and L2 occur for interlingual homographs differing in their relative frequency in L1 and L2. More specifically, we will test the prediction of BIA+ that in an English lexical decision task the MFS effects of Dutch on English are inhibitory in nature, while in a generalized lexical decision task they are facilitatory. Finally, we will go even further and try to demonstrate effects of the English morphological family size of interlingual homographs in a Dutch (L1) lexical decision task. Finding MFS effects would provide additional independent evidence supporting the language non-selective access hypothesis, because it would demonstrate that it is not just the stronger L1 that is affecting the weaker L2, but that there is a mutual effect between the two languages.

These predictions will be tested in two different parts. In the first part, we will test the first two predictions by re-analyzing two recent Dutch-English lexical decision experiments reported by Schulpen (2003). In the second part, we will conduct a new Dutch lexical decision experiment involving largely the same test materials to test the third prediction, i.e., the presence of L2 on L1 effects in the interlingual homographs.

Reanalyses of Two Earlier Studies

Experiment 1: English Visual Lexical Decision

Method

Participants. Nineteen students of the University of Nijmegen (mean age: 22.5 years) were paid to participate in the experiment. All were native speakers of Dutch.

Materials. In total, the stimulus set consisted of 252 items of which 126 were words and 126 nonwords. All word items were selected from the CELEX database (Baayen, Piepenbrock & Van Rijn, 1993) and had a length of 3-6 letters. Table 4.1 describes the words used in the three experiments. The current experiment included the 42 interlingual homographs from the table, i.e., words that are legal

both in Dutch and English. Homographs were chosen from three frequency categories (high English frequency - High Dutch frequency, high English frequency-low Dutch frequency, and low English frequency-high Dutch frequency). The experiment also included the 84 monolingual English words from the table. These monolingual words were divided in four groups, three groups of 14 words each (English Controls) were matched in English frequency with the three groups of interlingual homographs, additionally 42 words (English Open Range) were chosen from English low, mid and high frequency ranges (14 of each). 84 nonwords were constructed with an orthographic structure that is legal in English. The nonwords were divided in three categories. First, nonwords were derived from a low frequent English word by changing one letter. Second, other nonwords were derived from a middle frequent English word by also changing one letter. Finally, 42 nonwords were constructed that had a structure specifically belonging to neither the Dutch nor the English language (as rated on a scale from 1 to 7 by a group of 10 Ph.D. students).

Table 4.1: Materials used in Experiments 1, 2, and 3. The word frequency counts are in occurrences per million.

Materials	Frequency Range	Number of Words	Dutch Freq.	English Freq.	Experiment
English-Dutch homographs	HF English - HF Dutch	14	104	233	1, 2, 3
	HF English - LF Dutch	14	9	244	
	LF English - HF Dutch	14	114	32	
English monolingual words	HF English (HF Dutch)	14	-	233	1, 2
	HF English (LF Dutch)	14	-	244	
	LF English (HF Dutch)	14	-	32	
	Open range	42	-	5, 48, 415	
Dutch monolingual words	(HF English) HF Dutch	14	104	-	2, 3
	(HF English) LF Dutch	14	9	-	
	(LF English) HF Dutch	14	114	-	
	Open range	42	5, 40, 489	-	

Procedure. The design consisted of item blocks that were rotated across participants. The presentation order of items within a block was randomized individually with the restriction that no more than three words or nonwords were presented in a row. Each participant was tested individually. The presentation of the visual stimuli and the recordings of the RTs were controlled by an Apple Powerbook G3 400 MHz with 128 megabytes of working memory, with an external Multiplescan 15AV Display and using experimentation software was developed by the technical group of the University in Nijmegen. The participants were seated at a table with the computer monitor at a 60 cm distance. The visual stimuli were presented in

capital letters (24 points) in font New Courier in the middle of the screen on a white background. The participants performed an English visual lexical decision (EVLD) task. They first read an English instruction, telling them that they would see a letter string to which they were supposed to react by pressing the 'yes' button when it was an English word or the 'no' button when the letter string was a nonword. The participants were told to react as quickly as possible without making too many errors. Each trial started with the visual presentation of a fixation dot for 500 ms followed after 150 ms by the target letter string in the middle of the screen. The target letter string remained on the screen until the participant responded or until a maximum of 2000 ms. When the button was pressed, the visual target stimulus disappeared and a new trial was triggered immediately.

The experiment was divided in three parts of equal length. The first part was preceded by 24 practice trials. After the practice set the participant could ask questions. All communication between participant and experimentator was conducted in Dutch. After the experiment, the participants were asked to fill out two questionnaires, one on paper about their level of proficiency in the English language, and one on the computer evaluating their knowledge of the stimulus words used in the experiment on a 7-point scale. In total, each experimental session lasted about 45 minutes.

Results

Data cleaning procedures were based on error rates for items and participants. All participants performed with an error rate of less than 20 percent. Thirteen items elicited errors in more than 30 percent of the trials, and were removed from the analyses. All incorrect responses were removed from the data (9.90% of all trials). After removing the errors, 47 trials with RTs that were outside the range of two and a half standard deviations from the mean RT were considered as outliers and were discarded (3.33% of the remaining trials). Table 4.2 provides the means and standard deviations of the frequency counts, family size counts, and RTs after the data cleaning procedures had been applied, and Table 4.3 provides the ranges for those same counts.

For this dataset, Schulpen et al. (2003) report results using ANOVAs on different frequency conditions from a factorial design contrasting high and low Dutch and English frequency. However, because we intend to assess the influence of an additional variable (morphological family size) that had not been controlled for in the original experimental design, and given that the word frequency and morphologi-

Table 4.2: Means and standard derivation for the different counts in the data set from Experiment 1 (EVLD), and in the subsets of interlingual homographs and English monolingual words (after removing outliers). The word frequency counts are in occurrences per million.

	Total 113 items	Interlingual homographs 37 items	English monolingual words 76 items
English frequency	182.55 ± 314.05	190.31 ± 297.78	178.86 ± 323.54
English family size	16.51 ± 21.89	12.27 ± 14.99	18.56 ± 24.38
Dutch frequency	26.63 ± 85.91	81.32 ± 135.62	-
Dutch family size	10.64 ± 25.29	32.49 ± 35.50	-
Response latency	566 ± 46 ms	561 ± 39 ms	569 ± 49 ms

Table 4.3: Ranges of the different counts in the data set from Experiment 1 (EVLD), and in the subsets of interlingual homographs and English monolingual words (after removing outliers). The word frequency counts are in occurrences per million.

	Total 113 items	Interlingual homographs 37 items	English monolingual words 76 items
English frequency	$[1, 1981]$	$[3, 1351]$	$[1, 1981]$
English family size	$[0, 112]$	$[1, 70]$	$[0, 112]$
Dutch frequency	$[0, 724]$	$[2, 724]$	-
Dutch family size	$[0, 134]$	$[0, 134]$	-
Response latency	$[479, 694]$ ms	$[500, 635]$ ms	$[479, 694]$ ms

cal family size variables for both languages in this dataset follow smooth lognormal distributions according to Shapiro-Wilk normality tests (Royston, 1982) of the log counts ($W = 0.99, p = 0.40$ for English frequency, $W = 0.98, p = 0.13$ for English family size, $W = 0.98, p = 0.58$ for Dutch frequency of the homographs, and $W = 0.95, p = 0.08$ for Dutch family size of the homographs), we will report regression analyses on the experimental results, which are more adequate for analyzing this sort of data. In all cases, we report sequential analyses of variance on stepwise multilevel linear regression models (Alegre & Gordon, 1999; Baayen, Tweedie, & Schreuder, 2002; Lorch & Myers, 1990; Pinheiro & Bates, 2000).

We begin by assessing whether English word frequency and morphological family size have a significant influence on the response latencies to the 76 English monolingual words in our dataset. A stepwise multilevel regression model with RT as the dependent variable and English word frequency and English morphological family size as independent variables showed facilitatory main effects of English word frequency ($F(1, 1313) = 112.28, p < 0.0001$), and English morphological family size ($F(1, 1313) = 7.80, p = 0.0053$, after having partialled out the effect of English word frequency), with the interaction between frequency and family size approaching but not reaching significance ($F(1, 1312) = 3.62, p = 0.0573$).

In order to analyze the results for the interlingual homographs in this experiment, we investigate the possible influences of the different variables on the response latencies by means of correlations. As all variables are lognormally distributed, we will report correlations on their logarithms. Both English counts show negative correlation coefficients with RTs ($r = -0.21, p = 0.21$ for English word frequency, and $r = -0.36, p = 0.03$ for English morphological family size), for which we note that the correlation is not significant for English word frequency. In contrast, both Dutch counts show significant positive correlations with the response latencies ($r = 0.38, p = 0.02$ for Dutch frequency, and $r = 0.36, p = 0.03$ for Dutch family size). This indicates that while English word frequency and English family size exert more or less facilitatory influences, Dutch frequency and Dutch family size have inhibitory effects on the response latencies. This is in line with Schulpen et al. (2003) who report opposite effects of English frequency and Dutch frequency for this dataset in an ANOVA on the factorial design.

Note that the magnitude of the correlation coefficient of both family size counts is similar, differing mainly in the direction of the effect. This is reminiscent of the pattern reported by Moscoso del Prado Martín et al. (2003) for Hebrew, in which the semantically close and semantically distant family sizes appear to have effects

that are equal in magnitude but different in direction. Moscoso del Prado Martín and colleagues operationalized this in their analyses by taking the difference between the two logarithmic counts as the predictor variable in their analyses. Note here that a difference in logarithmic scale is equivalent to the logarithm of the ratio (in non logarithmic scale). This log-transformed ratio, henceforth the family size ratio, turned out to be the crucial predictor for the Hebrew data.

In our regression analyses, we will make use of a similar approach, by which we consider the English counts to have a facilitatory effect on the task (given that English was the relevant language in the experiment), and the Dutch counts to have an inhibitory effect of the same magnitude as their English counterparts. More specifically, we will use two ratio variables for the interlingual homographs: the English-Dutch frequency ratio, i.e., the difference between the log of the English frequency and the log of the Dutch frequency, and the English-Dutch family size ratio, i.e., the difference between the logarithm of the English family size and the logarithm of the Dutch family size. The usage of these ratios allows us to jointly analyze the effects of these four highly correlated variables, and it significantly reduces the collinearity in the data matrix.

A stepwise multilevel regression analysis with the RTs as the dependent variable and the English-Dutch frequency ratio and the English-Dutch family size ratio as independent variables revealed facilitatory main effects for the frequency ratio ($F(1, 612) = 18.47, p < 0.0001$) and the family size ratio ($F(1, 612) = 5.95, p = 0.0150$ after having partialled out the effect of the frequency ratio), and no significant interaction ($F(1, 611) = 1.45, p = 0.2297$).

Taken together, the analyses reported here clearly show effects of word frequency, for both the English monolingual words and the English-Dutch interlingual homographs. In the later case, as illustrated by the effect of the frequency ratio, the effect of word frequency seems to be a composite effect, which consists of a facilitation caused by the English word frequency (words with a high frequency in English are recognized faster), and an equivalent inhibitory effect of the Dutch frequency count (homographs with a high Dutch frequency are recognized slower). These frequency effects confirm the results reported by Schulpen et al. (2003). Additionally, the regression analyses reveal an effect of morphological family size for both the English monolingual words and the interlingual homographs. This effect shows characteristics similar to that of frequency, in that the size of the morphological paradigm of a word in the relevant language (i.e., English) facilitates the recognition of the word. Thus, words from large English morphological paradigms are

recognized faster, while at the same time, the size of the morphological paradigm in the language that is not relevant for the task inhibits target recognition.

Experiment 2: Generalized Visual Lexical Decision

Method

Participants. Eighteen students of the University of Nijmegen (mean age: 22.5 years) were paid to participate in the experiment. All were native speakers of Dutch.

Materials. The stimulus set consisted of 420 items of which 210 were words and 210 nonwords. The current experiment included the same 42 interlingual homographs and the 84 monolingual English words from Experiment 1. Additionally, the 84 monolingual Dutch words mentioned in Table 4.1 were included in the experiment. As it was the case for the English words, the set of monolingual Dutch words consisted of three groups of 14 words, each matched in frequency to one of the groups of interlingual homographs (Dutch Controls), plus 42 words in the low, medium, and high Dutch frequency ranges (Dutch Open Range). For the nonwords, the experiment contained the 126 nonwords with an orthography that would be legal in English, plus an additional set of 84 possible Dutch nonwords obtained by changing a letter in low and middle frequency Dutch words.

Procedure. The procedure was identical to that of Experiment 1, except that this time participants performed a generalized Dutch–English visual lexical decision task (GVLD). They were instructed to react by pressing the 'yes' button when the stimulus on the screen was either a legal English word or a legal Dutch word, and the 'no' button when the letter string was not a word in Dutch or English. All communication between participant and experimentator was conducted in Dutch. In total, each experimental session lasted about 60 minutes.

Results

All participants performed with an error rate of less than 20 percent. Ten items elicited errors in more than 30 percent of the trials, and were removed from further analyses. All incorrect responses were removed from the data (7.57% of all trials). After removing the errors, 166 trials with RTs that were outside the range of two and half standard deviations from the mean RT were considered as outliers and

were thus discarded (4.75% of the remaining trials). Table 4.4 provides the means and standard deviations of the frequency counts, family size counts, and RTs after the data cleaning procedures had been applied, and Table 4.5 provides the ranges for those same counts.

Table 4.4: Means and standard derivation for the different counts in the data set from Experiment 2 (GVLD), and in the subsets of interlingual homographs and English monolingual words (after removing outliers). The word frequency counts are in occurrences per million.

	Total 200 items	Interlingual homographs 42 items	English monolingual words 74 items	Dutch monolingual words 84 items
English frequency	103.52 ± 252.33	169.78 ± 284.63	183.43 ± 326.71	-
English family size	9.42 ± 18.33	11.40 ± 14.28	18.99 ± 24.58	-
Dutch frequency	69.27 ± 177.22	75.90 ± 128.39	-	126.97 ± 243.90
Dutch family size	22.80 ± 40.23	31.71 ± 34.12	-	38.43 ± 51.87
Response latency	542 ± 48 ms	525 ± 29 ms	574 ± 49 ms	522 ± 37 ms

Table 4.5: Ranges of the different counts in the data set from Experiment 2 (GVLD), and in the subsets of interlingual homographs and English monolingual words (after removing outliers). The word frequency counts are in occurrences per million.

	Total 200 items	Interlingual homographs 42 items	English monolingual words 74 items	Dutch monolingual words 84 items
English frequency	$[0, 1981]$	$[3, 1351]$	$[1, 1981]$	-
English family size	$[0, 112]$	$[1, 70]$	$[0, 112]$	-
Dutch frequency	$[0, 1370]$	$[2, 724]$	-	$[1, 1370]$
Dutch family size	$[0, 283]$	$[0, 134]$	-	$[0, 283]$
Response latency	$[454, 716]$ ms	$[472, 598]$ ms	$[484, 716]$ ms	$[454, 606]$

As we did in the analyses of previous Experiment, given that the word frequency and morphological family size variables in this dataset follow smooth lognormal distributions according to conservative Shapiro-Wilk normality tests of the log counts ($W = 0.99, p = 0.27$, for frequency, $W = 0.98, p = 0.14$ for family size in the English words, and $W = 0.98, p = 0.12$ for Dutch family size of the Dutch words,[1]) once more we report analyses of variance on stepwise multilevel linear regression models.

We first assessed whether word frequency and morphological family size have a significant influence on the response latencies to the monolingual English and Dutch words in our dataset. A stepwise multilevel regression model with RT as

[1]Dutch frequency appeared to be significantly non-lognormally distributed according to the Shapiro-Wilks test ($W = 0.97, p = 0.02$). However, visual inspection of a quantile-quantile plot showed that this deviation was quite small.

the dependent variable and English word frequency and English morphological family size as independent variables showed main effects of English word frequency ($F(1, 1166) = 98.22, p < 0.0001$), and English morphological family size ($F(1, 1166) = 18.51, p < 0.0001$, after having partialled out the effect of English word frequency), and a significant interaction between frequency and family size ($F(1, 1166) = 4.11, p = 0.0429$).

With respect to the monolingual Dutch words, a stepwise multilevel regression model with RT as the dependent variable, and Dutch word frequency and Dutch morphological family size as independent variables showed main effects of Dutch word frequency ($F(1, 1412) = 74.40, p < 0.0001$), but no significant effect of Dutch morphological family size ($F(1, 1410) = 2.29, p = 0.1301$, after having partialled out the effect of Dutch word frequency), nor any interaction between frequency and family size ($F < 1$).

As for the English visual lexical decision experiment, we analyse the results for the interlingual homographs by using correlations on the log counts to provide an overview of the influences of the different variables on the response latencies. Both English counts show significant negative correlation coefficients with RTs ($r = -0.41, p < 0.0001$ for English word frequency, and $r = -0.30, p = 0.0009$ for English morphological family size). In contrast to what we observed in the English visual lexical decision experiment, in the present generalized visual lexical decision experiment, both Dutch counts show significant correlations with the response latencies ($r = -0.38, p < 0.0001$ for Dutch frequency, and $r = -0.40$, $p < 0.0001$ for Dutch family size), which are now also negative. This indicates that both English and Dutch word frequency and family size counts have facilitatory influences on the response latencies. This finding was predicted by the BIA+ model, according to which in the generalized visual lexical decision task, the English and the Dutch reading of the homograph would compete with each other in a “race”. According to this “race” hypothesis, the counts from both languages should have facilitation effects on the response latencies. More in detail, the word with the most support from the two languages should win the recognition race. We will operationalize this prediction by taking as independent variables in our regression analyses the maximum frequency of a word, i.e., the largest of the Dutch and the English frequency counts for a homograph, and the maximum family size of a word, i.e., the largest of the Dutch and the English family size counts for a given word, all of them in logarithmic scale. This operationalization allows us to keep the collinearity in the data matrix under control while at the same time testing the predictions of the BIA+

model.

A stepwise multilevel regression analysis with the RTs as the dependent variable and the maximum logarithmic frequency and maximum logarithmic family size as independent variables revealed facilitatory main effects for the maximum frequency ($F(1, 690) = 11.28, p = 0.0008$) and the maximum family size ($F(1, 690) = 5.23, p = 0.0225$, after having partialled out the effect of the maximum frequency) and no significant interaction ($F < 1$). The effects of the maximum frequency and family size are consistent with the predictions of the BIA+ model. However, other analyses also reveal a similar pattern. For instance, the summed frequencies and summed family sizes of both readings of a homograph are also excellent predictor of the RTs, just as the maximum frequency and family size are. This shows that the pattern in the data is robust and independent of the specific theoretical framework of the BIA+ model.

These analyses show effects of word frequency, for both the English and Dutch monolingual words, while in the case of the English-Dutch interlingual homographs it is the maximum of the frequency of the Dutch reading and the frequency of the English reading of the homograph that predicts response latencies. This effect of the maximum frequency confirms the predictions of models of bilingual word recognition that postulate the existence of a race between the two readings of a homograph in this task. Additionally, the analyses reveal the presence of family size effects for the English monolingual words and the interlingual homographs. Again, in line with the predictions of the BIA+ model, in the case of the interlingual homographs, it is the maximum between the two family size counts that exerts an influence on the response latencies.

Discussion of Experiments 1 and 2

Our reanalyses of the two earlier experiments confirm and extend the original results of Schulpen et al. (2003). As they already reported for these same datasets, in bilinguals, the frequency of the L1 reading of an interlingual homograph has an influence on response latencies when the participant is performing a visual lexical decision task in L2. This supports the predictions of the BIA+ model that both readings of the homograph are activated simultaneously. The presence of a “race” between both readings of the homograph entails that, in cases of words with a high L1 frequency, it is the L1 reading of the homograph that “wins” the race. This results in inhibition when the L1 reading of the homograph is not relevant for the task (English visual lexical decision), and in facilitation when either of the two readings

represents a valid word (generalized visual lexical decision).

The regression analyses document for the first time the presence of a morphological family size effect in the processing of L2. This is an indication that the family size effect is a fundamental characteristic of human lexical processing, that is to a large extent independent of the point in which a morphological paradigm or group of paradigms is acquired. Crucially, the morphological family size effect in L2 shows very similar characteristics to its counterpart in L1. The analyses have shown that the morphologically related words that are related in meaning to the target provide facilitation, while those whose meanings are not related to a relevant possible reading of the target for a given task produce inhibition. This inhibitory effect is very similar to the inhibitory effect caused by semantically opaque Hebrew words (Moscoso del Prado Martín et al., 2003).

A New Experiment

The finding of cross-lingual effects of MFS, reported in the previous section, provides new and independent evidence that during the processing of interlingual homographs, both their L1 and L2 readings are activated. As in earlier papers, this finding can be interpreted as evidence for language non-selective access. However, in the English lexical decision task, the response must be based upon the English reading of the homograph, so the evidence relates only to the effect of the strong L1 on the weaker L2. Furthermore, the bilingual participants are Dutch native speakers immersed in a Dutch environment. It is therefore possible that, although there are L1 effects of morphological family size in L2 processing, the opposite is not true. This would be the case, for instance, if L1, being the participants native language, has a special status to L2. In other words, language nonselective access would not be general, but restricted to the native language. To investigate this possibility, we conducted an additional experiment testing if the contrasting effects of frequency and family size of the L2 (English) reading of a homograph also arise when participants make visual lexical decisions in their L1 (Dutch). We tested this in the strongest way possible by ensuring that the participants were not aware of the relevance of their second language while they were performing in their native language: The data from all participants who noticed the bilingual nature of interlingual homographs were excluded from analysis.

Experiment 3: Dutch Visual Lexical Decision

Method

Participants. Twenty-nine students of the University of Nijmegen were paid to participate in the experiment. All were native speakers of Dutch.

Materials. The stimulus set consisted of 252 items of which 126 were Dutch words and 126 were nonwords. As specified in Table 4.1, this experiment included the same 42 interlingual homographs used in Experiments 1 and 2,and the 84 monolingual Dutch words from Experiment 2. As nonwords we included the 84 Dutch nonwords from Experiment 2, and the 42 nonwords with a pattern valid both in English and Dutch that were used in Experiments 1 and 2.

Procedure. The procedure was very similar to that employed in Experiments 1 and 2. Participants performed a Dutch visual lexical decision task (DVLD). They received their instructions in Dutch and were instructed to react by pressing the 'yes' button when the stimulus on the screen was legal Dutch word, and the 'no' button when the letter string was not a word in Dutch. Words were presented on a NEC Multisync color monitor in white lowercase 24 point Arial letters. After the experiment, the participants were asked if they had noticed anything special about the experiment. Five participants reported having noticed the presence of some English words in the experiment and their data were excluded from the analyses to ensure that the results were not affected by any conscious strategies of the participants. In total, the experimental session lasted about 45 minutes.

Results and Discussion

The remaining 24 participants performed with an error rate of less than 20 percent. Two items elicited errors in more than 30 percent of the trials, and were removed from the analyses. All incorrect responses were removed from the data (3.72% of all trials). After removing the errors, 24 trials with RTs that were outside the range of two and a half standard deviations from the mean RT were considered as outliers and were discarded (0.82% of the remaining trials). Table 4.6 provides the means and standard deviations of the frequency counts, family size counts, and RTs after the data cleaning procedures had been applied, and Table 4.7 provides the ranges for those counts.

Table 4.6: Means and standard derivation for the different counts in the data set from Experiment 3 (DVLD), and in the subsets of interlingual homographs and English monolingual words (after removing outliers). The word frequency counts are in occurrences per million.

	Total 124 items	Interlingual homographs 40 items	Dutch monolingual words 84 items
English frequency	55.63 ± 182.71	172.45 ± 290.90	-
English family size	3.57 ± 9.70	11.07 ± 14.53	-
Dutch frequency	111.35 ± 214.71	78.52 ± 131.00	126.98 ± 243.90
Dutch family size	36.77 ± 46.83	33.27 ± 34.22	38.43 ± 51.87
Response latency	530 ± 36 ms	537 ± 41 ms	527 ± 33 ms

Table 4.7: Ranges of the different counts in the data set from Experiment 3 (DVLD), and in the subsets of interlingual homographs and English monolingual words (after removing outliers). The word frequency counts are in occurrences per million.

	Total 124 items	Interlingual homographs 40 items	Dutch monolingual words 84 items
English frequency	$[0, 1351]$	$[3, 1351]$	-
English family size	$[0, 70]$	$[1, 70]$	-
Dutch frequency	$[0, 1370]$	$[2, 724]$	$[0, 1370]$
Dutch family size	$[0, 283]$	$[0, 134]$	$[0, 283]$
Response latency	$[460, 657]$ ms	$[460, 657]$ ms	$[469, 636]$ ms

In order to compare our results with those reported by Schulpen (2003), we begin by analyzing the results of the interlingual homographs and their frequency matched Dutch controls (excluding the Dutch open frequency range items) in terms of the original orthogonal design contrasting the English and Dutch frequencies of the homographs. Table 4.8 describes the reaction times and errors in each frequency condition after applying the data cleaning procedures. By-participant and by-item analyses of variance revealed significant effects of Frequency Category (High English-High Dutch, High English-Low Dutch, or Low English-High Dutch; $F_1(2, 46) = 25.24, p < 0.0001; F_2(2, 76) = 6.81, p = 0.0019$), a less clear effect of Type of Target (Interlingual homograph vs. Dutch control; $F_1(1, 23) = 4.77, p = 0.0396; F_2(1, 76) = 2.80, p = 0.0983$), and an interaction only reaching significance in the by-participant analysis ($F_1(2, 46) = 5.15, p = 0.0096; F_2(2, 76) = 1.41, p = 0.2506$). When analyzing in more detail the effect of Frequency category, we found that only the words in the High English frequency - Low Dutch frequency were significantly slower than the rest ($t = 2.17, p = 0.0330$).

A logistic regression on the ratio of incorrect to correct responses revealed significant effects of both Type of Target ($\chi^2_{1,80} = 28.76, p < 0.0001$), and Frequency Category ($\chi^2_{2,78} = 20.37, p < 0.0001$), with no significant interaction ($\chi^2_{2,76} = 0.83, p = 0.6610$). As in the analyses of RTs, the only frequency category that differed from the rest was the high English frequency-low Dutch frequency that gave rise to significantly more errors ($Z = 3.17, p = 0.0016$).

Table 4.8: Means and standard derivation of the reaction times, and error percentages for the different frequency categories of interlingual homographs and their frequency matched Dutch controls from Experiment 3 (DVLD), after applying data cleaning procedures.

	RT		SD		Errors	
	Homographs	Controls	Homographs	Controls	Homographs	Controls
HFE-HFD	535 ms	526 ms	43 ms	28 ms	3.66%	1.20%
LFE-HFD	514 ms	515 ms	29 ms	24 ms	2.61%	0.60%
HFE-LFD	562 ms	534 ms	41 ms	31 ms	10.34%	1.84%

In summary, these analyses show that interlingual homographs are significantly slower and produce more errors than their frequency matched controls, and words with a low Dutch frequency are slower than the rest. These analyses already reveal an interesting difference with the predictions of the BIA+ model, as this model does not predict the observed inhibitory effect from English on the interlingual ho-

mographs. Moreover, one would expect an interaction between the Type of Target and the Frequency Category, such that the homographs with a higher English frequency (or a lower Dutch to English frequency ratio) showed a greater inhibition with respect to their corresponding Dutch controls. However, this interaction was only significant in the by-participant analysis.

Having completed the factorial analysis of the frequency effects, we now turn to correlational and regression analysis of the data, including family size as independent variable. As in the previous experiments, the word frequency and morphological family size variables for both languages in this dataset follow lognormal distributions ($W = 0.96, p = 0.14$, for English frequency of the homographs, $W = 0.95, p = 0.08$ for English family size of the homographs, $W = 0.97, p = 0.02$ for Dutch frequency,[2] and $W = 0.99, p = 0.23$ for Dutch family size).

Both Dutch counts showed negative correlation coefficients with RTs ($r = -0.40$, $p < 0.0001$ for Dutch word frequency, and $r = -0.43, p < 0.0001$ for Dutch morphological family size). In contrast, both English counts showed positive correlations with the response latencies ($r = 0.29, p = 0.0737$ for English frequency,[3] and $r = 0.45, p = 0.0037$ for English family size). This pattern is precisely the opposite of that observed in the English visual lexical decision task. While Dutch word frequency and Dutch family size have facilitatory influences on the response latencies, English frequency and English family size exert inhibitory influences on the response latencies. Note that the correlation coefficients for English and Dutch are again similar.

The correlation analyses indicate that there is an effect of the English frequency of the homographs. Figure 4.1 provides further evidence for this English frequency effect, using non-parametric regression (robust locally weighted regression, Cleveland, 1979). The horizontal axis displays log frequency. The vertical axis plots the pairwise difference in the response latencies for the homographs and their controls. The solid line represents the effect of English frequency on this difference, the dashed line the effect of Dutch frequency. Note that as the English frequency increases, the difference in response latency between homograph and control word increases as well, with perhaps a ceiling effect for the highest-frequencies. Con-

[2] Although Dutch frequency appeared to be significantly non-lognormally distributed, visual inspection of a quantile-quantile plot showed that the deviation from lognormality was very small. Moreover, separate normality tests revealed that neither the subset of interlingual homographs ($W = 0.97, p > 0.05$), nor the subset of Dutch monolingual words ($W = 0.98, p = 0.60$), deviated significantly from lognormality.

[3] Although this correlation was only marginally significant, it reached full significance by a non-parametric Spearman correlation ($r_s = 0.37, p = 0.0189$).

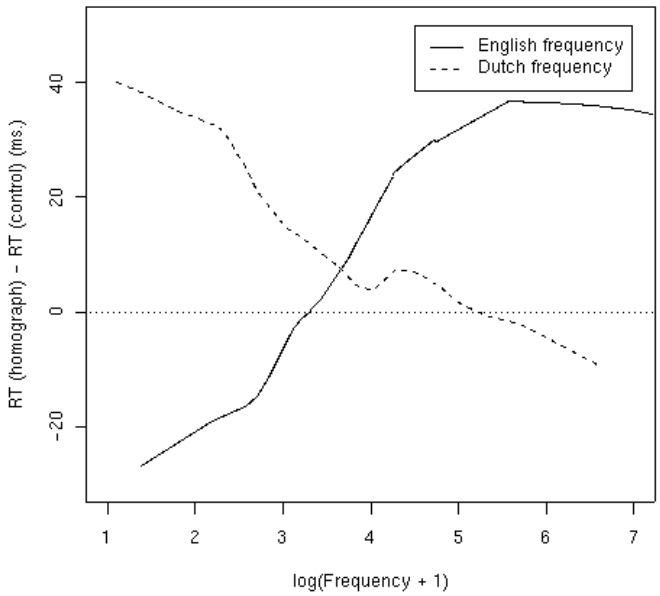

Figure 4.1: Nonparametric regression lines showing the effects of English word frequency (solid line) and Dutch word frequency (dashed line) on the average difference in RT between a homograph and its frequency-matched control. Frequency counts are in logarithmic scale.

versely, the difference in response latency decreases steadily with increasing Dutch frequency. The crossover of the two regression lines bears elegant witness to the inverse effects of English and Dutch frequency for interlingual homographs in Dutch visual lexical decision.

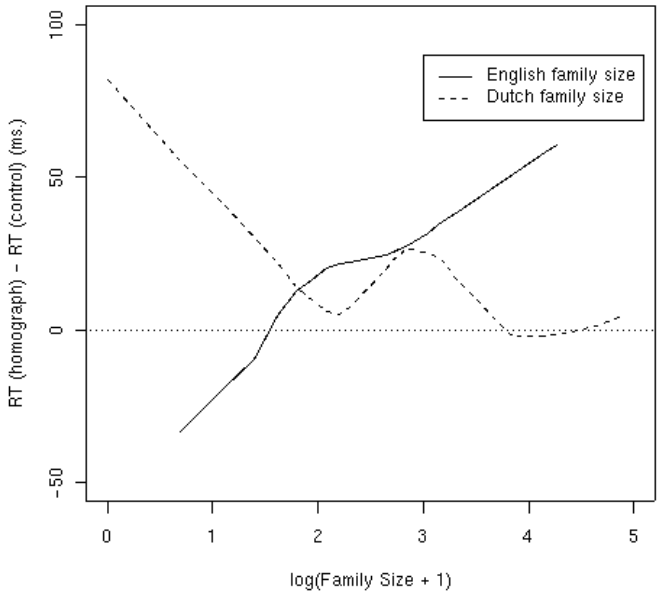

Figure 4.2: Nonparametric regression lines showing the effects of English morphological family size (solid line) and Dutch morphological famil size (dashed line) on the average difference in RT between a homograph and its frequency-matched control. Family size counts are in logarithmic scale.

Figure 4.2 shows a similar crossover between the effects of Dutch and English family size counts on the difference of RTs between homographs and controls. The horizontal axis now plots the logarithm of the family size counts, the vertical axis again displays the difference between the response latencies to a homograph and its control. The solid line representing the correlation between English family size and the difference in response latencies increases steadily with increasing family size. The greater the English family size is, the more time it takes to respond to a homograph relative to its control. The dashed line representing the correlation for Dutch suggests a floor effect for the words with larger families.

Having studied the correlations between frequency and family size with the difference in response latencies between the homographs and their controls, we now return to the response latencies to the homographs by themselves. As a first step, we fitted a stepwise multilevel regression model to the 80 Dutch monolin-

gual words with RT as the dependent variable, and log Dutch word frequency and log Dutch morphological family size as independent variables. We obtained facilitatory main effects of Dutch word frequency ($F(1, 1925) = 48.87, p < 0.0001$), and Dutch morphological family size ($F(1, 1925) = 8.42, p = 0.0038$, after having partialled out the effect of Dutch word frequency), with a small inhibitory interaction ($F(1, 1925) = 8.05, p = 0.0046$).

As in the preceding analyses, we extend the model to take account of English frequency and family size by means of frequency and family size ratios. The Dutch-English frequency ratio captures the difference between the log of the Dutch frequency and the log of the English frequency. Similarly, the Dutch-English family size ratio accounts for the difference between the logarithm of the Dutch family size and the logarithm of the English family size. A stepwise multilevel regression analysis with the RTs of the interlingual homographs as the dependent variable and the Dutch-English frequency ratio and the Dutch-English family size ratio as independent variables revealed facilitatory main effects for the frequency ratio ($F(1, 847) = 43.06, p < 0.0001$) and the family size ratio ($F(1, 847) = 20.20, p < 0.0001$, after having partialled out the effect of the frequency ratio), and no significant interaction ($F < 1$).

Figure 4.3 provides further validation for the use of the frequency and family size ratios. The horizontal axis plots the ratios, which on the logarithmic scale on which they are defined range from -4 to +4. The vertical axis plots the difference in RTs between the homograph and its control. The solid line visualizes the correlation between the latency difference and the frequency ratio, the dashed line represents this correlation for the family size ratio. Note that for both ratios, the relation with latency difference is roughly linear with a negative slope that is perhaps slighly larger for the family size ratio. A linear regression with the average different between RTs to the homographs and RTs to their control and the Dutch-English frequency ratio and the Dutch-English family size ratio as independent variables confirmed this linear relations. The frequency ratio had a significant effect on the magnitude of this difference ($F(1, 37) = 5.19, p = 0.0304$) and so did the family size ratio ($F(1, 37) = 4.16, p = 0.0488$, after partialling out the effect of the frequency ratio, with no significant interaction between them ($F < 1$).

The pattern of results obtained for the Dutch visual lexical decision data are consistent with those obtained for the English visual lexical decision data of Experiment 1. In both experiments, the frequency and family size ratios are key predictors of the response latencies for interlingual homographs. Crucially, however, the ratios

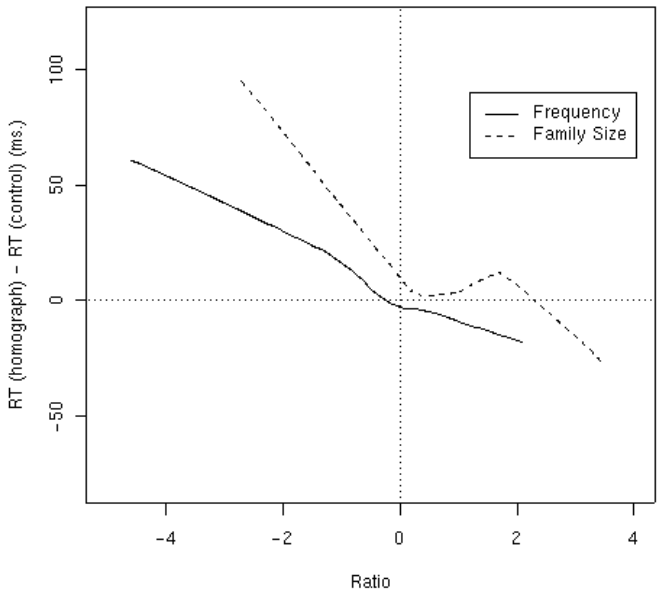

Figure 4.3: Nonparametric regression lines showing the combined effects of Dutch and English frequency (solid line), and Dutch and English morphological famil size (dashed line) on the average difference in RT between a homograph and its frequency-matched control. English and Dutch counts are combined using the Dutch-English frequency and family size ratios.

are defined with the English measures in the numerator and the Dutch measures in the denominator for Experiment 1, while for Experiment 3, the Dutch measures are in the numerator and the English measures in the denominator. In other words, the effects reverse when the task is changed from English lexical decision to Dutch lexical decision.

An additional factor that might influence the processing of the interlingual homographs amount of phonological overlap between the Dutch and English readings of a homograph. Some homographs have very similar phonological realizations, e.g., SLIM is pronounced as /slɪm/ in both English and Dutch. Other homographs have quite dissimilar phonolical realizations. A case in point is ANGEL, realized as /ɑŋəl/ in Dutch but as /eɪndʒəl/ in English. Although it is unlikely that the MFS effect would be confounded with phonological distance, we conclude our analyses by addressing this possibility.

To do so, a measure of phonological overlap between the English and Dutch readings of a homophone is required. We therefore calculated the distance between these forms using the Accumulation of Expectations (AoE; Moscoso del Prado Martín, Schreuder, & Baayen, 2003). This technique uses recurrent networks to obtain a multidimensional vector representation for word forms. Differences between word forms are formalized as the cosine distance in this multidimensional vector space. Since English and Dutch have overlapping but not identical phoneme inventories, we decided to restrict our analysis to those words that consist of phonemes common to both English and Dutch.

A sequential analysis of variance with response latency as dependent variable and with phonological distance, frequency ratio, and family size ratio as predictors, revealed an additional inhibitory effect of the phonological distance ($F(1, 671) = 23.17, p < 0.0001$) and a facilitatory interaction between phonological distance and family size ratio ($F(1, 671) = 12.77, p = 0.0001$). The effects of phonological distance and the ratio measures remained significant irrespective of the order in which the terms were entered in the regression.

Note, however, that longer words are likely to have larger phonological distances between their two pronunciations, simply because they have more phonemes that might differ. We therefore ran a final analysis in which we entered word length into the regression model before phonological distance. Although word length had a clear effect on the RTs ($F(1, 670) = 5.86, p < 0.0157$), its inclusion in the model did not eliminate the effect of phonological distance ($F(1, 670) = 17.28, p < 0.0001$), although it did interact with it ($F(1, 670) = 6.81, p = 0.0093$). In this analysis, both the

frequency and family size ratios remained robust predictors ($F(1, 670) = 42.34, p < 0.0001$ for the frequency ratio, and $F(1, 670) = 21.23, p < 0.0001$, for the family size ratio).

General Discussion

The present study shows that neither words nor languages as a whole should be considered as isolated building blocks in the organization of the mental lexicon. The recognition of words does not just depend on the characteristics of the items themselves (e.g., their frequency, length, or language membership), but also on lexical context, in our case the number of morphologically complex words that the word is related to. For words that exist in two languages, the morphological family sizes (MFS) in both languages play a role.

The experiments we presented led to several important new findings. First, in bilinguals, the MFS in L1 and L2 both affect lexical processing. Second, for interlingual homographs, MFS effects from both languages are present simultaneously. The direction of these effects is task dependent. In an English specific lexical decision experiment, English family size is facilitatory for homographs, while Dutch family size is inhibitory in nature. In contrast, in a generalized (Dutch-English) lexical decision task, both English and Dutch family size had a facilitatory effect.

The pervasiveness of the MFS was revealed by a Dutch lexical decision task. Participants performing in their native language, and unaware of the importance of their second language (English), nevertheless revealed inhibitory effects of the non-target language MFS were found on the RTs to the target language reading of interlingual homographs.

A striking finding is that the direction of the MFS effects is in line with that of word frequency effects. In accordance with earlier studies by Dijkstra et al. (1998) and Schulpen (2003), we reported that in the present English specific lexical decision study (Experiment 1), the RTs to interlingual homographs were inhibited to an extent that depended on the relative frequency of the two readings of these items. If the items had a low word frequency in English and a high word frequency in Dutch, RTs were slower than in a one-language control condition consisting of purely English words. In contrast, in the generalized Dutch-English lexical decision experiment (Experiment 2), word frequency in both languages exerted a facilitatory effect on the RTs.

Analogously to the English specific lexical decision experiment, in Experiment 3

(Dutch lexical decision), the RTs to the Dutch reading of interlingual homographs were faster when word frequency was higher in Dutch, and slower when it was higher in English (LFE-HFD condition). This finding is in agreement with the data from the Dutch lexical decision experiment (Experiment 2) by De Groot, Delmaar, and Lupker (2000).

Although the direction of MFS and word frequency effects was generally the same, the MFS effect in our data was not a mere frequency effect. The effects of the MFS remained significant after partialling out the word frequency effects. Still, the origin of both types of effects may to some extent comparable. Word frequency effects have been attributed to both orthographic and conceptually-semantic levels (e.g., Bradley & Foster, 1987, Morton, 1969, attribute frequency effects to the orthographic level, while Becker, 1979, Stanovich & West, 1981, Borowsky & Besner, 1993, Plaut & Booth, 2001 argue for it being a conceptual-semantic effect), and the same may hold for MFS effects. In the introduction, we already reviewed the empirical evidence supporting that view that the MFS effect has a strong semantic component (Bertram et al., 2000; De Jong et al., 2000; Moscoso del Prado Martín, Deutsch et al., 2003).

These findings and conclusions are compatible with both the monolingual model for the recognition of morphologically complex words that has been proposed to account for MFS effects by De Jong, Schreuder, & Baayen (in press) and with the BIA+ model of bilingual word recognition (Dijkstra & Van Heuven, 2002). Indeed, they suggest that an optimal integration of both models can easily be established.

The model accounting for morphological family size effects proposed by De Jong et al. (2003), the morphological family resonance model (FMRM) is an interactive activation model in which there is a cumulative build-up of activation resonating between lemmas (Levelt, 1989; Schreuder & Baayen, 1995) and the semantic and syntactic representations to which these lemmas are linked. If a lemma is linked to a semantic representation that itself is linked to a great many other lemmas, as is the case for a word with a large morphological family, this semantic representation will co-activate its associated lemmas, which in turn will contribute to the activation level of this semantic representation. Over time, this resonance within the morphological family speeds up the rate at which the activation of the target lemma increases. The greater the morphological family, the greater the rate will be with which the target lemma is activated. In the FMRM, a lexical decision response is initiated once a lemma has reached a threshold activation level. For a formal definition of the FMRM and some simulation studies, the reader is referred to De Jong et al.

(2003).

The MFS effects documented for the bilingual lexicon can be explained in the MFRM framework along the following lines. Given a homograph as orthographic input, e.g., ROOM, two lemmas are activated in parallel: the English lemma 'room' (meaning chamber), and the Dutch lemma 'room' (meaning cream). Both lemmas activate their families simultaneously through the resonance mechanism. Consequently, depending on their respective family sizes, the activation levels of the two lemmas increase exponentially over time. If the task is language-specific visual lexical decision, the appropriate lemma has to be selected. This might be accomplished by means of, for instance, the Luce choice rule. If the task is generalized visual lexical decision, the lexical decision can be made at the time the first representation reaches the threshold activation level. Extended in this way, the MFRM becomes very similar in spirit to the BIA+ model.

The BIA+ model is also an IA model assuming interactive links between orthography and meaning levels within the lexicon, but it has focussed on the bilingual recognition of monomorphemic words. The BIA+ model has succesfully modelled a range of word frequency effects in the bilingual lexicon, an area that the MFRM has not addressed at all. By assuming the more complex linkage system between lemmas and meaning that the MFRM proposes, the BIA+ model can be extended to account for the findings of the present study with respect to the MFS effects.

By incorporating the MFRM within the BIA+ architecture, a richer modelling framework is obtained that has interesting consequences for the processing of those interlingual homographs that share their meaning across languages, namely the homographs known as "cognates". An example of a form-identical cognate is the word FILM that shares its orthography, and to a large extent its semantics and phonology across languages. Non-identical cognates such as TOMAAT (Dutch) - TOMATO (English) also exist.

In the past, researchers such as Kirsner (e.g., Lalor & Kirsner, 2000) and Sánchez-Casas (e.g., Sánchez-Casas, Davis, & García-Albea, 1992) have proposed that cognates can be considered as morphological representations that are shared between languages. The extended BIA+ model provides a clear account of how this could work. The members of word pairs such as FILM/FILM and TOMATO/TOMAAT share most of their links to conceptual, semantic, and (partially) orthographic representations across languages. Given the process of morphological resonance we discussed above, the overlap may lead to the "semantic" facilitation effects that are so often observed for these types of items (e.g., Van Hell & Dijkstra, 2002). As in

the monolingual domain, the strength of the effect will depend on the transparency of the mappings between orthography and meaning within and across languages. It follows that cross-linguistic morphological priming effects should be obtained for items such as REGENACHTIG (Dutch for ”rainy”) and RAIN (REGEN in Dutch).

To conclude, starting from a strong language non-selective access assumption and the recent findings of morphological family size effects in the monolingual domain, we predicted similar within-language and between-language effects with respect to bilingual word recognition. We did not only find the expected effects in a reanalysis of two experiments that had been performed with a completely different aim, but also in a new study that specifically investigated the effect of the L2 family size on L1 homograph recognition under circumstances where the participants were processing in their strongest language and were unaware of the bilingual nature of the experiment. It further turned out that these empirically innovative data could be interpreted in an interesting theoretical integration of a monolingual model for the recognition of morphologically complex words and a bilingual model for bilingual word recognition. The new modeling framework, furthermore, allows a reinterpretation of earlier proposals about the representation of cross-linguistically ambiguous words such as interlingual homographs and cognates.

References

Alegre, M. and Gordon, P.: 1999, Frequency effects and the representational status of regular inflections, *Journal of Memory and Language* **40**, 41–61.

Baayen, R. H., Dijkstra, T. and Schreuder, R.: 1997, Singulars and plurals in Dutch: Evidence for a parallel dual route model, *Journal of Memory and Language* **37**, 94–117.

Baayen, R. H., Lieber, R. and Schreuder, R.: 1997, The morphological complexity of simplex nouns, *Linguistics* **35**, 861–877.

Baayen, R. H., Tweedie, F. J. and Schreuder, R.: 2002, The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon, *Brain and Language* **81**, 55–65.

Becker, C. A.: 1979, Semantic context and word frequency effects in visual word recognition, *Journal of Experimental Psychology: Human Perception and Performance* **5**, 252–259.

Bertram, R., Schreuder, R. and Baayen, R. H.: 2000, The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26**, 419–511.

Borowsky, R. and Besner, D.: 1993, Visual word recognition: A multistage activation model, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**, 813–840.

Bradley, D. C. and Forster, K. I.: 1987, A reader's view of listening, *Cognition* **25**, 103–134.

De Groot, A. M. B., Delmaar, P. and Lupker, S. J.: 2000, The processing of interlexical homographs in a bilingual and a monolingual task: Support for nonselective access to bilingual memory, *Quarterly Journal of Experimental Psychology* **53**, 397–428.

De Jong, N. H.: 2002, *Morphological families in the mental lexicon*, PhD thesis, University of Nijmegen, Nijmegen, The Netherlands.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: 2000, The morphological family size effect and morphology, *Language and Cognitive Processes* **15**, 329–365.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: *to appear*, Morphological resonance in the mental lexicon, *Linguistics*.

Dijkstra, A., Grainger, J. and Van Heuven, W. J. B.: 1999, Recognition of cognates

and interlingual homographs: The neglected role of phonology, *Journal of Memory and Language* **41**, 496–518.

Dijkstra, A. and Van Heuven, W. J. B.: 2002, The architecture of the bilingual word recognition system: From identification to decision, *Bilingualism: Language and Cognition* **5**, 175–197.

Dijkstra, A., Van Jaarsveld, H. and Ten Brinke, S.: 1998, Interlingual homograph recognition: Effects of task demands and language intermixing, *Bilingualism: Language and Cognition* **1**, 51–66.

Jared, D. and Kroll, J.: 2001, Do bilinguals activate phonological representations in one or both of their languages when naming words?, *Journal of Memory and Language* **44**, 2–31.

Lalor, E. and Kirsner, K.: 2000, Cross-lingual transfer effects between english and italian cognates and noncognates, *International Journal of Bilingualism* **4**, 385–398.

Levelt, W. J. M.: 1989, *Speaking. From intention to articulation*, The MIT Press, Cambridge, Mass.

Lorch, R. F. and Myers, J. L.: 1990, Regression analyses of repeated measures data in cognitive research, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**, 149–157.

Morton, J.: 1969, The interaction of information in word recognition, *Psychological Review* **76**, 165–178.

Moscoso del Prado Martín, F., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H. and Baayen, R. H.: 2003, Changing places: A cross-language perspective on frequency and family size in Hebrew and Dutch, *Manuscript submitted for publication, Max Planck Institute for Psycholinguistics*.

Pinheiro, J. C. and Bates, D. M.: 2000, *Mixed-effects models in S and S-PLUS*, Statistics and Computing, Springer, New York.

Plaut, D. C. and Booth, J. R.: 2000, Individual and developmental differences in semantic priming: Empirical and computational support for a single mechanism account of lexical processing, *Psychological Review* **107**, 786–823.

Royston, P.: 1982, An extension of Shapiro and Wilk's W Test for Normality to large samples, *Applied Statistics* **31**, 115–124.

Sánchez-Casas, R., Davis, C. W. and García-Albea, J. E.: 1992, Bilingual lexical processing: Exploring the cognate/non-cognate distinction, *European Journal*

of Cognitive Psychology **4**, 311–322.

Schreuder, R. and Baayen, R. H.: 1995, Modeling morphological processing, *in* L. B. Feldman (ed.), *Morphological Aspects of Language Processing*, Lawrence Erlbaum, Hillsdale, New Jersey, pp. 131–154.

Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **37**, 118–139.

Schulpen, B. J. H.: 2003, *Explorations in bilingual word recognition: Cross-modal, cross-sectional, and cross-language effects*, PhD thesis, University of Nijmegen, Nijmegen, The Netherlands.

Stanovich, K. E. and West, R. F.: 1981, The effect of sentence context on ongoing word recognition: Test of a two-process theory, *Journal of Experimental Psychology: Human Perception and Performance* **7**, 658–672.

Van Hell, J. G. and Dijkstra, A.: 2002, Foreign language knowledge can influence native language performance, *Psychonomic Bulletin & Review* **9**(4), 780–789.

Van Heuven, W. J. B., Dijkstra, A. and Grainger, J.: 1998, Orthographic neighborhood effects in bilingual word recognition, *Journal of Memory and Language* **39**, 458–483.

Information-Theoretical Characterization of Morphological Paradigms

CHAPTER 5

This chapter has been submitted as Fermín Moscoso del Prado Martín, Aleksandar Kostić, and R. Harald Baayen: Putting the bits together: An information-theoretical perspective on morphological processing.

Abstract

In this study we present an information theoretical approach to understanding the emergence of type-based and token-based effects in morphological processing. We propose a probabilistic measure of the informational complexity of a word, its information residual, which considers the combined influences of the amount of information contained by the target word and the amount of information carried by its morphological paradigms. By means of re-analyses of previously published data on Dutch words we show that the information residual outperforms the combination of traditional token-based and type-based counts in predicting response latencies in visual lexical decision.

Introduction

In the present study, we develop a new measure of the morphological complexity of a word: its information residual. This measure is inspired by the approach to processing of inflected morphology proposed by Kostić (Kostić, 1991, 1995, 2003; Kostić, Marković, & Baucal, in press). Our measure accounts for the combined effects of surface frequency, base frequency, family size and cumulative root frequency in a single framework. More importantly, it provides us with a parsimonious treatment of monomorphemic, polymorphemic, and compound words.

Several token-based counts (i.e., counting the number of occurrences of a word, a base form, or a root in a sufficiently large corpus), are reported to affect response latencies in the visual lexical decision task. Surface frequency, the number of times that an inflected form appears in a corpus, is negatively correlated with response latencies to monomorphemic words in visual lexical decision (Taft, 1979; Whaley, 1978). Also base frequency, that is, the summed frequency of all inflected variants of a word, has been shown to correlate positively with response latencies, even after partialling out the effect of surface frequency (Baayen, Dijkstra, & Schreuder, 1997; Taft, 1979). Similarly, Hay (2001) showed that the logarithm of the ratio between the surface frequency and the base frequency of a word, its inflectional ratio, correlates negatively with response latencies. Finally, the summed base frequency of all words derived from the same stem, its cumulative root frequency, has also been reported to have a negative correlation with response latencies after partialling out the effects of surface and base frequency (Colé, Beauvillain, & Seguí, 1989; Taft, 1979).

In contrast to these token-based counts, Schreuder and Baayen (1997) introduced a type-based measure that has an independent effect on responses to visual lexical decision in Dutch. The morphological family size of a word is the number of other polymorphemic words and compounds in which it appears as a constituent, independently of their frequencies of occurrence. For instance, the morphological family of the word *fear* contains the words *fearful*, *fearfully*, *fearfulness*, *fearless*, *fearlessly*, *fearlessness*, *fearsome*, and *godfearing*, according to the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), so the family size of *fear* equals 9. A facilitatory effect of family size has also been documented in a range of languages other than Dutch (Baayen, Lieber, & Schreuder, 1997; Ford, Marslen-Wilson, & Davis, in press; Lüdeling & De Jong, 2001; Moscoso del Prado Martín, Deutsch, Frost, Schreuder, De Jong, & Baayen, 2003, Moscoso del Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, 2003), and appears to arise at the level of

semantic processing (cf., De Jong, 2002).

Morphological family size is highly correlated with cumulative root frequency. In general, the more family members a word has, the higher their summed frequency will be. However, Schreuder and Baayen (1997) report that, when the effect of family size is controlled for there is no effect of cumulative root frequency. However, a re-analysis of their experimental results in (Baayen, Tweedie, & Schreuder, 2002) using a more sensitive linear mixed effect model (Pinheiro & Bates, 2000) revealed a small inhibitory effect of cumulative root frequency after having partialled out the facilitatory effect of family size.[1]

Kostić (Kostić, 1991, 1995, 2003; Kostić et al., in press) addressed the relevance of token- and type-based counts for inflectional processing using an information theoretical framework. In a series of lexical decision experiments with Serbian inflected nouns, he demonstrated that the amount of information carried by an inflected noun form is inversely proportional to its processing latency. The amount of information, however, was not derived from the token counts alone but from the ratio of the surface frequency to the type count of syntactic functions and meanings carried by an inflected form within a given paradigm (e.g., feminine nouns). This predictor accounted for almost all processing variability of inflected noun forms of all three Serbian grammatical genders.

The previous results seem to indicate that inflectional and derivational paradigms affect the recognition of a word in different ways. While for inflectional paradigms one should consider the summed frequency of all the members of the paradigm (base frequency), or the information content of the paradigms themselves, both of which are mainly token-based counts, it appears that the influence of derivational paradigms is best quantified by morphological family size, a completely type-based count, with a small additional effect of token-based cumulative root frequency. This picture is further complicated by compound words. De Jong, Feldman, Schreuder, Pastizzo, and Baayen (2002) reported that the morphological paradigm of a compound is only formed by those compounds that share the right constituent as a right constituent or the left constituent as a left constituent. This claim is based on analyses where they observed that it was the positional family size and positional cumulative root frequency, that is, restricted to the paradigm members in which the right constituent of the compounds appears as a right constituent, or left constituent

[1]Note here that, in their original studies, Baayen and colleagues referred to family frequency, which is the cumulative root frequency minus the base frequency of a word. However, since the frequency of the word is partialled out in all analyses, we will consider cumulative root frequency as equivalent to family frequency. In the analyses, we report the best result from using either of the counts, referring to both of them as cumulative root frequency.

appears as a left constituent, that best accounted for response latencies.

The question arises whether the two classes of counts (i.e., type-based and token-based) tap into different properties of the cognitive system, or reflect aspects of a single process. In this study, we develop a measure that reconciles the apparently contradictory findings as to the respective predictive powers of each of the counts across various types of words, including monomorphemic words, polymorphemic words, and compound words. In this way, our measure offers a parsimonious treatment of inflectional, derivational, and compounding relations, using a single measure, therefore eliminating the need for different cognitive processes dealing with each of these types of relations.

In what follows, we begin by formulating the information residual measure and discussing its relationship to different counts that have been described in the literature. We continue by evaluating the performance of this measure in predicting the response latencies of three previously published Dutch visual lexical decision experiments, as compared to using the traditional type and token frequency counts. Finally, we conclude with an outline of the implications of this measure for models of lexical processing.

The Information Residual of a Word

In this section we provide the general formulation of the information residual of a word. Our formulation will be guided by the different morphological effects that have been shown to affect response latencies in visual lexical decision. Table 5.1 provides a summary of the different variables from which we derive the information residual.

Amount of Information Contained by a Surface Form

The surface frequency of a word can be expressed by Shannon's amount of information, that is, the minimum number of bits that would be necessary to encode the word in an optimal binary coding of all the words in the lexicon (Shannon, 1948; for an extensive introduction to information theory see, e.g., Cover & Joy, 1991). In this way the *amount of information*($I_s(w)$) of a surface form w, with a frequency $F(w)$ in a corpus of size N is:

$$I_s(w) = -\log_2 p(w) \simeq -\log_2 \frac{F(w)}{N}, \tag{5.1}$$

Table 5.1: Summary of the morphological variables from which we derive the information residual.

Variable	Description	Type- or token-based	Correlations with RT reported in the literature
Surface Frequency	Number of times that a word appears in a corpus.	token-based	negative
Base Frequency	Sum of the surface frequencies of all inflectional variants of a word.	token-based	negative
Inflectional Ratio	Quotient between surface and base frequency of a word.	token-based	negative
Morphological Family Size	Number of different words that contain the same stem (excluding inflectional variants)	type-based	negative
Cumulative Root Frequency	Summed based frequencies of all words sharing a stem	token-based	negative/positive
Positional Family Size	Number of compounds containing the same left or right constituent.	type-based	negative
Positional Cumulative Root Frequency	Summed frequency of compounds containing the same left or right constituent.	token-based	negative

where $p(w) \simeq F(w)/N$ is the probability of encountering w in a corpus. Note that, according to (5.1), the amount of information is inversely proportional to the log frequency of the word. As it is established that logarithmic frequency correlates negatively with lexical decision latencies (Rubenstein & Pollack,1963; Scarborough, Cortese, & Scarborough, 1977; Shapiro, 1969), the amount of information of a word will show a positive correlation with lexical decision latencies.

Morphological Paradigms

Kostić (2003) showed that the average amount of information of an inflectional paradigm (e.g. feminine nouns) is inversely related to processing speed of individual inflected forms that constitute the paradigm – high average amounts of information are paralleled by shorter processing speed per one bit in a given experiment. Kostić used the average amount of information in the individual forms of the inflectional paradigm, without weighting each of these amounts of information by its probability of occurrence, to estimate the amount of information contained by an inflectional paradigm. Instead, we propose using a more standard measure to estimate the amount of information in an inflectional paradigm, its entropy (Shannon, 1948). We can consider the inflectional paradigm of a word to be a random variable whose possible values are the different inflected forms that a base word can take. Hence, we can calculate the entropy of the inflectional paradigm, its *inflectional entropy*. In general, the entropy of a paradigm $\mathcal{P}$ with $V(\mathcal{P})$ members $\{x_1, \ldots, x_{V(\mathcal{P})}\}$, each of which has a probability of occurrence of $p(x_i|\mathcal{P}) \simeq F(x_i)/F(\mathcal{P})$, is:

$$H(\mathcal{P}) = -\sum_{x \in \mathcal{P}} p(x|\mathcal{P}) \log_2 p(x|\mathcal{P}) \simeq -\sum_{x \in \mathcal{P}} \frac{F(x)}{F(\mathcal{P})} \log_2 \frac{F(x)}{F(\mathcal{P})}, \tag{5.2}$$

where $F(\mathcal{P})$ is base frequency of the inflectional paradigm, and $F(x_i)$ is the surface frequency of the word. Note that this measure is related to the base frequencies and inflectional ratios (the inflectional entropy of a paradigm is the weighted sum of the inflectional ratios of its members). Therefore, we predict that the inflectional entropy will correlate negatively with response latencies, in line with the negative correlations with response latencies found for base frequency and inflectional ratios.

For example, the inflectional paradigm of the base form *car* consists of the forms “car” and “cars” with surface frequencies $F(\text{“car”})$ and $F(\text{“cars”})$.[2] The base fre-

[2]Throughout this paper we will use double quotes to refer to surface forms.

quency of the inflectional paradigm *car* is $F(\text{car}) = F(\text{“car”}) + F(\text{“cars”})$ and the probabilities of the inflected form being “car” or “cars” within the inflectional paradigm of *car* are $p(\text{“car”}) = F(\text{“car”})/F(\text{car})$ and $p(\text{“cars”}) = F(\text{“cars”})/F(\text{car})$. The entropy of the inflectional paradigm of *car* will then be $H(\text{car}) = -p(\text{“car”}) \log_2 p(\text{“car”}) - p(\text{“cars”}) \log_2 p(\text{“cars”})$.

We can also conceive of derivational paradigms as random variables, for which we can calculate the entropy. Once again, we can estimate the probability of occurrence of each of word inside the paradigm by its derivational ratio: the word's base frequency divided by the summed frequency of all members of the paradigm (the cumulative root frequency), and then calculate the entropy according to (5.2).

As one can deduce from (5.2), in general, the greater the number of members in a morphological paradigm, the greater the entropy of the paradigm will tend to be. Everything else being equal, increasing the number of members will decrease their probability of occurrence within the paradigm, thus increasing the number of bits required to represent each of them. In fact, family size gives us an estimation of the maximum entropy situation (i.e., the case when all paradigm members have the same probability of occurrence). Figure 5.1 compares the number of bits of the entropy of the paradigm for all Dutch monomorphemic words in the CELEX Lexical Database (Baayen et al., 1995) in the family size range $[1, 200]$, with the maximum entropy for words with that morphological family size. The fact that both curves follow very similar patterns ensures that both measures will correlate similarly with reaction times.

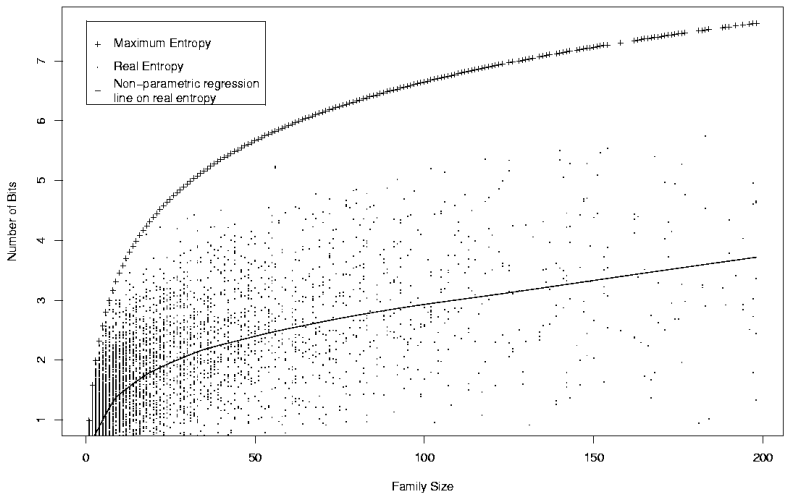

Figure 5.1: Comparison between real and maximum entropy for all monomorphemic Dutch words in the family size range $[1, 200]$.

To illustrate the way in which the derivational entropy measure also accounts for the inhibitory effects of cumulative root frequency after having partialled out the effect of family size, consider Figure 5.2. It shows the frequency histograms of the derivational paradigms of the two Dutch words *barbaar* (*barbarian*) and *faam* (*fame*) taken from the dataset where Baayen and colleagues found the inhibitory effect of cumulative root frequency. Both of them have five members in their derivational paradigms, and both have similar base frequencies (228 for *barbaar* and 218 for *faam*),[3] however, the cumulative root frequency is greater in the paradigm of *faam* (676) than in the paradigm of *barbaar* (566). The increase in the frequencies of the members of the paradigm of *faam* with respect to the frequencies of those of the paradigm of *barbaar* is not constant, but consists of a very strong increase in the frequency of the most frequent paradigm member, together with smaller increases (in this case even decreases) in the frequencies of the other members, in line with the Zipfian distributions of morphological paradigms. This results in a more skewed distribution of frequencies in the paradigm with the greatest cumulative root frequency, and thus in a lower entropy value for that paradigm. Correspondingly, the entropy of the paradigm of *faam* ($H(\mathrm{faam}) = 1.13$) is slightly lower than that of the paradigm of *barbaar* ($H(\mathrm{barbaar}) = 1.36$). As the entropy of the paradigm is negatively correlated with response latencies, this will result in words from the paradigm of *faam* receiving less facilitation from their derivational paradigm than words from the paradigm of *barbaar* receive from theirs, resulting in their being recognized slower.

Hierarchical Structure of the Paradigms

Inflectional paradigms are nested within derivational paradigms, which in turn are nested within each other forming tree-like structures. This hierarchical structure entails that a single word can belong simultaneously to several morphological paradigms. For instance, Figure 5.3 shows the morphological paradigms which include the polymorphemic word “thinkers”. “Thinkers” and its singular form “thinker” belong to the inflectional paradigm of *thinker*, which in turn belongs to the derivational paradigm of $[\mathrm{thinker}]$. $[\mathrm{thinker}]$ itself is a paradigm of the larger paradigm of $[\mathrm{think}]$, which includes other elements such as $[\mathrm{rethink}]$ or the inflectional paradigm of *think*.[4]

We need to specify the joint entropy of the paradigms to which a given word

[3] The frequency counts are based on a corpus of 42 million words.

[4] We will use square brackets to refer to derivational paradigms.

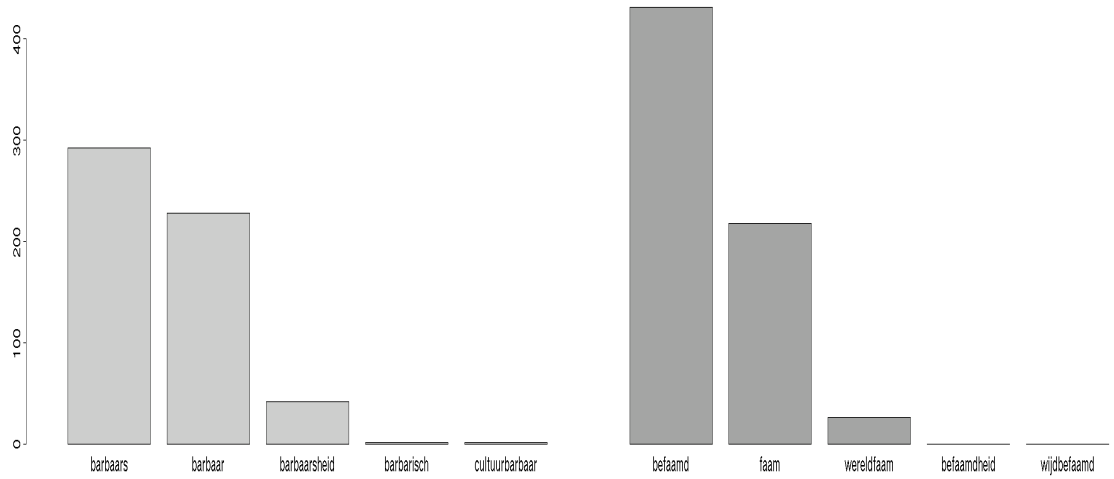

Figure 5.2: Comparison of the frequencies of the members of the derivational paradigms of the Dutch words *barbaar* and *faam*.

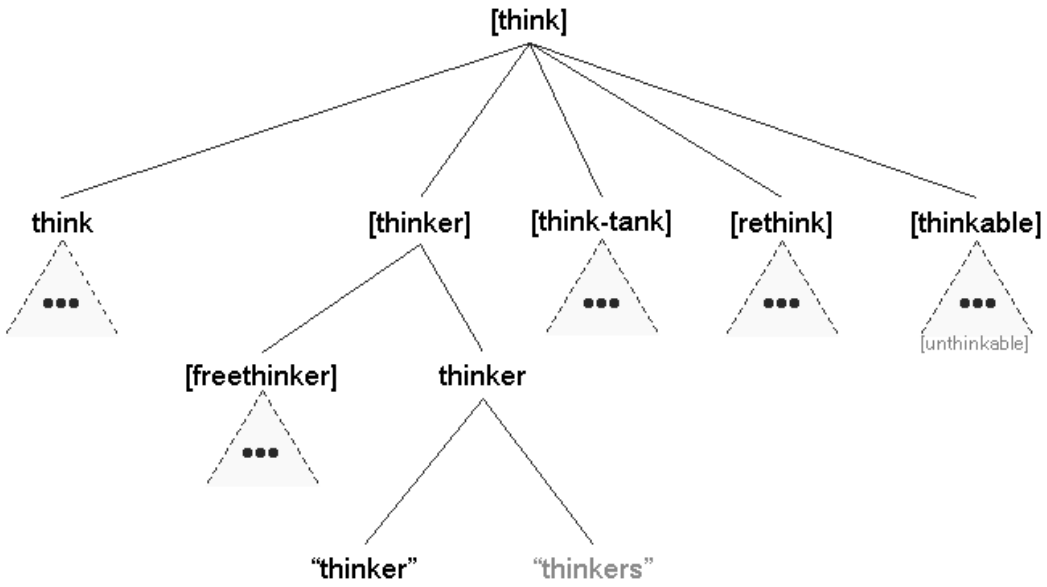

Figure 5.3: Morphological paradigms for the polymorphemic word *"thinkers"*.

belongs. In general, except for compound words, the joint entropy of the morphological paradigms can be calculated as the sum of the entropies of all the morphological paradigms that dominate it in the tree. This is because the paradigms are nested within each other. In this case the joint probability of the nested paradigms is the product of the probabilities of each of them and, in logarithmic scale, this product turns into a sum. We will refer to the *paradigmatic entropy* ($H_{tot}(w)$) of a word as the joint entropy of all the morphological paradigms that dominate it in the morphological tree.

For instance, in the above example, the joint entropy of the paradigms to which "thinkers" belongs can be calculated as the sum of the individual entropies of the paradigms:

$$\begin{aligned} H(thinker, [\text{thinker}], [\text{think}]) &= H([\text{think}]) + H([\text{thinker}]|[\text{think}]) \\ &\quad + H(\text{thinker}|[\text{thinker}], [\text{think}]) \\ &= H(\text{thinker}) + H([\text{thinker}]) + H([\text{think}]). \end{aligned} \quad (5.3)$$

Compound Words

Figure 5.4a shows the typical paradigmatic structure of a compound. In this case, to calculate the joint paradigmatic entropies of the levels that dominate $[\text{think-tank}]$ is problematic because it has two direct ancestors. We cannot calculate the joint entropy of the two paradigms by addition – considering the two paradigms of $[\text{think}]$ and $[\text{tank}]$ separately and adding them up. The members of the paradigms of $[\text{think}]$ and $[\text{tank}]$ are mutually exclusive except for $[\text{think-tank}]$ itself. This means that we can consider the union of the $[\text{think}]$ and $[\text{tank}]$ paradigms as a single random variable, as shown in Figure 5.4b, and then simply calculate the entropy according to (5.2).

With respect to the positional restriction on compound families suggested by De Jong et al. (2002), an analysis of the distribution of the cumulative root frequencies of the left and right constituents of the compounds in their study revealed that, in their experimental dataset, left constituents tend to appear as left constituents in other compounds more often than they appear as right consituents of other compounds (paired $t = -7.8203, df = 109, p < 0.0001$), and vice-versa , right constituents tend to appear as right constituents in other compounds (paired $t = 5.0184, df = 111, t < 0.0001$). Hence, the positional family effects that they reported may arise from the inner distribution of the families.

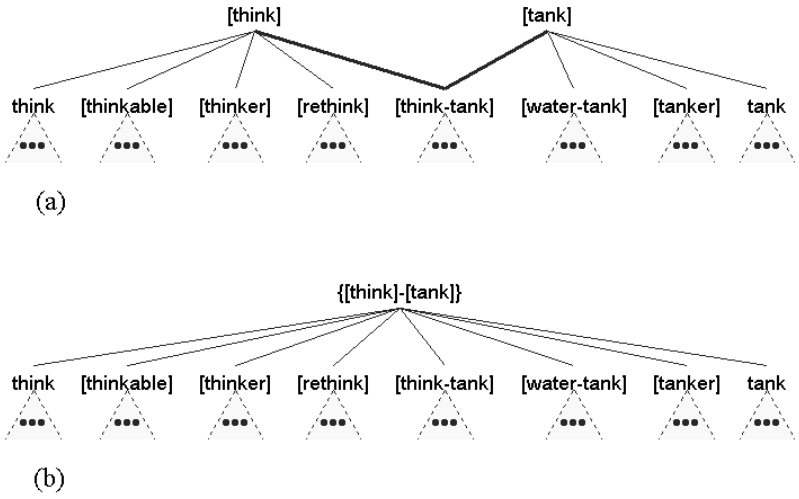

Figure 5.4: (a) Derivational paradigms of $[\text{think}]$ and $[\text{tank}]$ (b) Derivational paradigm of $\{[\text{think}], [\text{tank}]\}$.

Putting the Bits Together

At the beginning of this section, we showed how surface frequency can be described as the amount of information contained by a word, $I_s(w)$. On the one hand, $I_s(w)$ gives us an estimation of how difficult it is to recognize a word by itself. The greater the amount of information contained by the word, the more costly it will be to recognize. On the other hand, the total paradigmatic entropy $H_{tot}(w)$ gives us an estimation of how much support that word receives from the morphological paradigms to which it belongs. As both I_s and H_{tot} are measured in logarithmic scale, we can express the quotient between the cost of recognizing a word and the support provided by its paradigms as a difference, *the information residual of a word*:

$$I_R(w) = I_s(w) - H_{tot}(w). \tag{5.4}$$

Note that we predict the information residual will be positively correlated with processing latencies. This is a direct consequence of I_s being positively correlated with response latencies, and H_{tot} being negatively correlated with response latencies.

Re-analyses of Previously Published Experiments

In this section, we compare the performance of our information residual measure in predicting response latencies in visual lexical decision experiments on the one hand, to the performance of the traditional measures (surface and base frequency, family size, and cumulative root frequency) on the other.

Methods

Materials. We obtained the datasets from three previously published visual lexical decision experiments on monomorphemic words (Schreuder & Baayen, 1997; Experiment 3), polymorphemic words (Neijt, Schreuder & Baayen, 2003), and compounds (De Jong et al., 2002; Experiment 1b). The experiment on monomorphemic words contrasted morphological family size, controlling for frequency and cumulative root frequency. The experiment on polymorphemic words consisted of Dutch words with different feminine agentive suffixes (such as the *-ess* in English *countess*), and the authors were investigating the effects of the different origins of the suffixes (Latinate or Germanic) on visual word recognition. As this distinction goes beyond the scope of this paper, we will limit ourselves to partial out that variance by adding an additional variable “suffix type” into our regression analyses. Finally, the compound words dataset was an experiment in which the authors used compounds with constant left constituents that were controlled for base frequency and word length in letters (of the right constituent and the compounds), and they contrasted the morphological family size of the right contituents of the compounds.

In order to obtain comparable results in all experiments, we recalculated all frequency, family size, and cumulative root frequency measures using the morphological parses and word frequency counts available in the CELEX lexical database (Baayen et al., 1995). In order to provide the maximum possible accuracy for items where the CELEX parsing was erroneous according to standard dictionary-based parses, we corrected these parses by hand. We calculated the information residual value for all the words in the three datasets using the method described in the previous section. Before performing any analyses we excluded from the datasets all trials and items with reaction times above or below two and a half standard deviations from the mean (in logarithmic scale).

Procedure. We fitted two multiple regression models (each including a by participant and a by-item regression) to each of the three datasets. One of the regressions

had the logarithms of the traditional counts (word frequency, morphological family size, and cumulative root frequency) as independent variables, while the other had only the information residual as an independent variable. Both regressions had the logarithm of the response latencies as the dependent variable. The regressions on the polymorphemic word dataset (Neijt et al., 2003) had suffix type as an additional independent variable, whose values were the different feminine agentive suffixes present in the experimental items.

Results and Discussion

In all cases we report sequential analyses of variance on by-participant multilevel regression analyses (Alegre & Gordon, 1999; Baayen et al., 2002; Pinheiro & Bates, 2000; Table 5.2) and traditional by-item linear regression models (Table 5.3).

The first three columns of Tables 5.2 and 5.3 present the results of the by-participant and by-item analyses on each of the datasets. The upper section of the tables concerns the analyses using the traditional counts, while the middle section contains the results of the analyses using the information residual measure. In all analyses, the effects are reported in the order in which they were entered into the regressions, so the significance of each of the effects is calculated after partialling out the contribution of the effects reported above it. The signs between brackets represent the direction of the effects (the sign of their coefficient in the models), for the effects that were significant in each of the analyses. The bottom rows of the tables compare the amount of variance explained by the regression using the information residual, with the amount of variance explained by the traditional type- and token-based counts. Additionally, in the bottom row of Table 5.3, we report analyses of variance testing whether the models are significantly different (it is not possible to do this on the by-participant multilevel regressions).

If we consider the results of the analyses using the traditional counts, we find a facilitatory effect of frequency in all three experiments. Next, in all datasets, we also find a facilitatory effect of morphological family size. Finally the effect of cumulative root frequency and positional family size only reaches significance in the by-participant regressions (except in the second dataset, where it does not reach significance at all). Interestingly, as we predicted in the previous section, once the facilitatory effect of morphological family size has been partialled out, cumulative root frequency, when present, shows an inhibitory effect.

In contrast to the different effects that are found in each experiment using the traditional counts, we find that the information residual count has a consistent in-

Table 5.2: Summary of by-participant multi-level regressions. Freq. refers to frequency, Cum. Freq. refers to Cumulative root frequency, Fam. Size to morphological family size, I_R to the information residual, I'_R to the modified information residual, and Suffix to the effects of suffix type. The signs in brackets in front of the effect represent the direction of the effects. Significance codes are: $^{+}: p < 0.1$, $^{*}: p < 0.05$, $^{**}: p < 0.005$, and $^{***}: p < 0.0005$

	Schreuder & Baayen, 1997 – Exp.3	Neijt et al., 2003	De Jong et al., 2003 – Exp.1b
Traditional analyses	(–) Freq.: $F(1, 956) = 562.57^{***}$ (–) Fam. Size: $F(1, 956) = 129.13^{***}$ (+) Cum. Freq.: $F(1, 956) = 68.29^{***}$	(–) Freq.: $F(1, 1036) = 187.12^{***}$ (–) Fam. Size.: $F(1, 1287) = 39.73^{***}$ Cum. Freq.: $F < 1$ Suffix: $F(10, 1036) = 2.66^{**}$	(–) Freq.: $F(1, 1287) = 97.39^{***}$ (–) Pos. Fam. Size.: $F(1, 1287) = 33.69^{***}$ (+) Pos. Cum. Freq.: $F(1, 1287) = 5.90^{*}$
Explained variance (r^2)	44%	47%	39%
I_R analyses	(+) I_R: $F(1, 958) = 923.37^{***}$	(+) I_R: $F(1, 1037) = 212.25^{***}$ Suffix: $F(10, 1037) = 2.90^{**}$	(+) I_R: $F(1, 1289) = 60.96^{***}$ (+) I'_R: $F(1, 1289) = 168.49^{***}$
Explained variance (r^2)	48%	47%	35% 40%
Comparison of models	+4%	0%	–4% +1%

Table 5.3: Summary of by-item regressions. Freq. refers to frequency, Cum. Freq. refers to Cumulative root frequency, Fam. Size to morphological family size, I_R to the information residual, I'_R to the modified information residual, and Suffix to the effects of suffix type. The signs in brackets in front of the effect represent the direction of the effects. Significance codes are: $^{+}: p < 0.1$, $^{*}: p < 0.05$, $^{**}: p < 0.005$, and $^{***}: p < 0.0005$

	Schreuder & Baayen, 1997 – Exp.3	Neijt et al., 2003	De Jong et al., 2003 – Exp.1b
Traditional analyses	(–) Freq.: $F(1, 33) = 18.10^{***}$ (–) Fam. Size: $F(1, 33) = 4.40^{*}$ Cum. Freq.: $F(1, 32) = 2.31$	(–) Freq.: $F(1, 37) = 52.87^{***}$ (–) Fam. Size: $F(1, 37) = 8.19^{**}$ Cum. Freq: $F < 1$ Suffix: $F < 1$	(–) Freq.: $F(1, 108) = 18.57^{***}$ (–) Pos. Fam. Size: $F(1, 108) = 7.54^{**}$ Pos. Cum. Freq.: $F < 1$
Explained variance (adjusted r^2)	37%	60%	18%
I_R analyses	(+) I_R: $F(1, 34) = 32.92^{***}$	(+) I_R: $F(1, 38) = 58.09^{***}$ Suffix: $F < 1$	(+) I_R: $F(1, 109) = 13.95^{***}$ (+) I'_R: $F(1, 109) = 39.26^{***}$
Explained variance (adjusted r^2)	48%	59%	11% 26%
Comparison of models	+11%, $F(33, 34) = 4.80^{*}$	–1%, $F(38, 39) = 1.78$	–7%, $F(108, 109) = 10.89^{**}$ +8%, $F(108, 109) = 9.40^{**}$

hibitory effect in all analyses. The consistency of the information residual effect represents an advantage over the changing effects of the traditional counts. If we compare the amount of variance explained by the information residual with the variance explained using the traditional counts, we find that the information residual analyses significantly outperforms the traditional counts in the first dataset, and that its predictive power is not significantly different for the second dataset.

Interestingly, in the case of the compounds, the traditional counts clearly outperform the information residual analyses in terms of explained variance. This underperformance for the compounds led us to investigate in more detail the contribution to the effect of each of the components that form the information residual, that is, the information content of the surface form, and the paradigmatic entropies at the different levels. We did this by fitting regression models with log reaction time as the dependent variable and amount of information of the word and the paradigmatic entropies at the different levels as independent variables. This decomposition showed that the amount of information of the word and the paradigmatic entropies up to and including the joint entropy between the two constituents were showing effects in the predicted direction (inhibitory for the amount of information and facilitatory for the paradigmatic entropies). However, the paradigmatic entropies of the levels in the tree located above the node where the joint entropy of the compound constituents is calculated were showing an inhibitory effect, opposite to the direction we had predicted. This inhibitory effect was significant both in the by-participant ($F(1, 1288) = 60.45, p < 0.0001$) and in the by-item regressions ($F(1, 108) = 12.26, p = 0.0007$). The inhibitory effect of these 'upper' paradigms of a compound suggests a refinement in our description of the total paradigmatic entropy of a word is required.

The members of the paradigms at the 'upper' levels show much weaker semantic relations to the target than the members of the more immediate paradigms that are dominated by the level at which the constituents of the compound are joined. The presence of such 'upper' paradigms produces paradigms that are highly heterogeneous with respect to the meanings of their members. For instance, in our dataset at these levels we find words that bear very distant semantic relations to the targets, such as *vangen* (*to catch*) in the paradigm of *gevangenispsychiater* (*prison psychiatrist*), and words that are even completely unrelated to the target, such as *boter* (*butter*) in the paradigm of *avondboterham* (*evening sandwich*). It has been shown that the effects of morphological paradigms arise as a consequence of the semantic relations that bind the members of a morphological paradigm (cf.

De Jong, 2002). Interestingly, Moscoso de Prado Martín, Deutsch, et al. (2003) reports an inhibitory effect for semantically distant family members in Hebrew morphological families whose members belong to heterogeneous semantic fields. The contribution of the morphological paradigms to the I_R measure should therefore be divided into two separate components. On the one hand there is the joint paradigmatic entropy of the paradigms whose members bear very close semantic relations to the target ($H_{related}(w)$). The information in these paradigms provides support for the recognition of the word. On the other hand, the joint entropy of more distant paradigms ($H_{unrelated}(w)$) does not support the recognition of the word, but rather makes it more difficult. Introducing this change in the information residual measure from (5.4) we obtain:

$$I_R(w) = I_s(w) - H_{related}(w) + H_{unrelated}(w). \tag{5.5}$$

We recalculated the value of the information residual of the compounds according to (5.5). For compounds with polymorphemic constituents, we considered the paradigms above the node where the compounds constituents are joined to be semantically distant, while the rest of the levels below and including that node were considered semantically close. The fourth columns in Tables 5.2 and 5.3 report the results of the regression analyses including this modification on the information residual measure. Observe that the information residual measure (labelled I'_R in the tables to distinguish it from the previous value) now significantly outperforms the traditional measures in terms of explained variance.

General Discussion

In this paper we have shown that the effects on visual lexical decision response latencies by frequency counts such as surface frequency and base frequency, inflectional ratio, cumulative root frequency, and morphological family size, can be accounted for in a more parsimonious manner using the information residual of a word. This measures the cost of recognizing a word considering the reductions and increases in uncertainty provided by the morphological paradigms to which it belongs.

The information residual measure performs at least as well as a combination of the traditional counts in predicting response latencies in visual recognition of monomorphemic words, polymorphemic words, and compounds. This is interest-

ing because no other single measure is able to quantify the processing cost of the wide range of different types of words we have studied (monomorphemic, polymorphemic, and compounds). The use of the information residual measure presents several theoretical and practical advantages. First, on the practical side, the previously described frequency counts are strongly correlated with one another. This causes very high degrees of collinearity in any regression model that includes them separately. Such high collinearity makes regression models unreliable, as it makes it difficult to assess the real magnitude and direction of the effects (Belsley, 1991). However, the present measure presents a way of combining the previous measures into a single count, hence avoiding the collinearity problem.

On the theoretical side, an important advantage of the information residual measure above the traditional counts is that it provides a parsimonious description of inflectional and derivational paradigms. In our approach, the effects on response latencies of the different counts, such as type-based morphological family size for derivational paradigms and token-based base frequency for inflectional paradigms, arise as by-products of the quantitative differences that are found between the statistical distributions within the paradigms. Also, the uniform account of nesting paradigm structures offers a straightforward approach to deal with languages in which even inflectional paradigms are nested within each other, as is the case in strongly inflectional languages such as those of the Romance and Slavic families, or in agglutinative languages such as Finnish.

Additional evidence for frequency counts reflecting informational complexity has been put forward by McDonald and Shillcock (2001). They report that a measure they call Contextual Distinctiveness is a better predictor of reaction times than is word frequency. Interestingly, their measure is based on the entropy of the distribution of the different contexts in which a word is used, what we could call its syntagmatic entropy. We believe that this measure is highly related to the first component of the information residual, the amount of information contained by the word. It will be interesting to investigate whether that component can be substituted for contextual distinctiveness in order to express the complexity of recognizing a word in terms of a combination of its paradigmatic and syntagmatic entropies.

The effect of the information residual of a word has important consequences for theories of morphological processing. First, and most evidently, the information residual is a purely probabilistic measure, and reflects how the human language processing system is extremely sensitive to stochastic factors. A second consequence comes from the markedly hierarchical nature of morphological paradigms.

The effect is not driven by individual relations between pairs of words, but by relations between hierarchically organized paradigms. The third consequence arises from the sensitivity to graded semantic similarity between paradigms, as indicated by the opposite influences of the morphologically closer and more distant paradigms to which a compound word belongs. This is also in accordance with the converging evidence for the graded influence of semantic similarity in inflectional (Baayen & Moscoso del Prado Martín, 2003; Kostić et al., in press; Ramscar, 2002) and derivational processes (De Jong, 2002; Moscoso del Prado Martín, Bertram, et al., 2003). These properties emerge naturally in distributed connectionist models of lexical processing (e.g., Gaskell & Marslen-Wilson, 1997; Plaut & Booth, 2000; Plaut & Gonnerman, 2000; Seidenberg & Gonnerman, 2000), which view morphological relations as the simultaneous effect of similarity between distributed representations of form and meaning. In this respect, McKay (2003) indicates that information theory is indeed the apropriate tool for analyzing the behavior of a neuronal network. Additionally, Moscoso del Prado Martín and Baayen (2003) shows how paradigmatic effects of this kind arise in a connectionist model that is trained to map distributed representations of form into distributed representations of meaning.

In summary, our approach succeeds in extending the ideas of Kostić (Kostic, 1991; 1995; 2003; Kostic et al., in press) from inflectional to derivational morphology and compounds while at the same time using a more standard measure of the amount of information carried by a paradigm. Nevertheless, further refinements of the technique are anticipated to account for the fine-grained sensitivity to semantic relatedness as revealed by the semantic constraints on morphological family size and inflectional morphology. Ideally, paradigm membership should not a be binary function, but should instead be defined as a continuous function of semantic relatedness and form similarity.

References

Alegre, M. and Gordon, P.: 1999, Frequency effects and the representational status of regular inflections, *Journal of Memory and Language* **40**, 41–61.

Baayen, R. H. and Moscoso del Prado Martín, F.: 2003, Questioning the unquestionable: Semantic density and past-tense formation in three Germanic languages, *Manuscript submitted for publication, Max Planck Institute for Psycholinguistics*.

Baayen, R. H., Lieber, R. and Schreuder, R.: 1997, The morphological complexity of simplex nouns, *Linguistics* **35**, 861–877.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX lexical database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baayen, R. H., Tweedie, F. J. and Schreuder, R.: 2002, The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon, *Brain and Language* **81**, 55–65.

Belsley, D. A.: 1991, *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, Wiley, New York.

Colé, P., Beauvillain, C. and Segui, J.: 1989, On the representation and processing of prefixed and suffixed derived words: A differential frequency effect, *Journal of Memory and Language* **28**, 1–13.

Cover, T. M. and Joy, A. T.: 1991, *Elements of Information Theory*, John Wiley & Sons, New York.

De Jong, N. H.: 2002, *Morphological Families in the Mental Lexicon*, MPI Series in Psycholinguistics, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M. and Baayen, R. H.: 2002, The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects, *Brain and Language* **81**, 555–567.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: 2000, The morphological family size effect and morphology, *Language and Cognitive Processes* **15**, 329–365.

Ford, M. A., Marslen-Wilson, W. D. and Davis, M. H.: in press, Morphology and frequency: Contrasting methodologies, *in* R. H. Baayen and R. Schreuder (eds), *Morphological structure in language processing*, Mouton de Gruyter, Berlin.

Gaskell, M. G. and Marslen-Wilson, W.: 1997, Integrating form and meaning: A distributed model of speech perception, *Language and Cognitive Processes* **12**, 613–656.

Hay, J.: 2001, Lexical frequency in morphology: Is everything relative?, *Linguistics* **39**, 1041–1070.

Kostić, A.: 1991, Informational approach to processing inflected morphology: Standard data reconsidered, *Psychological Research* **53**(1), 62–70.

Kostić, A.: 1995, Informational load constraints on processing inflected morphology, *in* L. B. Feldman (ed.), *Morphological Aspects of Language Processing*, Lawrence Erlbaum Inc. Publishers, New Jersey.

Kostić, A.: 2003, The effects of the amount of information on processing of inflected morphology, *Manuscript submitted for publication, University of Belgrade*.

Kostić, A., Marković, T. and Baucal, A.: in press, Inflectional morphology and word meaning: orthogonal or co-implicative domains?, *in* R. H. Baayen and R. Schreuder (eds), *Morphological structure in language processing*, Mouton de Gruyter, Berlin.

Lüdeling, A. and De Jong, N. H.: 2002, German particle verbs and word-formation, *in* N. Dehé, R. Jackendoff, A. McIntyre and S. Urban (eds), *Verb-particle explorations*, Mouton de Gruyter, Berlin, pp. 315–333.

McDonald, S. and Shillcock, R.: 2001, Rethinking the word frequency effect: The neglected role of distributional information in lexical processing, *Language and Speech* **44**, 295–323.

McKay, D. J.: 2003, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge, U.K.

Moscoso del Prado Martín, F. and Baayen, R. H.: 2003, BC–DC: A broad-coverage distributed connectionist model of visual word recognition, *Manuscript, Max Planck Institute for Psycholinguistics*.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R. and Baayen, R. H.: 2003, Morphological family size in a morphologically rich language: The case of finnish compared to dutch and hebrew, *Manuscript submitted for publication, Max Planck Institute for Psycholinguistics*.

Moscoso del Prado Martín, F., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H. and Baayen, R. H.: 2003, Changing places: A cross-language perspective on frequency and family size in Hebrew and Dutch, *Manuscript submitted for publication, Max Planck Institute for Psycholinguistics*.

Neijt, A., Schreuder, R. and Baayen, R. H.: 2003, Verpleegsters, ambassadrices, and masseuses. stratum differences in the comprehension of Dutch words with feminine agent suffixes, *in* L. Cornips and P. Fikkert (eds), *Linguistics in the Netherlands 2003*, Benjamins, Amsterdam, pp. 117–127.

Pinheiro, J. C. and Bates, D. M.: 2000, *Mixed-effects models in S and S-PLUS*, Statistics and Computing, Springer, New York.

Plaut, D. C. and Booth, J. R.: 2000, Individual and developmental differences in semantic priming: Empirical and computational support for a single mechanism account of lexical processing, *Psychological Review* **107**, 786–823.

Plaut, D. C. and Gonnerman, L. M.: 2000, Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing?, *Language and Cognitive Processes* **15**(4/5), 445–485.

Ramscar, M.: 2002, The role of meaning in inflection: Why the past tense doesn't require a rule, *Cognitive Psychology* **45**, 45–94.

Rubenstein, H. and Pollack, I.: 1963, Word predictability and intelligibility, *Journal of Verbal Learning and Verbal Behavior* **2**, 147–158.

Scarborough, D. L., Cortese, C. and Scarborough, H. S.: 1977, Frequency and repetition effects in lexical memory, *Journal of Experimental Psychology: Human Perception and Performance* **3**, 1–17.

Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **37**, 118–139.

Seidenberg, M. S. and Gonnerman, L. M.: 2000, Explaining derivational morphology as the convergence of codes, *Trends in Cognitive Sciences* **4**(9), 353–361.

Shannon, C. E.: 1948, A mathematical theory of communication, *Bell System Technical Journal* **27**, 379–423.

Shapiro, B. J.: 1969, The subjective estimation of word frequency, *Journal of verbal learning and verbal behavior* **8**, 248–251.

Taft, M.: 1979, Recognition of affixed words and the word frequency effect, *Memory and Cognition* **7**, 263–272.

Whaley, C. P.: 1978, Word-nonword classification time, *Journal of Verbal Language and Verbal Behavior* **17**, 143–154.

Accumulation of Expectations

CHAPTER 6

A revised version of this chapter will appear as Fermín Moscoso del Prado Martín, Robert Schreuder and R. Harald Baayen: Using the structure found in time: Building real-scale orthographic and phonetic representations by Accumulation of Expectations, *in* H. Bowman and C. Labiouse (eds.), *Connectionist Models of Cognition, Perception and Emotion: Proceedings of the eighth Neural Computation and Psychology Workshop*. Singapore: World Scientific

Abstract

This study introduces the Accumulation of Expectations technique to build vectorial representations of the orthographic and phonetic forms of all the words in a language. We demonstrate how this technique can be used to build realistic orthographic representations for all Dutch and English words from the CELEX database, and realistic phonetic representations for all Dutch words in CELEX.

Introduction

The representation of the orthographic or phonetic forms of words to be used as input and output has long been a problem for connectionist models of language processing. The inherently sequential nature of human language is a factor that needs to be taken into account by paradigms to represent word forms. Some authors (e.g., Gaskell & Marslen-Wilson, 1997; Harm & Seidenberg, 1999, 2001; Plaut & Gonnerman, 2000; Plaut & Booth, 2000; Plunkett & Juola, 1999, Shillcock, Ellison, & Monaghan, 2000; Shillcock & Monaghan, 2001) have made use of templates that predefine slot-based templates for the sequence of letters or phonemes of which words consist. Such a technique limits the maximum length of a word to the number of slots representing letters, phonemes or morphological constituents that are predefined in the representation. Additionally, templates assume a predefined 'possible word' structure that requires preprocessing of the words in order to fit them into the templates. This introduces the additional problem of alignment by which, depending of the goal of the model, the words need to be aligned to their beginnings, endings or word centers. Pinker & Ullman (2002) have criticized this approach for implicitly assuming symbolic processing. Other authors have tried different schemas, such as variants of the 'wickelgraph' (cf., Wickelgren, 1969) used in many models (e.g., Mozer, 1987; Rumelhart & McClelland, 1986; Seidenberg & McClelland, 1989), which introduce units corresponding to sequences of letters, phonemes, or phonetic features. These approaches are not only unrealistic, but also they are not capable of unambiguously representing all words in a language (Prince & Pinker, 1988). In our studies, we require a paradigm to represent unambiguously the formal properties of all words in a language. This paradigm needs to yield word representations that encode in a realistic way the form similarities between the words, and needs to be suitable to be used as input or output of backpropagation networks. Additionally, the similarity spaces must be created automatically, without manual pre-classiffication of the different segments, and need to be language specific, to capture the language-specific way in which two words may be similar. Finally, a continuous measure of word similarity between the representations of pairs of words is required. For these reasons, we developed the Accumulation of Expectations paradigm. This paradigm produces representations of word forms that fulfil our requirements.

General Description

Although the number of possible sequences of letters or phonemes that can form words in a given language is virtually unbounded, there are clear statistical regularities that make some sequences more likely to occur than others. Simple Recurrent Networks (SRN; Elman, 1990; Jordan, 1986) are classical three-layered associative networks (Rumelhart, Hinton, & Williams, 1986) to which an additional layer of units called a “context” layer, has been added (see Figure 6.1). The purpose of this context layer is to provide a copy of the activation values of the “hidden” layer in the previous time step, which are used as additional inputs to the network. This allows the network to respond to a particular set of inputs taking into account historical information on the previous activation states of the network. Elman (1990; 1993) showed that an SRN trained on predicting the next element (phoneme, letter, or word) on a linguistic sequence develops sensitivity to the possible sequences in a language.

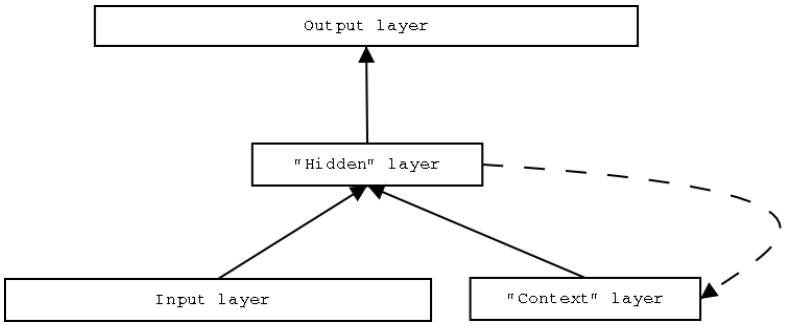

Figure 6.1: General architecture of a Simple Recurrent Network. The solid lines represent trainable all-to-all connections between the units in two layers. The dashed line indicates fixed-weight (non-trainable) one-to-one connections between the units of context and hidden layers.

Our approach to representing the orthographic and phonetic forms of words draws on the language-specific regularities that can be learned by an SRN. Both for the phonetic and the orthographic representations, we constructed SRN’s similar to the ones described by Elman, and we trained them on predicting the next letter or phoneme in the sequence using the whole set of phonetic or orthographic forms in English or Dutch. In order to speed up training, each word was presented to the network a number of times proportional to its logarithmic frequency of occurrence in the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995).

Once a network had learned about the regularities of word sequences in a particular language, we used the representations that were developed in its hidden layer

to create the vectors corresponding to individual words. Recall that the context layer allows a network to maintain a memory about the events (phonemes or letters) to which it had been exposed in the past, and that this memorized information about past-experience is fed into the hidden layer concurrently with the activation of each letter in the input. Therefore, ideally, the activation of the units in the hidden layer after the sequential presentation of all of a word's letters or phonemes should contain detailed information about all the previous letters or phonemes. However, this is not the case. Stojanov (2001) noted that SRNs do not perform well at reproducing Dutch monomorphemic words one phoneme at a time. For polymorphemic words, the problem is accentuated. As the network is trained on predicting letter or phoneme sequences from a particular language, the limitation of representational space in its hidden layer drives the networks to represent only that information about the past items in the sequence which is potentially useful for predicting the upcoming elements. This has the beneficial effect of giving rise to language-specific generalizations, but comes at the cost of actually 'forgetting' many details about the items that were encountered in the past when these details are not particularly informative about the elements that are to appear in the future. For instance, consider the morphologically complex English word *disclosure*. During presentation of this word to a network that has been trained in letter prediction on English words, the activation values at the network's hidden layer change in three stages. After presentation of the letters 'd' and 'i' at the beginning of the word, the network assigns a high probability to 's' being the next letter to appear in the input. This is because *dis* is a common prefix in English. However, after recognition of this prefix, i.e., immediately after presentation of 's' to the network, the amount of uncertainty increases again because there are many stems that could appear after *dis*, thus making it impossible to predict the following letter to come. This uncertainty decreases again after presentation of the 'c', the 'l' and the 'o', at which point the network assigns a high probability to the 's' in concordance with the many English words that can have *close* as their stem. Note at this point that for the prediction of the 's' in *close* the network would not need much information about *dis* having appeared before it. Therefore, because of the limitation of representational space, most of the information about the prefix *dis* would have disapeared at the point when the network encounters the 'u'. A similar process takes place here, once it is known that the stem was *clos*, the network would assign high probabilities to the initial letters of the possible suffixes that *close*, or a verbal stem of the same type, can take, and does not need to store detailed information about the preceding string (*disclos*).

In order to avoid this problem, instead of taking the activation of the hidden layer after presentation of the last letter or phoneme of the word as a word's representation, we represent words as the average of the activation values in the hidden layer after the presentation of each of the word's letters or phonemes. This approach ensures that information about each particular letter or phoneme is considered in the final representation, while at the same time giving different, context-sensitive representations to all the elements in the sequence.

At this point we need to address a technical problem. The activation of the hidden layer always contains at least some residual information about its activation values at the previous time steps. Consequently, plain accumulation of the activation values in the hidden layer would result in the information about letters or phonemes that appear early in the sequence receiving more weight in the accumulated representation than the letters or phonemes that appear later in the sequence. This has the undesirable effect of creating representations that give more importance to word beginnings than word centers, and more importance to word centers than word endings.

The solution to this technical problem is straightforward: We need a weighted averaging process such that the relative importance of the activation values increases with time. In other words, we modulate the contribution of the hidden layer at each time step to the accumulated representation by giving a higher weight to elements that appear late in the sequence than to elements that appear early in the sequence.

Technical Specification of the Network used for Orthographic Representation

Network Architecture

The network that was used for orthographic representations consisted of twenty-six input and output units each corresponding to one letter in the alphabet. The output layer contained an additional binary unit representing the end of a word. The network's hidden and context layers consisted of forty units each, with one-to-one non-trainable connections from the units in the hidden layer to those in the context layer. The units in the input and context layers had all-to-all trainable connections to the units in the hidden layer, which had themselves all-to-all connections to the units in the output layer.

Network Training

The networks were trained on letter prediction for all words in the CELEX database, that is, 297,690 words in the case of the Dutch network, and 154,063 words in the case of the English network. Letters were presented sequentially in the input by setting to one the value of the corresponding unit, while setting all other units to zero. At each time step, the network was trained to predict the next letter in the sequence (or the end of word) by activating the corresponding output unit. Errors were calculated using the divergence between the actual output of the network (after normalization) and the desired output (the next letter in the sequence). We trained the networks by the simple recurrent backpropagation through time algorithm with modified momentum descent (Rohde, 1999), with a learning rate of 0.04, a momentum of 0.9, and a weight decay of $1 \cdot 10^{-6}$.

Training proceeded by presentation of 10^6 words to the network. Words were chosen randomly, each having a probability of being presented proportional to the logarithm of its CELEX frequency. The activation values of the units of the context layer were reset to 0.5 after presentation of each word.

We trained two different networks, one on letter prediction for all Dutch words in the CELEX database, and the other on letter prediction for all English words in CELEX.

Technical Specification of the Network used for Phonetic Representation

Network Architecture

For the phonetic representations we built an SRN with fifteen input and output units each corresponding to one of the phonetic features described by Moscoso del Prado, Ernestus, and Baayen (2003). The output layer contained an additional binary unit representing the end of a word. The network's hidden and context layers consisted of forty units each, with one-to-one non-trainable connections from the units in the hidden layer to those in the context layer. The units in the input and context layers had all-to-all trainable connections to the units in the hidden layer, which had themselves all-to-all connections to the units in the output layer.

Network Training

We trained the networks on phoneme prediction for all 297,690 Dutch words in the CELEX database. Phonemes were presented sequentially in the input by setting to one the values of the units corresponding to their positive features, and to zero all remaining units. The training regime and parameters with which this network was trained were identical to those used in the networks for orthographic representations, with the only difference that we used the cross-entropy between the actual output of the network and the desired output features (the next phoneme in the sequence) as the error measure, instead of the divergence that we used in the orthographic network. [1]

Building the Orthographic and Phonetic Representations

After training the networks, the representation for a word was formed by presenting the word letter by letter (or phoneme by phoneme) at the network's input, and accumulating the activation values in the network's hidden layer at each time step. For words shorter than three elements (characters or phonemes), the activation values in all time steps received equal weights. If the word had $l > 3$ characters, the activation values of the hidden layer at time step i was weighted by:

$$w_i = 1 + \frac{\left(\frac{l-3}{l}\right)^2 \cdot i^3}{l^4 + l^3} \tag{6.1}$$

Before applying the weights, the values of the weights for all time-steps corresponding to a word were normalized to sum up to 1.0. The weighting schema was determined empirically, by testing the effect of different weighting schemas on clustering large number of words by prefix, suffix, and stem. The elements that appear at the end of the word receive a higher weighting than the rest. This increase in the weights given to the late elements is more marked in long words than in short words, and non-existent in words of three or less segments.

[1] This was done because, as the feature-based representation of a phoneme requires the activation of more than one unit per phoneme, the pattern of activation at the output layer does not correspond anymore to a simple distribution of probability.

Evaluation of the Representations

In order to use these representations in computational models of morphological processing, we need to ascertain whether the information contained in the vectors is sufficiently fine-grained to distinguish between words that share an affix or the stem. For each of the representations that we built, English and Dutch orthographic vectors, and Dutch phonetic vectors, we therefore examine whether the representations of words sharing a prefix, a suffix, or a word stem are systematically similar. Figures 6.2 to 6.10 compare the orthographic and phonetic representations of English and Dutch words sharing selected prefixes, suffixes, and word stems, using Principal Components Analysis (PCA), an unsupervised technique for tracing similarities between elements in multidimensional space. Each figure plots two sets of words, with each of the sets being composed of words that share an affix or a stem, in the plane defined by the two first principal components obtained from a PCA on their vectors. Figures 6.2 to 6.4, and Figures 6.5 to 6.7 show comparisons for orthographic representations of Dutch and English respectively. Figures 6.8 to 6.10 show clusterings arising from phonetic representations of Dutch words. Note that Figures 6.4 and 6.10 contrast the Dutch prefixes *on-* and *ont-* in their orthographic and phonetic representations. These two prefixes are extremely similar, with the first one being a substring of the second. Therefore there is a great deal of overlap in their representations. This is not surprising given that there are many Dutch words (e.g., *ontegenzeggelijk* – “indisputably”) that start with the prefix *on-* followed by a form starting with a ’t’, thus being exactly the same as if the word had had *ont-* as a prefix. Even in these very similar cases, the representations of words having these two suffixes are systematically different, as can be observed in the figures by the much coarser distribution occupied by the words starting with the prefix *ont-*. In fact, the examination of the values of the first principal component reveals differences in the distributions of both kinds of words, both for the orthographic representations ($t = -16.52, df = 347.73, p < 0.0001$), and for the phonetic representations ($t = 7.61, df = 213.11, p < 0.0001$).

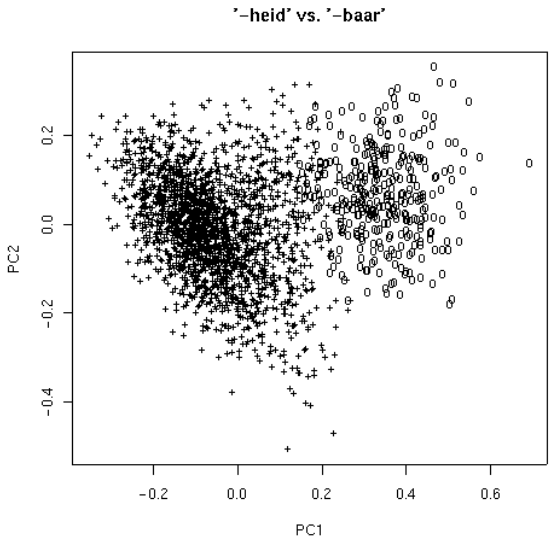

Figure 6.2: First two principal components contrasting the **orthographic** vectors of Dutch words ending in *-heid* (crosses) with those of Dutch words ending in *-baar* (circles).

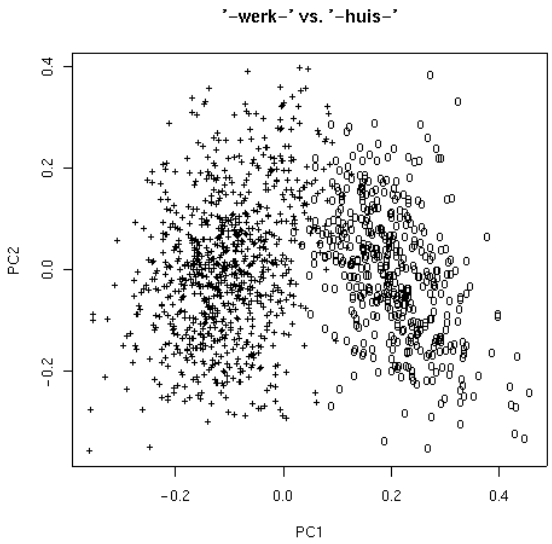

Figure 6.3: First two principal components contrasting the **orthographic** vectors of Dutch words that contain *-werk* (crosses) with those of Dutch words containing in *-huis-* (circles).

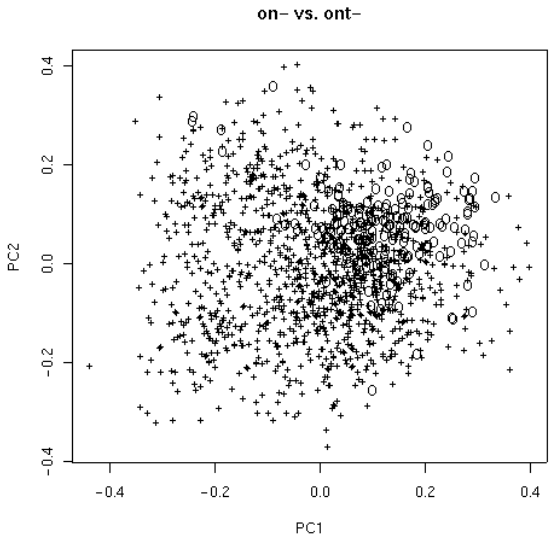

Figure 6.4: First two principal components contrasting the **orthographic** vectors of Dutch words starting with the Dutch prefix *on-* (crosses) with those of Dutch words starting with the Dutch prefix *ont-* (circles).

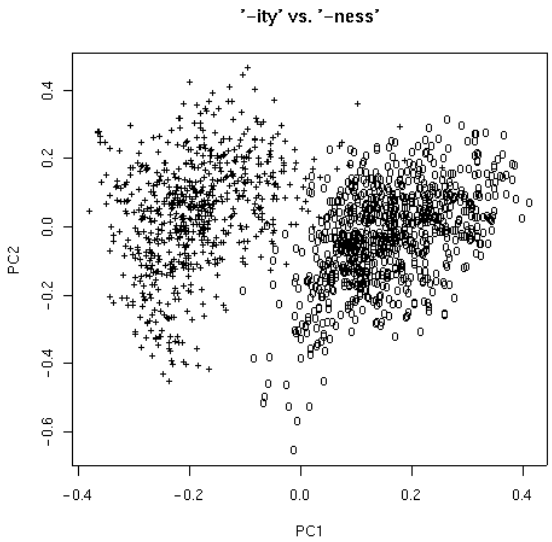

Figure 6.5: First two principal components contrasting the **orthographic** vectors of English words ending in the derivational suffix *-ity* (crosses) with those of English words ending in *-ness* (circles).

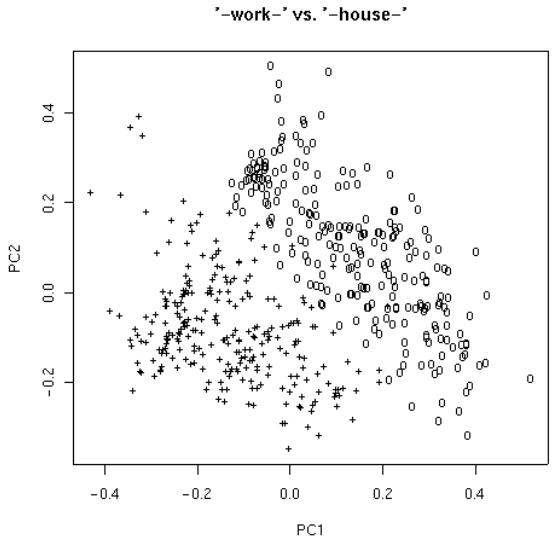

Figure 6.6: First two principal components contrasting the **orthographic** vectors of English words that contain *-work-* (crosses) with those of English words containing in *-house-* (circles).

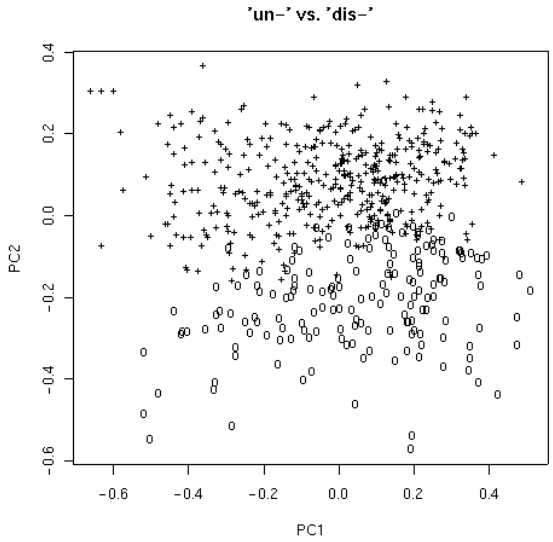

Figure 6.7: First two principal components contrasting the **orthographic** vectors of English words starting with the prefix *un-* (crosses) with those of English words starting in *dis-* (circles) (For visibility reasons we plot only a random sample of 600 points from the full data matrix).

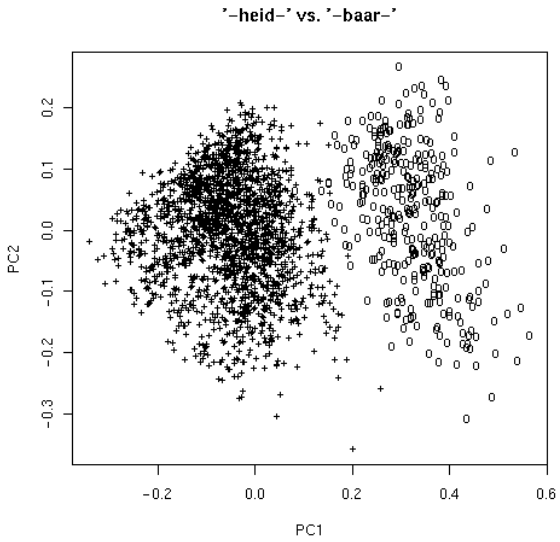

Figure 6.8: First two principal components contrasting the **phonetic** vectors of Dutch words ending in *-heid* (crosses) with those of Dutch words ending in *-baar* (circles).

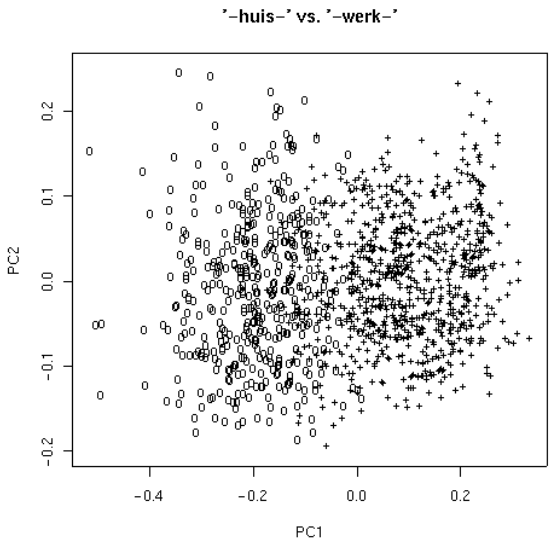

Figure 6.9: First two principal components contrasting the **phonetic** vectors of Dutch words that contain *-werk-* (crosses) with those of Dutch words containing in *-huis-* (circles).

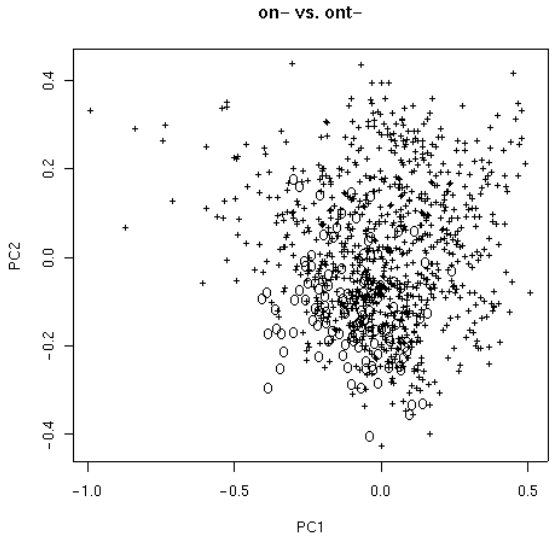

Figure 6.10: First two principal components contrasting the **phonetic** vectors of Dutch words starting with the Dutch prefix *on-* (crosses) with those of Dutch words starting with the Dutch prefix *ont-* (circles).

Conclusions

In this study we have shown that the Accumulation of Expectations paradigm provides a useful method for representing orthographic and phonetic word forms in a way that can be used in connectionist models, overcoming the word length and preprocessing problems of previous template-based approaches. This paradigm extends the work of Elman (1990; 1993) on SRNs to cover all words in a language. Additionally, accumulating the activation values of the hidden layer allows us to create a single vector representation for every word in a language. This enables us to calculate the distance between two word forms using a traditional distance measure such as the cosine. Stojanov (2001) showed that SRN's do not perform well in reproducing long words at their output one phoneme at a time. Instead we use an SRN to build the Accumulation of Expectations representations as they incrementally unfold, segment by segment. As a result we obtain a time-independent representation that can be used to reproduce a form in the output in a single time-step, while at the same time keeping track about the position dependent details of the segments. The Accumulation of Expectations technique provides representations that can be successfully used in connectionist models of language processing, as shown by the models described by Moscoso del Prado et al. (2003) and Moscoso del Prado and Baayen (2003). Our technique also provides a continuous, language specific, measure of word similarity that is capable of predicting human responses as used by Dijkstra, Moscoso del Prado, Schulpen, Schreuder, and Baayen (2003).

References

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX lexical database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Dijkstra, T., Moscoso del Prado Martín, F., Schulpen, B., Schreuder, R. and Baayen, R.: 2003, Family size effects in bilinguals, *Manuscript, University of Nijmegen*.

Elman, J. L.: 1990, Finding structure in time, *Cognitive Science* **14**, 179–211.

Elman, J. L.: 1993, Learning and development in neural networks: The importance of starting small, *Cognition* **48**, 71–99.

Gaskell, M. G. and Marslen-Wilson, W.: 1997, Integrating form and meaning: A distributed model of speech perception, *Language and Cognitive Processes* **12**, 613–656.

Harm, M. W. and Seidenberg, M. S.: 1999, Phonology, reading acquisition, and dyslexia: Insights from connectionist models, *Psychological Review* **106**, 491–528.

Harm, M. W. and Seidenberg, M. S.: 2001, Are there orthographical impairments in phonological dyslexia?, *Cognitive Neuropsychology* **18**, 71–92.

Jordan, M. I.: 1986, Serial order: A parallel distributed approach, *Institute for Cognitive Science Report 8604*, University of California, San Diego.

Moscoso del Prado Martín, F. and Baayen, R. H.: 2003, A broad-coverage distributed connectionist model of visual word recognition, *Manuscript, Max Planck Institute for Psycholinguistics*.

Moscoso del Prado Martín, F., Ernestus, M. and Baayen, R. H.: 2003, Do type and token effects reflect different mechanisms: Connectionist modelling of dutch past-tense formation and final devoicing, *Manuscript submitted for publication, Max Planck Institute for Psycholinguistics*.

Mozer, M. C.: 1987, Early parallel processing in reading: A connectionist approach, *in* M. Coltheart (ed.), *Attention and Performance XII: The psychology of reading*, Erlbaum, London, pp. 83–104.

Pinker, S. and Ullman, M.: 2002, The past and future of the past tense, *Trends in the Cognitive Sciences* **6**(11), 456–462.

Plaut, D. C. and Booth, J. R.: 2000, Individual and developmental differences in semantic priming: Empirical and computational support for a single mechanism account of lexical processing, *Psychological Review* **107**, 786–823.

Plaut, D. C. and Gonnerman, L. M.: 2000, Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing?, *Language and Cognitive Processes* **15**(4/5), 445–485.

Plunkett, K. and Juola, P.: 1999, A connectionist model of English past tense and plural morphology, *Cognitive Science* **23**(4), 463–490.

Prince, A. and Pinker, S.: 1988, Wickelphone ambiguity, *Cognition* **30**, 189–190.

Rohde, D. L. T.: 1999, LENS: The light, efficient network simulator, *Technical Report CMU-CS-99-164*, Carnegie Mellon University, Pittsburg, PA.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: 1986, Learning internal representations by error propagation, *in* D. E. Rumelhart and J. L. McClelland (eds), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, The MIT Press, Cambridge, Mass., pp. 318–364.

Rumelhart, D. E. and McClelland, J. L.: 1986, On learning the past tenses of English verbs, *in* J. L. McClelland and D. E. Rumelhart (eds), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models*, The MIT Press, Cambridge, Mass., pp. 216–271.

Seidenberg, M. S. and McClelland, J. L.: 1989, A distributed, developmental model of word recognition and naming, *Psychological Review* **96**, 523–568.

Shillcock, R., Ellison, T. M. and Monaghan, P.: 2000, Eye-fixation behaviour, lexical storage and visual word recognition in a split processing model, *Psychological Review* **107**, 824–851.

Shillcock, R. and Monaghan, P.: 2001, The computational exploration of visual word recognition in a split model, *Neural Computation* **13**, 1171–1198.

Stojanov, I. P.: 2001, *Connectionist Lexical Processing*, PhD thesis, Rijksuniversiteit Groningen, Groningen, The Netherlands.

Wickelgren, W. A.: 1969, Context-sensitive coding, associative memory, and serial order in (speech) behavior, *Psychological Review* **76**, 1–15.

Models of Dutch Past-Tense Formation and Final Devoicing

CHAPTER 7

This chapter will appear as Fermín Moscoso del Prado Martín, Mirjam Ernestus and R. Harald Baayen: Do type and token effects reflect different mechanisms? Connectionist modelling of Dutch past-tense formation and final devoicing. *Brain and Language*.

Abstract

In this paper, we show that both token and type based effects in lexical processing can result from a single, token-based, system, and therefore do not necessarily reflect different levels of processing. We report three Simple Recurrent Networks modelling Dutch past-tense formation. These networks show token-based frequency effects and type-based analogical effects closely matching the behavior of human participants when producing past-tense forms for both existing verbs and pseudo-verbs. The third network covers the full vocabulary of Dutch, without imposing predefined linguistic structure on the input or output words

Introduction

Dutch regular past-tenses are formed by adding *de* (/də/) or *te* (/tə/) to the verbal stem. The choice of the allomorph depends on the last phoneme of the stem by a simple rule: Verb stems ending in an underlyingly unvoiced obstruent take *te*, while all other verbs take *de*. For instance, as the stem of the verb *harken*, /hɑrkən/, "to rake" is /hɑrk/ ending in a unvoiced /k/, its singular past-tense form is *harkte*, /hɑrktə/, "raked". Similarly, the verb *zorgen*, /zɔrɣən/, "to care", with its stem /zɔrɣ/, underlyingly ending in voiced /ɣ/, has the singular past-tense form *zorgde*, /zɔrɣdə/, "cared".

Final devoicing introduces a complication to the rule of Dutch past-tense formation. In Dutch, all obstruents in word-final positions are realized as unvoiced (except before voiced plosives), independently of their underlying voice specification. This makes it impossible to infer the underlying voice specification of word-final obstruents from their acoustic realization. For example, due to final devoicing, the acoustic form [ɪk laːt] could be the first person singular form of either the verb *laden*, /laːdən/, "to load", or *laten*, /laːtən/, "to let". As a consequence, it is impossible to know which is the correct regular past-tense allomorph for a new or unknown Dutch obstruent-final verb, when the verb stem is presented auditorily without being followed by a vowel-initial suffix. Thus, the past-tense form of the pseudo-verb [dɑp] could either be *dabde* or *dapte*.

Ernestus and Baayen (2001, 2003) investigated how speakers of Dutch decide which is the past-tense form for existing verbs and pseudo-verbs. They asked Dutch participants to write down the regular past-tense forms of existing Dutch verbs and pseudo-verbs. For example, after hearing the pseudo-verb [ɪk dɑp], participants had to write down *ik dapte* or *ik dabde*, depending on which underlying voice specification they attributed to the final [p]. Similarly after hearing [ɪk dʏp] ("I doubt"), participants had to write down its correct regular past-tense form *ik dubde* ("I doubted"). Ernestus and Baayen (2003) found that Dutch speakers based their interpretation of the final obstruent of pseudo-verbs on the phonologically similar existing words. In general, for pseudo-words, participants interpret a final obstruent as voiced when the majority of the words in their lexicon with similar phonological properties end in a voiced obstruent, and they consider it to be unvoiced in the opposite case. In the [ɪk dɑp] example, most participants opted for *te*, since most final bilabial plosives following short vowels in existing words are unvoiced. The proportion of the *de* responses to pseudo-words reflected very closely the proportion of similar words ending in voiced obstruents in the lexicon. Ernestus and Baayen describe this ana-

logical effect as type-based, i.e., the proportion of *de* responses depends on the number of similar words ending in a voiced obstruent, independently of their frequencies of occurrence. Existing Dutch verbs were also sensitive to this analogical effect. For instance, 43% of the participants produced *ik dubte* as the past tense form for [ɪk dʏp], by analogy with the majority of existing forms in the lexicon such as *hup*, [hʏp], *stap*, [stɑp], *klap*, [klɑp], *stop*, [stɔp], . . . In addition to this type-based analogical effect, there was a token-based effect. High frequency past-tense forms were less sensitive to the analogical effect, producing fewer errors (Ernestus & Baayen, 2001).

The presence of both a type-based analogical effect and a token-based frequency effect can be interpreted as a reflection of two separate processing mechanisms. Some authors (e.g., Clahsen, 1999; Pinker, 1999; Pinker & Prince , 1994, 1998) propose a dual route system in which one route is a symbolic rule application mechanism that deals with the regular forms, and the other is an associative memory that stores the irregular forms. In such a model only the associative memory mechanism would be sensitive to token frequency effects. The mechanism in charge of the morphological generalizations (either rules or statistical analogies) has been argued to learn on the sole basis of type frequencies (e.g., Albright & Hayes, 2003; Bybee, 1995, 2001). These theories predict the existence of a rule application or analogical mechanism operating on types, which would be in charge of producing the regular forms, in combination with a token frequency sensitive mechanism which would be used to store and retrieve the memorized forms.

In contrast to the predictions of the traditional dual route mechanism, token frequency effects have been shown to affect the processing of morphologically regular forms as well (e.g., Baayen, Dijkstra, & Schreuder, 1997; Baayen, Schreuder, De Jong, & Krott, 2002, Schreuder, De Jong, Krott, & Baayen, 1999). This finding questions the clean separation between type-based application of regularities, and token-based memory storage of irregular forms. Additionally, Moscoso del Prado, Kostić, and Baayen (2003) report that an information-theoretical approach subsumes both type-based and token-based effects under a single measure of uncertainty. Although this measure is calculated on the basis of the token frequencies, it shows properties similar to those arising from type-based counts. Taken together, these two findings lead us to question the necessity of separating token sensitive and type sensitive mechanisms, indicating that both types of effects can be consequences of uncertainty in a purely token based approach.

Can token-based and type-based effects also arise as consequences of one

single processing mechanism, in line with the so-called ‘single route’ theories of lexical processing (e.g., McClelland & Patterson, 2002a, Plunkett & Juola, 1999, Rumelhart & McClelland, 1986)? In the present study, we address this question by modelling the experiments described by Ernestus and Baayen (2001, 2003) using a single route model. Ernestus and Baayen (2003) already studied several, non-connectionist, single route models that account for the analogical type-based effect. None of these models accounts for the token frequency effects reported by Ernestus and Baayen (2001) for the existing verbs.

Although previous connectionist systems have proved capable of learning morphological generalizations on the basis of token frequencies (e.g., Rumelhart & McClelland, 1986), these models have done so with restricted, very limited vocabularies, such as only monomorphemic or even monosyllabic stems. Additionally, most of the models have used training regimes in which words are not presented with their actual frequencies of occurrence, but on transformed frequency counts (e.g., the logarithm) that have the effect of mitigating the frequency differences between high and low frequency words. Moreover, most of these models use input-output templates, which impose predefined structure on their inputs, prescribing all the possible words in a language to conform to a given pattern (e.g., CCCVVCCC, etc.). These templates provide language-specific knowledge to the system, knowledge that ideally should be acquired from the input. Additionally, the templates require reclassifying and aligning the segments of the input words as onset consonants, vowels, or coda consonants, before presenting them to the system. All this built-in linguistic knowledge considerably oversimplifies the past-tense formation problem (Pinker & Ullman, 2002a).

In the present study, we describe connectionist models that deal with the full verbal past-tense formation system of Dutch. We describe three connectionist models of Dutch past-tense formation and evaluate their performance by having them produce past-tense forms for the stimuli in the experiments by Ernestus and Baayen (2001, 2003). Finally, we discuss the implications of the results of our models for theories of lexical processing.

Simulation 1

In the first simulation, we modelled Dutch past-tense formation with a Simple Recurrent Network (SRN; Elman, 1990, 1993). This network allows us to represent the input as a sequence of phonemes, without word length restrictions and without

any built-in assumptions about syllable structure.

Figure 7.1 shows the basic architecture of the SRN that we used in our simulations. This network consists of an input layer of 15 units, each of which represents a binary phonetic feature, plus an additional ‘end-of-word’ bit for triggering the output, and an output layer of 26 units representing the letters of the alphabet. Between the input and output layers, there is a small hidden layer of 10 units. The outputs of all the units in the input layer are connected to the inputs of all units in the hidden layer, and the outputs of all units in the hidden layer are connected to the inputs of all units in the output layer. There is an additional context layer, which represents a copy of the hidden layer in the previous state in time. The outputs of the units in the context layer have all-to-all connections with the inputs of the units in the hidden layer. This single recurrent loop allows the network to maintain a memory of the activation of the hidden layer in the previous time steps (Elman 1990, 1993).

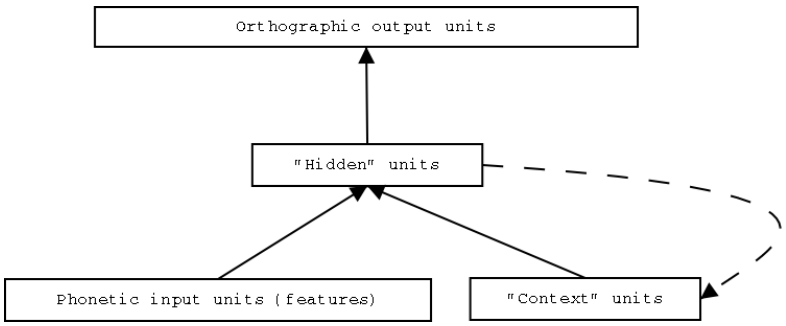

Figure 7.1: Modular architecture of a Simple Recurrent Networks used in the simulations. The boxes correspond to layers of units. The solid arrows represent sets of trainable ‘all-to-all’ connections between the units in two layers. The dashed arrow stands a for fixed ‘one-to-one’ not trainable connection between two layers. These connections have the function of copying the activation of the hidden units into the context units at every time step.

Presenting words phoneme by phoneme at the input simulates auditory input to the network. A word is represented by a sequence of phonemes presented at consecutive time steps. Each phoneme is presented by activating its corresponding phonetic features. Table 7.1 shows the phonetic features that we employed in our simulations. They distinguish between all Dutch phonemes, and take the phonology of Dutch into account. The feature matrix is similar to the one presented by Booij (1995) except that we replaced [coronal] by [alveolar] and [palatal] in order to be able to distinguish /s/ and /z/ from /ʃ/ and /ʒ/. We added the feature [tense] to express the difference in quality between long (tense) vowels and short (lax) vowels. Finally, we omitted the [aspiration] feature since /h/, the only phoneme for which it

is positive, can be uniquely defined without it.

Table 7.1: Phonetic features used for coding the input in the simulations.

Phoneme	consonant	sonorant	continuant	voicing	nasal	approximant	labial	coronal	palatal	dorsal	lateral	high	mid	back	round	tense
p	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
b	+	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-
t	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
d	+	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-
k	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
g	+	-	-	+	-	-	-	-	-	+	-	-	-	-	-	-
ŋ	+	+	+	+	+	-	-	-	-	+	-	-	-	-	-	-
m	+	+	+	+	+	-	+	-	-	-	-	-	-	-	-	-
n	+	+	+	+	+	-	-	+	-	-	-	-	-	-	-	-
l	+	+	+	+	-	+	-	+	-	-	+	-	-	-	-	-
r	+	+	+	+	-	+	-	+	-	-	-	-	-	-	-	-
f	+	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-
v	+	-	+	+	-	-	+	-	-	-	-	-	-	-	-	-
s	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-
z	+	-	+	+	-	-	-	+	-	-	-	-	-	-	-	-
ʃ	+	-	+	-	-	-	-	+	+	-	-	-	-	-	-	-
ʒ	+	-	+	+	-	-	-	+	+	-	-	-	-	-	-	-
x	+	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-
ɣ	+	-	+	+	-	-	-	-	-	+	-	-	-	-	-	-
j	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	-
h	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
w	+	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-
i	-	+	+	+	-	-	-	-	-	-	-	+	-	-	-	+
y	-	+	+	+	-	-	-	-	-	-	-	+	-	-	+	+
e	-	+	+	+	-	-	-	-	-	-	-	+	+	-	-	+
ø	-	+	+	+	-	-	-	-	-	-	-	+	+	-	+	+
a	-	+	+	+	-	-	-	-	-	-	-	-	-	+	-	+
o	-	+	+	+	-	-	-	-	-	-	-	+	+	+	+	+
u	-	+	+	+	-	-	-	-	-	-	-	+	-	+	+	+
ɪ	-	+	+	+	-	-	-	-	-	-	-	+	+	-	-	-
ɛ	-	+	+	+	-	-	-	-	-	-	-	-	+	-	-	-
ɑ	-	+	+	+	-	-	-	-	-	-	-	-	-	+	-	-
ɔ	-	+	+	+	-	-	-	-	-	-	-	-	+	+	+	-
ʏ	-	+	+	+	-	-	-	-	-	-	-	+	+	+	+	-
ə	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-

The network's output is also represented by a sequence, which in this case, is a sequence of letters. Although Dutch orthography is to a large extent transparent, with a quite regular grapheme to phoneme mapping, there are some irregularities making the orthographic transcription of one sound dependent on the following sounds. As a consequence, it is impossible to synchronize the network's input with the network's output, that is, having the system output a letter right after the corresponding phoneme is presented at the input. We therefore chose to start producing the output only after the full input had been received, and we added an additional trigger bit signaling the end of the word to the input layer of phonetic features. Only after this trigger node has been activated, we began recording (and training) the network's output.

A training regime aiming to reproduce the full letter sequence at the output would impose a tremendous amount of memory load on the network. Because of this memory load, SRN's do not perform well in reproducing full Dutch words one letter at a time, starting after having received the full input form (Stojanov, 2001). As our first investigation is concerned only with the interpretation of stem-final, neutral obstruents and the choice of the regular past-tense allomorph, we reduce the

memory load by training our network to produce only the last letter of the verb stem and the two letters of its past-tense allomorph. In this way, the network needs to store only those characteristics of the words that are relevant for the interpretation of the final obstruent and the choice of the past-tense allomorph, which leads to a drastic reduction in task complexity.

A model of past-tense formation can only be realistic if its training input is similar to the input that human speakers receive. It therefore has to be exposed to past-tense forms such that each form is presented a number of times that is proportional to its frequency of occurrence. Hence, in Simulation 1a, we used a token-based training strategy. In contrast, in Simulation 1b, we used a type-based training regime; that is, all verbs were presented to the network an equal number of times, independently of their frequency of occurrence. These two simulations allow us to investigate in detail the effects of using type-based and token-based training regimes.

Method

Materials. The phonemic representations for the words as available in CELEX (Baayen, Piepenbrock and Gulikers, 1995), determined the phonetic features activated in the input, according to Table 7.1. However, we made some systematic adjustments to the CELEX phonemic representations to make them more realistic phonetically. We doubled all phonetically long vowels in stressed positions: /a/, /e/, /o/, and /ø/ in stressed syllables were substituted for [aa], [ee], [oo] and [øø], respectively (cf., Rietveld, Kerkhoff and Gussenhoven, 1999). Moreover, we simulated the diphthongization of /e/, /ø/, and /o/ (Ernestus, 2000), by averaging the second part of the long vowel with [j] in the case of /e/, and with [w] in the cases of /o/ and /ø/. However, when the long stressed vowels preceded an /r/, we did not average the final part of the vowel with the glide, as in this context the vowels tend more to end in a schwa.

We used all 121,529 Dutch forms present in the CELEX lexical database in a re-training phase. The purpose of this phase was to provide the networks with some basic information about Dutch orthography before training it on the actual past-tense formation task. The orthographic output form of the word on which the network was trained was reduced to its last letter in this pre-training phase.

For the training phase itself, we used all 2,957 first person singular present-tense Dutch verb forms with regular past-tenses and a frequency greater than zero in CELEX. The outputs at this phase were the last letter of the verb stem followed by

the two letters that form its regular past-tense allomorph (*de* or *te*).

Finally, for testing the networks, we used the 165 existing Dutch verbs from Ernestus and Baayen (2001) (which all had CELEX frequencies grater than zero and thus also appeared in the training set), and the 145 pseudo-verbs from Ernestus and Baayen (2003).

Procedure. We modelled the networks using the Light Efficient Network Simulator (LENS; Rohde, 1999). Training was done using the modified momentum descent algorithm described by Rohde and Plaut (1999), using cross-entropy as the error measure on normalized outputs. For the training, we used a learning rate of 0.04, and a momentum of 0.90.

In the pre-training phase, the words from the pre-training dataset were presented to both networks in a pseudo-randomized order, each word being presented to the networks a mean number of times that was equal to its logarithmic surface CELEX frequency. In total, the number of word tokens presented to the networks equaled the number of word types in the data set, that is, 121,529. This pre-training went on for 100 epochs: The networks were presented with 121,529 chosen word tokens for 100 times. During this pre-training phase, the input and output weights to and from three of the ten hidden units were frozen, and thus were not affected by pre-training. In this way, we guaranteed that not all of the networks' memory would be used in learning Dutch orthography, leaving some space for the actual past-tense formation problem. The weights to and from the remaining seven units were adjusted on the basis of their orthographical outputs for the final letters of the words.

After the pre-training phase, the weights of all units in the hidden layer were released, allowing training to proceed in all of them during the training phase. Then, one network was trained with a type-based regime, and the other with a token-based regime. Both networks were trained to produce the last letter of the verb stem and the letters of the past-tense suffix for all the verbs in the training set. The networks received the first person singular present-tense forms of these verbs phoneme by phoneme, one at a time. The networks' weights were adjusted on the basis of their outputs for the stem-final segments and the past-tense allomorphs.

The network from Simulation 1a (token-based) was trained for eight epochs, in which examples were randomly chosen from the experimental list according to their frequencies. After eight epochs, the training was stopped because further training seemed to impoverish the network's performance (over-training), as appeared from a small test set of zero-frequency verbs from the CELEX database (which were not

present in the training set or in the experimental datasets). After training, the mean square error per output unit on the training set was 0.0043 (equivalent to plus or minus 6.56% of each unit's correct activation value).

The network from Simulation 1b (type-based) was trained for seven epochs, after which its performance on the verbs in the small testing set seemed to drop. In this Simulation, words were presented randomly according to a uniform probability distribution (all words were on average presented an equal number of times per epoch). After training, the mean square error per output unit on the training set was 0.0045 (equivalent to plus or minus 6.71% of each unit's correct activation value).

Testing proceeded in the same way in both networks. We presented the verbs from the two experimental datasets, introducing the first person singular present-tense form one phoneme at a time. Once the trigger bit had been activated, we started recording the activation of the output units in that time-step and the two following steps. These three time-steps corresponded to the last letter of the stem and the two letters of the past-tense suffix. For the first letter of the past-tense suffix, given that the output could only be 't' or 'd', it was not necessary to record the activation at the nodes representing the letters other than 't' or 'd', which was zero in all cases. Note that the activations of the 't' and 'd' nodes gave us an estimation of the probabilities of choosing between the *de* and *te* suffixes as estimated by the network.

Results and Discussion

We started our evaluation by attributing *de* as the network's response when the activation of the 'd' output node was greater than the activation of the 't' output node, and *te* in the opposite cases. We compared these networks' choices with the majority choices of the participants in the experiments reported in Ernestus and Baayen (2001, 2003). Both networks showed above chance agreement with the participants according to the κ statistic for inter-rater agreement (Guggenmoos-Holzman, 1996). The token-based model showed a coherence score with the participants' majority choices on the pseudo-words of 79% ($\kappa = 0.50, SE = 0.08, Z = 6.44, p < 0.0001$) and 78% on the existing words ($\kappa = 0.51, SE = 0.07, Z = 7.28, p < 0.0001$). The network that received a type-based training outperformed the one that received a token-based training on the pseudo-verbs. It showed a coherence score on pseudo-words of 91% ($\kappa = 0.76, SE = 0.08, Z = 9.14, p < 0.0001$) and of 78% on the existing words ($\kappa = 0.53, SE = 0.07, Z = 7.23, p < 0.0001$).

The Spearman rank correlation coefficients between the activation of the net-

works' 'd' output nodes and the proportion of *de* responses were very similar for the two networks, both for the pseudo-verbs ($r_s = 0.63, p < 0.0001$ token-based, versus $r_s = 0.65, p < 0.0001$ type-based), and the existing Dutch verbs ($r_s = 0.70, p < 0.0001$ token-based, versus $r_s = 0.69, p < 0.0001$ type-based), indicating that the two networks provide an equally good fit to the participants' responses. We conclude that both training regimes result in similar performances in terms of reproducing the participants' behavior on the experimental tasks, with a slight advantage on the pseudo-words for the network that was trained with a type-based training regime. Since both networks perform well on the pseudo-verbs, we can conclude analogical behavior, like the participants did.

When we investigated whether the networks also showed the purely token-based surface frequency effects found by Ernestus and Baayen (2001) for the existing verbs, we obtained very different results. While the log frequency of a past-tense form correlated with the average number of non-standard responses produced by the participants for that form ($r_s = -0.24, p = 0.0016$), this correlation only reached significance in the token-based training regime ($r_s = -0.30, p = 0.0002$), and not in the type-based training regime ($r_s = -0.08, p = 0.32$).

These two models have a number of limitations. The type-based training is unnatural, as actual word frequencies are not uniformly distributed. In addition, both models under performed in producing past-tenses for existing verbs: The average coherence score between a given participant's choice of past-tense allomorph and the allomorph chosen by the majority of the other participants was 90%, which is significantly above the performance of both networks. This is probably due to the small memory of the networks, in which only three hidden units bear most of the burden of past-tense formation. This substantial limitation on representational space does not allow the networks to form individual lexical representations for existing verbs, and analogical generalizations at the same time. This also explains why the type-based training regime showed an astonishing performance on pseudo-words, indicating that it succeeded in capturing the analogical effects, while it had a paradoxically lower performance on producing the past-tenses of the words on which it had been trained.

The present type and token-based models are limited also in other respects. By training the models on producing the last letter of the stem followed by the past-tense allomorph, we have implicitly provided the information that the last segment of the stem is crucial for the choice of the past-tense allomorph. This simplified the task by inducing the networks to base their choice of past-tense allomorph

primarily on the last letter of the stem, without taking other properties of the verbs into account.

In addition, the input to the models during training was restricted to singular regular past-tenses. The exclusion of the irregulars decreased memory load, but oversimplified the Dutch past-tense formation problem. At the same time, the exclusion of regular plural forms from the networks' training entailed that the networks had no access to a potential natural source of information on past-tense formation. Dutch regular plural present-tense forms consist of the verb stem followed by the suffix *en*. The stem-final obstruent before this suffix is not devoiced, and therefore it accurately predicts the appropriate past-tense allomorph for all forms of that particular verb (when unvoiced the allomorph is *te*, otherwise it is *de*). Given the strong degree of similarity between the singular and plural regular past-tense forms of the same verb (they only differ orthographically in the presence of the letter 'n' at the end of the plural), the information about the plural facilitates formation of the past-tense for the singular.

We conclude that a more realistic model of Dutch past-tense formation requires a larger memory, that it should produce full verbal forms, and, finally, that it should also be trained to produce plural and irregular past-tenses as well.

Simulation 2

There are two possibilities for a network that is required to output full verbal forms. Either the output has to be sequential, which is problematic (Stojanov, 2001), or the full form has to be output in a single time-step. Some representational paradigms have been used for single time-step output. These paradigms use either templates (e.g., Plunkett & Juola, 1999) or 'wickelfeatures' (e.g., Rumelhart & McClelland, 1986). Templates include crucial information about word structure specific to a particular language, and require unrealistic preprocessing of the inputs by alignment, etc.. They also impose explicit constraints on word length, which makes them unsuitable for our task. Wickelfeatures have been claimed to represent strings of unlimited length but they are in fact incapable of representing unambiguously all strings in a language (Pinker & Prince, 1988).

We therefore made use of the 'Accumulation of Expectations' technique (AoE; Moscoso del Prado, Schreuder, & Baayen, 2003). This technique is inspired by the simulations described by Elman (1990,1993), which show that, when an SRN is trained on predicting the next letter in a sequence, using a large enough corpus for

training, it develops in its hidden layer a detailed representation of the orthography of the language. The AoE for a word is the activation of the hidden units of an SRN trained to predict the next letter in a sequence of letters, summed across the presentation of each letter of the word. This vector thus gives a distributed representation of the orthographic form of the word. The AoE technique allows us to create vectors representing the orthographical forms of Dutch words (and pseudo-words) of any length, implicitly incorporating the generalizations of the orthographical system of Dutch.

In contrast to Simulation 1, the training input contained all first and third person present and past verbal forms, regulars or irregulars. We used a past-tense formation SRN very similar to the ones used in Simulation 1. The phonetic input layer remained exactly the same as in Simulation 1, using the same feature-based representations of the input together with the trigger bit. We extended the memory to 200 units in the hidden and context layers, as the model has to learn mappings for full words, including irregulars. The output layer consisted of 50 units that represent the AoE of the input words.

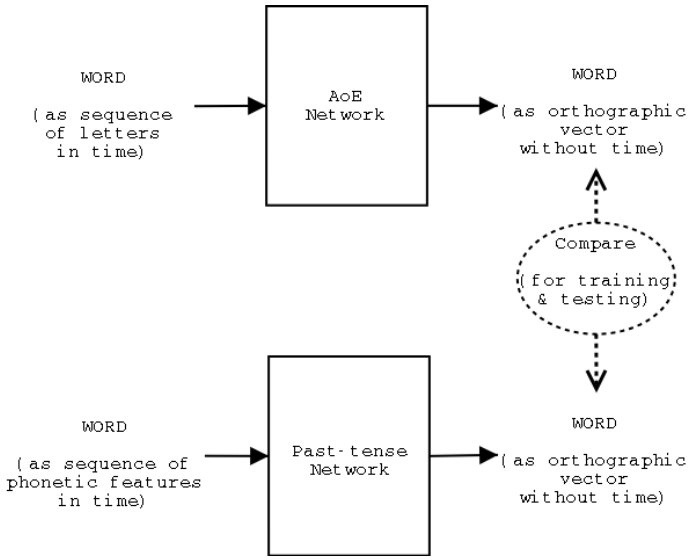

Figure 7.2: Outline of the complete model used in Simulation 2. The upper half of the diagram represents the AoE network that created orthographic vectors corresponding to sequences of letters. These vectors were used for training and evaluating the performance of the past tense network depicted in the lower part of the diagram

Figure 7.2 provides an outline of the complete model that was used in the present simulation. The model consisted of two parts: An AoE network, that was used to create flat orthographic representations for the words and pseudo-words (upper

part of the figure), and a past-tense formation network similar to the one employed in Simulation 1. The representations created by the AoE network were employed for evaluating the outputs of the past-tense formation network, that is, for training, testing, and interpreting the outputs of this network.

Method

Materials. The AoE network was an SRN with 50 units in its hidden and context layers. The input and output layers consisted of 26 units, each of these corresponding to one letter of the Dutch alphabet. We trained this network on predicting the next letter, using all 297,690 Dutch words in CELEX as training set. The words were presented letter by letter, and at each time-step the AoE network was trained on predicting the next letter in the word. After training the AoE network, we ran through it all Dutch words in the CELEX database, and we accumulated the activation vectors produced in the hidden layer after each phoneme. In this way, we obtain a vector of 50 numbers for each word, with values in the interval $[0.0, 1.0]$. The past-tense formation network in this simulation was trained to produce these same vectors as its output orthographic representations. A more detailed description of this AoE network can be found in Moscoso del Prado, Schreuder, and Baayen (2003).

For the pre-training phase of the past-tense formation network, we used again all 121,529 Dutch words from the CELEX lexical database with frequencies higher than zero. The phonological coding of the input words was the same as in Simulation 1. Outputs were coded using the AoE technique.

For the training phase of the past-tense formation network, we selected all 10,750 present and past-tense, regular and irregular, Dutch first and third person verbal forms that appear in the CELEX database with a frequency higher than zero. The input consisted of the phonological form corresponding to the present-tense forms. The coding of the inputs was done according to the method described in Simulation 1, with an additional bit that was set to zero at the last phoneme when an identity mapping was the required task, in which case, the model performs a pure ‘dictation’ task. If the required output was the corresponding past-tense form, the bit was set to one.

Again, for testing the network, we used the 145 pseudo-verbs from Ernestus and Baayen (2003) and the 165 existing Dutch verbs from Ernestus and Baayen (2001). Input and output were coded in the same manner as in the training phase.

Procedure. We modeled the network using LENS. Training was done using the modified momentum descent algorithm (Rohde and Plaut, 1999), this time using cosine as our error measure. In all training phases, we used a learning rate of 0.04 and a momentum of 0.90.

The network first went through a pre-training phase of 100 epochs. This phase was identical to the one in Simulation 1, except that the network was trained to produce the full orthographic forms, instead of just the last letter. Each item was presented a number of times proportional to its logarithmic frequency. The input and output weights to and from one hundred of the two hundred hidden units were frozen, and thus were not affected by pre-training. Error was back-propagated after the presentation of the last phoneme of each word.

After the identity mapping pre-training phase, the weights of all units in the hidden layer were released, allowing training to proceed throughout the network. Training items were presented to the network in a random order a number of times that was directly proportional to the frequency of occurrence of the output form (present or past-tense). In this way we simulated the frequencies of production of present and past-tense forms. The network was trained for 2000 epochs. Training was stopped at this point because further training did not seem to improve performance, which was tested on the same set of zero-frequency verbs from Simulation 1. After training, the network's average cosine error on the training set was $0.0343\,(\pm 0.0254)$.

For testing, we presented the verbs from both experimental datasets, introducing the first person singular present-tense form (for both the pseudo-verbs and the existing verbs) one phoneme at a time. Once the output triggering bit had been activated, we recorded the activation of the output units in that time-step.

Results and Discussion

Once more, we evaluated our model by comparing the past-tense forms chosen by the majority of the participants with the preferred output of the model. Since the output of this network was a distributed AoE representation, instead of a localistic one, determining the preferred output of the network was more complicated than in Simulation 1. Using the AoE network described above, we created 50-element vectors for the two possible orthographic forms of the regular past-tense (ending in *te* or *de*) for each of the experimental items. For example, given the experimental stimulus [kɛit], the AoE network produced 50-element vectors for the possible output strings *keidde* and *keitte*. Additionally, to investigate the effects of irregularization, we also created vectors for two irregular forms for each item, one ending

in a voiced obstruent and the other ending in a unvoiced obstruent. The selection of these irregular past-tenses for the experimental items (that were either regular or non-existing verbs) was done by choosing the most likely vowel change pattern for the final vowel and consonant clusters of the verb (or pseudo-verb) according to the majority of existing irregular verbs in CELEX. For instance, for the experimental item [kɛit], we created AoE vectors corresponding to the orthographical forms *keet* and *keed*, because the majority of existing irregular verbs that contain the diphthong /ɛi/ undergo the /ɛi/ to /eː/ vowel change. We calculated the cosine distance between the output of the past-tense formation network for each experimental item, and the corresponding regular past-tenses ending in *de* and *te* as well as the two created irregular forms (which in most cases do not correspond to any existing Dutch word). Out of these four options, the form with the vector that had the smallest distance to the network's output was selected as the network's preferred choice.

In order to assess whether the network was indeed producing as its output one of these four possibilities, we also calculated the cosine distance between the output vector produced by the network for each experimental item (words and pseudo-words), and the orthographical vectors of all the 30,740 word forms in the CELEX database with a surface frequency of at least one occurrence per million. The cosine distance between the network's output and the closest form among the four predefined choices was smaller (or equal in cases of existing verbs) than the distance to the closest existing word from CELEX in 85% of the existing words and 63% of the pseudo-words. The outputs for the remaining 15% of words and 37% of pseudo-words were marked by deviant spellings, often caused by ambiguity in Dutch phoneme to grapheme mappings (e.g., the diphthongs /ɑu/ and /ɛi/ can be respectively spelled as *au* or *ou*, and *ei* or *ij* in Dutch). For some pseudo-words, the network produced intermediate representations between the different possibilities because these cases can only be resolved by memorizing the correct spelling for each word. Other errors involved small 'misperceptions', that is, pseudo-words were mistaken for similar-sounding existing words. For instance, when presented with [bɑus], the network produced the existing verb *bouwde* ("built"), instead of inflected versions of *baus* or *bous*. Finally, for a small proportion of the verbs, the network produced 'impossible' morphological forms, such as attaching *te* after a voiced consonant (e.g., *bembte*), or *de* after an unvoiced one (e.g., *daantde*). Interestingly, Ernestus and Baayen (2001) reported that participants also produced this type of errors.

The model chose an irregular past-tense for forty experimental items by changing the vowel without affixation of a past-tense allomorph. For instance, it created *keed* instead of *keidde* or *keitte*, as the past tense form of [kɛit], in analogy with existing Dutch verbs. Interestingly, although the training was token-based, and the irregular past-tenses were consequently more frequent than the regular ones, the network's preferred past-tense was irregular for only 25 out of 145 (14%) pseudo-verbs and 15 out of the 165 (9%) existing Dutch regular verbs, in line with percentages of irregularization reported in the literature (e.g., Albright and Hayes (2003) report 18.5% irregularization in two experiments on English pseudo-verbs). Some of the irregularizations are in fact correct as they correspond to existing homophonic irregular verbs (e.g., the network produced *liet*, the past-tense form of the verb *laten*, "to let", instead of *laadde*, the past-tense form of *laden*, "to load", upon presentation of [laːt]). Irregularization occurred more often when the irregularized form was also an existing (usually unrelated) word. In Ernestus and Baayen's experiments participants were explicitely instructed to produce regular past-tense forms in all cases. We therefore excluded these 40 irregularized items from the following analyses.

The comparison of the network's preferred choices ending in *te* or *de* with the participants' majority choices gave significantly above chance coherence scores of 68% for the pseudo-verbs ($\kappa = 0.33, SE = 0.07, Z = 4.5, p < 0.0001$) and 85% for the existing Dutch verbs ($\kappa = 0.70, SE = 0.08, Z = 8.6, p < 0.0001$). Interestingly, the participants in Ernestus and Baayen's experiments showed similar average coherence scores with each other: $74\%(\pm 5\%)$ for the pseudo-verbs, and $90\%(\pm 5\%)$ for the existing verbs. The coherence scores of the participants did not differ significantly from the coherence scores showed by the network, neither for the pseudo-words ($Z = -1.20, p = 0.1151$), nor for the existing verbs ($Z = -1.00, p = 0.1587$).

We also calculated the correlations between the logits for the participants' responses, that is, the logarithmic ratio between the number of *de* responses and the number of *te* responses for a given verb, and an estimation of the logit values for the network outputs. We estimated the network logits using $\hat{L} = \log \frac{d_{te}}{d_{de}}$, with d_{de} being the cosine distance between the *de* form and the network output, and d_{te} being the cosine distance between the *te* form and the network output. The Spearman rank correlations between the logits of the number of *de* and *te* responses produced by the participants, and the L values were $r_s = 0.50$ for the pseudo-verbs and $r_s = 0.76$ for the existing Dutch verbs ($p < 0.0001$ in both cases). The correlation between the participants' logits and our estimated network logits is illustrated in Figure 7.3. The regression line shows that the network's outputs replicate the participants' behav-

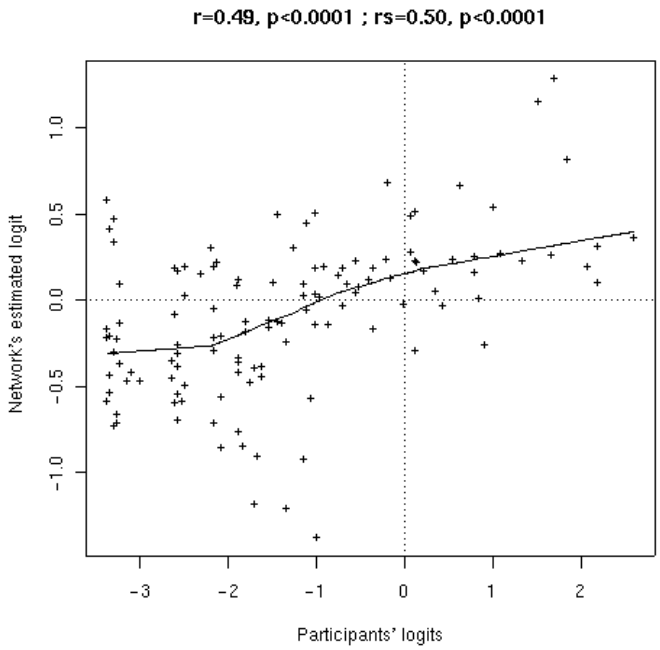

Figure 7.3: Comparison between the logit values of the proportion of *de* and *te* responses to pseudo-verbs by the participants in Ernestus and Baayen (2003), with the estimated logit values of the responses of the network in Simulation 2. The line represents a non-parametric regression (Cleveland, 1979).

ior: A linear increase of the network's logits is proportional to the linear increase of the participants' logits.

The correlation between the network's estimated logits and the participants' logits suggests that our network is showing the type-based analogical effects described by Ernestus and Baayen (2001, 2003). Figure 7.4 provides a more detailed account of these analogical effects. The figure compares the estimated logit values produced by the network, with the logits of the participant responses to different groups of pseudo-words classified by type of final obstruent of the pseudo-verb (upper panel), type of segment preceding the final obstruent of the pseudo-verb (middle panel), and the quantity of the final vowel (lower panel). Note that the network replicates accurately the patterns showed by the participants, with only small differences of scaling.

The network's error for each verb was estimated by the absolute value of the estimated network logit ($|\hat{L}|$). This estimate represents how confident the network was in its choice of past tense. High values should correspond to few errors produced by the participants. The Spearman correlation between this error estimate and the number of errors produced by the participants for the existing verbs was $r_s = -0.44$ ($p < 0.0001$). This correlation shows that the network was replicating not only the participants' preferred choices, but also the participants' certainty about a particular choice. Finally the network showed higher confidence for the more frequent verbs ($r_s = -0.55; p < 0.0001$), thus replicating the token-based frequency effect in the participants' responses.

To explore the network's performance on irregulars verbs, we selected the 153 Dutch monomorphemic irregular verbs from the CELEX lexical database (excluding the verbs that could have more than one past-tense form), and we ran them through the network in their first person singular present-tense form, producing their-past tenses. We compared the output of the network with the AoE vectors corresponding to the correct irregular past-tense form, and with the two possible regularizations (using the *te* or *de* allomorphs) for each verb. In 88% percent of the cases, the distance between the output of the network and the correct irregular past-tense form was smaller than the distance to any of the two possible regularizations. Additionally, we found that the network's output distance to the chosen past-tense form, was in 85% of the cases smaller or equal than the distance to the closest existing Dutch word with a frequency of at least one occurrence per million.

This new model succeeds in capturing type-based analogical effects and token-based frequency effects in a single system with a token-based training regime. This

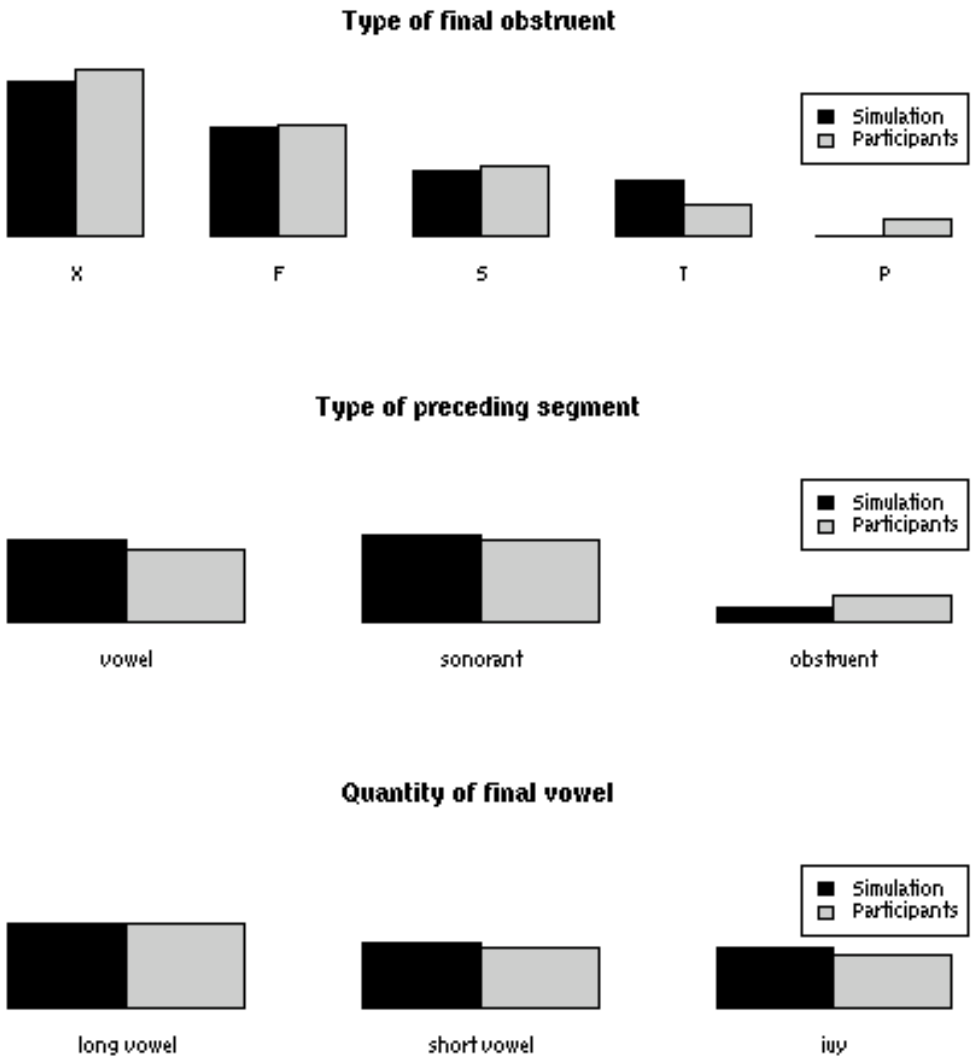

Figure 7.4: Comparison between the standardized logit values of the proportion of *de* and the *te* responses to pseudo-verbs by the participants in Ernestus and Baayen (2003), with the standardized logit values of the responses of the network in Simulation 2. The pseudo-verbs are classified by type of final obstruent (upper panel), type of segment preceding the final obstruent (middle panel), and quantity of the final vowel (lower panel).

shows that the combined presence of token-based and type-based effects does not necessarily imply two different mechanisms. The network's performance is remarkable given that, in this simulation, the task was much more complicated than in Simulation 1, requiring the learning of the full Dutch phoneme to grapheme mappings and the whole past-tense formation system, including irregulars and plurals, with a still very limited amount of memory.

General Discussion

In this paper, we discussed three neural networks modelling past-tense formation in Dutch. The first two networks dealt only with regular past-tense formation in the singular, one receiving a type-based training, the other one a token-based training. The inputs for both models consisted of featural phonetic representations of verbal stems simulating auditory input. As outputs, the models produced the final letters of the verbal stems followed by their past-tense allomorphs. The token-based model replicated the type and token based effects reported by Ernestus and Baayen (2001, 2003). The model that received a type-based training regime matched the experimental results for the pseudo-verbs in more detail, but it failed to replicate the token-based frequency effects for the existing verbs. In fact, both of the models showed a relatively low performance on modelling the participants' responses to existing verbs. We concluded that accurate modelling of past-tense formation is only possible when a not too small memory is available, allowing both for the storage of individual items, and for the formulation of generalizations.

The third model received only a token-based training regime. Its memory was much larger, and the training set contained both regular and irregular, and both singular and plural past-tense forms. Furthermore, the model not only chose a past-tense allomorph, but also provided full orthographic forms. This model displays the type-based analogical effect for both the existing verbs and the pseudo-verbs, together with the token-based frequency effect for the existing verbs, closely replicating humans' responses to those same items. We conclude that type and token based effects in morphological processing do not necessarily imply the existence of separate processing mechanisms. Type-based analogical effects can arise as a consequence of uncertainty in token-based probability distributions as it was proposed by Moscoso del Prado, Kostić, and Baayen (2003).

Our system contributes to the ongoing debate on single and dual route models for regular and irregular past-tense formation (for reviews see McClelland and

Patterson, 2002a, 2002b, Pinker and Ullman, 2002a, 2002b) by showing that both regulars and irregulars can be captured by a simple model trained on a realistic amount of different verbs according to the best available estimates of their frequencies. Our system produced regular past-tense forms for the great majority of pseudo-verbs, showing a default rule for past-tense formation. At the same time, it produced irregular past-tenses for existing irregular verbs, and for only a few pseudo-verbs with strong analogical support for the irregular form. This shows that analogical processing does not exclude rule-like generalizations. Default rules, in the sense of Pinker (1999), can arise in an exemplar-based connectionist system.

As far as we know, this is the first model of past-tense formation that covers all verbs of a language, irrespective of their length, regularity, or morphological complexity. Pinker and Ullman (2002a, 2002b) argue that previous connectionist models of past tense formation (e.g., Plunkett & Juola, 1999, Plunkett & Marchman, 1993, Rumelhart & McClelland, 1986) only succeed in this task because a great deal of linguistic knowledge was built into their systems, thus making them more similar to a symbolic model. In contrast, our third simulation succeeded in this task using only phonetic representations. It acquired the remaining structural knowledge by statistical generalizations over the phonological and orthographical sequences present in words, without making use of any implicit symbolic processing mechanism.

References

Albright, A. and Hayes, B.: 2003, Rules vs. analogy in English past tenses: A computational/experimental study, *Manuscript UCLA*.

Baayen, R. H., Dijkstra, T. and Schreuder, R.: 1997, Singulars and plurals in Dutch: Evidence for a parallel dual route model, *Journal of Memory and Language* **37**, 94–117.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX lexical database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baayen, R. H., Schreuder, R., De Jong, N. H. and Krott, A.: 2002, Dutch inflection: the rules that prove the exception, *in* S. Nooteboom, F. Weerman and F. Wijnen (eds), *Storage and Computation in the Language Faculty*, Kluwer Academic Publishers, Dordrecht, pp. 61–92.

Booij, G. E.: 1995, *The phonology of Dutch*, Clarendon Press, Oxford.

Bybee, J. L.: 1995, Diachronic and typological properties of morphology and their implications for representation, *in* L. B. Feldman (ed.), *Morphological Aspects of Language Processing*, Lawrence Erlbaum Associates, Hillsdale, N. J., pp. 225–246.

Bybee, J. L.: 2001, *Phonology and language use*, Cambridge University Press, Cambridge.

Clahsen, H.: 1999, Lexical entries and rules of language: a multi-disciplinary study of German inflection, *Behavioral and Brain Sciences* **22**, 991–1060.

Cleveland, W. S.: 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

Elman, J. L.: 1990, Finding structure in time, *Cognitive Science* **14**, 179–211.

Elman, J. L.: 1993, Learning and development in neural networks: The importance of starting small, *Cognition* **48**, 71–99.

Ernestus, M.: 2000, *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface*, LOT, Utrecht.

Ernestus, M. and Baayen, H.: 2001, Choosing between the Dutch past-tense suffixes *-te* and *-de*, *in* T. van der Wouden and H. de Hoop (eds), *Linguistics in the Netherlands 2001*, Benjamins, Amsterdam, pp. 81–93.

Ernestus, M. and Baayen, R. H.: 2003, Predicting the unpredictable: Interpreting neutralized segments in Dutch, *Language* **79(1)**, 5–38.

Guggenmoos-Holzmann, I.: 1996, The meaning of kappa: probabilistic concepts of reliability and validity revisited, *Journal of Clinical Epidemiology* **49**(7), 775–782.

McClelland, J. L. and Patterson, K.: 2002a, Rules or connections in past-tense inflections: what does the evidence rule out, *Trends in the Cognitive Sciences* **6**(11), 465–472.

McClelland, J. L. and Patterson, K.: 2002b, 'words or rules' cannot exploit the regularity in exceptions: Reply to Pinker and Ullman, *Trends in the Cognitive Sciences* **6**(11), 464–465.

Moscoso del Prado Martín, F., Kostić, A. and Baayen, R. H.: 2003, Putting the bits together: An information theoretical perspective on morphological processing, *Manuscript submitted for publication, Max Planck Institute for Psycholinguistics*.

Moscoso del Prado Martín, F., Schreuder, R. and Baayen, R. H.: 2003, Using the structure found in time: Building real-scale orthographic and phonetic representations by Accumulation of Expectations, *Manuscript submitted for publication, Max Planck Institute for Psycholinguistics*.

Pinker, S.: 1999, *Words and Rules: The Ingredients of Language*, Weidenfeld and Nicolson, London.

Pinker, S. and Prince, A.: 1988, On language and connectionism, *Cognition* **28**, 73–193.

Pinker, S. and Prince, A.: 1994, Regular and irregular morphology and the psychological status of rules of grammar, *in* S. Lima, R. Corrigan and G. Iverson (eds), *The Reality of Linguistic Rules*, John Benjamins, Amsterdam.

Pinker, S. and Ullman, M.: 2002a, Combination and structure, not gradedness, is the issue: Reply to McClelland and Patterson, *Trends in the Cognitive Sciences* **6**(11), 472–474.

Pinker, S. and Ullman, M.: 2002b, The past and future of the past tense, *Trends in the Cognitive Sciences* **6**(11), 456–462.

Plunkett, K. and Juola, P.: 1999, A connectionist model of English past tense and plural morphology, *Cognitive Science* **23**(4), 463–490.

Plunkett, K. and Marchman, V.: 1993, From rote learning to system building: acquiring verb morphology in children and connectionist nets, *Cognition* **48**, 21–69.

Rietveld, T., Kerkhoffs, J. and Gussenhoven, C.: 1999, Prosodic structure and vowel duration in Dutch, *in* J. Ohala, Y. Hasegawa, M. Ohala, D. Granville and A. Baily (eds), *Proceedings of the 14th International Congress of Phonetic Sci-*

ences, San Francisco, 1–7 August 1999, Linguistic Department, University of California, Berkeley, pp. 463–466.

Rohde, D. L. T.: 1999, LENS: The light, efficient network simulator, *Technical Report CMU-CS-99-164*, Carnegie Mellon University, Pittsburg, PA.

Rohde, D. L. T. and Plaut, D. C.: 1999, Language acquisition in the absence of explicit negative evidence: how important is starting small?, *Cognition* **72**(1), 67–109.

Rumelhart, D. E. and McClelland, J. L.: 1986, On learning the past tenses of English verbs, *in* J. L. McClelland and D. E. Rumelhart (eds), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models*, The MIT Press, Cambridge, Mass., pp. 216–271.

Schreuder, R., De Jong, N. H., Krott, A. and Baayen, R. H.: 1999, Rules and rote: beyond the linguistic either-or fallacy, *Behavioral and Brain Sciences* **22**, 1038–1039.

Stojanov, I. P.: 2001, *Connectionist Lexical Processing*, PhD thesis, Rijksuniversiteit Groningen, Groningen, The Netherlands.

Automatic Construction of Morpho-Syntactic Representations

CHAPTER 8

A revised version of this chapter will appear as Fermín Moscoso del Prado Martín and R. Harald Baayen: Unsupervised extraction of high-dimensional lexical representations from corpora using simple recurrent networks, *in* A. Lenci, S. Montemagni and V. Pirrelli (eds.), *The Acquisition and Representation of Word Meaning*, Pisa: Linguistica Computazionale.

Abstract

Existing techniques for vector-space semantic representations have provided useful tools for the automatic building of semantic type systems. However, these models tend to pay little attention to the position of each element in the sequence of words. This leads to the loss of valuable information. On the other hand, Simple Recurrent Neural Networks have been used to capture precisely this word order information. We present Simple Recurrent Networks trained on a word prediction task on medium-sized corpora of Dutch and English, followed by an evaluation of the amount of lexical knowledge acquired by the networks during the course of training. Our results suggest that Neural Networks can be used as a complement to traditional vector-space techniques, avoiding the need to use pre-constructed lexical resources such as parsers and taggers, that might be difficult to obtain in some languages.

Introduction

This study is part of a research project investigating the morphological family size effect, the finding that in language comprehension the speed and accuracy with which a word is recognized is co-determined by the number of words in the mental lexicon in which that word occurs as a constituent (Schreuder & Baayen, 1997). Interestingly, which members of the morphological family contribute to the facilitating effect of the family size depends on the immediate morphological and syntactic context (De Jong, Schreuder, Baayen, in press), suggesting that the semantic percept of a word is co-determined by both context and by the morphologically related words in the lexicon.

Is the morphological family size effect driven by strictly morphological relations in the mental lexicon, or are general semantic and syntactic relations involved? To answer this question, we need lexical semantic representations that are independent of formal properties of words and that are dynamically adjustable by context. Such representations would be useful for obtaining estimations of word similarity and/or densities of the semantic space around a word, both in behavioral experiments and in computational models of lexical processing. Currently, psycholinguistic research involving measures of semantic relations among words is hampered by several facts. First, most experiments addressing this question are limited to studying the effects of the relations among a very limited number of words, that in one way or another must appear explicitly in the experimental materials. Research based on the use of hand-crafted lexical resources, such as *WordNet* (Miller, 1990) is limited to small sets of semantic relations. Third, vector-space semantic techniques such as HAL (Lund & Burgess, 1996) or LSA (Landauer, Foltz, & Laham, 1998), although very powerful, lack the ability to capture the information that is contained in the order in which words appear in the text. In their current state, these methods are not very sensitive to the immediate sentential context of word occurrences.

Vector Space Semantic Representations

There have been many attempts to capture the semantic information that is present in the word co-occurrence statistics of large corpora. These techniques generally collect matrices with the frequencies with which words co-occur in small windows within a large corpus (Lund & Burgess, 1996), or in full documents (Landauer, Foltz, & Laham, 1998; Kanerva, Kristofersson, & Holst, 2000). After the raw frequencies have been collected, dimensional reductions techniques such as Singular Value

Decomposition (SVD) or Principal Component Analysis (PCA) are applied. This results in a collection of high-dimensional vectors representing the meanings of words, and a distance structure that attempts to capture the relationships between these vectors.

Results from this approach have been extremely fruitful in providing high quality representations of word similarity that are both extensive in lexical coverage and cheap to build from large untagged corpora. These techniques have been successfully applied to a wide variety of domains such as: Information Retrieval (Schütze, 1994, Foltz & Dumais, 1992), modelling of priming experiments (Lund, Burgess, & Atchley, 1995; Livesay & Burgess, 1997), automatic marking of student papers (Foltz, Laham, & Landauer, 1999) or inference of morphological relationships (Schone & Jurafsky, 2001). The techniques have also been evaluated using a range of different tests such as synonym identification parts of the Test of English as a Foreign Language (Landauer & Dumais, 1997). Recent experimental work has also shown that subjects are very sensitive to word co-occurrence patterns when presented with novel words, and that their similarity judgments can be manipulated by altering the contextual distributions in which the words appear. Importantly, this work has also shown that vector space semantic representation techniques successfully model these differences (McDonald & Ramscar, 2001).

These systems treat texts as if they were plain bags of words, disregarding the importance of the sequential order in which words appear. Methods as HAL (Lund, Burgess, & Atchley, 1995) take word positions into account by weighting the contribution of co-occurrence frequency with the distance at which the words occur within the sliding window. However, as these weights are added, this will only modify the final frequency value, that is, the co-occurrence of two words that co-occur seldom at a close distance would be the same as that of words that co-occur often at a far distance. Landauer, Laham, Rehder, and Schreiner (1997) argue that, although word order information is one of the cues used by people to understand texts, its importance is small when compared to the information contained by the words themselves, and that little information is lost when they are considered as “in a bag” instead of “as in a sequence”. Although this might be the case for getting the meaning of the majority of texts or sentences, the problem nevertheless arises as we consider the meaning of individual words. For individual words, the syntactical structures in which they appear co-determine their semantics.

In what follows, we explore to what extent Simple Recurrent Networks can be used to obtain high quality semantic representations.

Simple Recurrent Networks and Lexical Representations

Simple Recurrent Neural Networks (SRN; Jordan, 1986; Elman, 1990) are a class of multilayer back-propagation artificial neural networks which include a single recurrent loop in the 'hidden' units. An additional context layer is added to a traditional three-layered neural network. These units are an exact copy of the activation of the "hidden" layer in the previous time tick. Their outputs are connected again to the "hidden" layer as if they were normal input units. Networks of this type implicitly code temporal information.

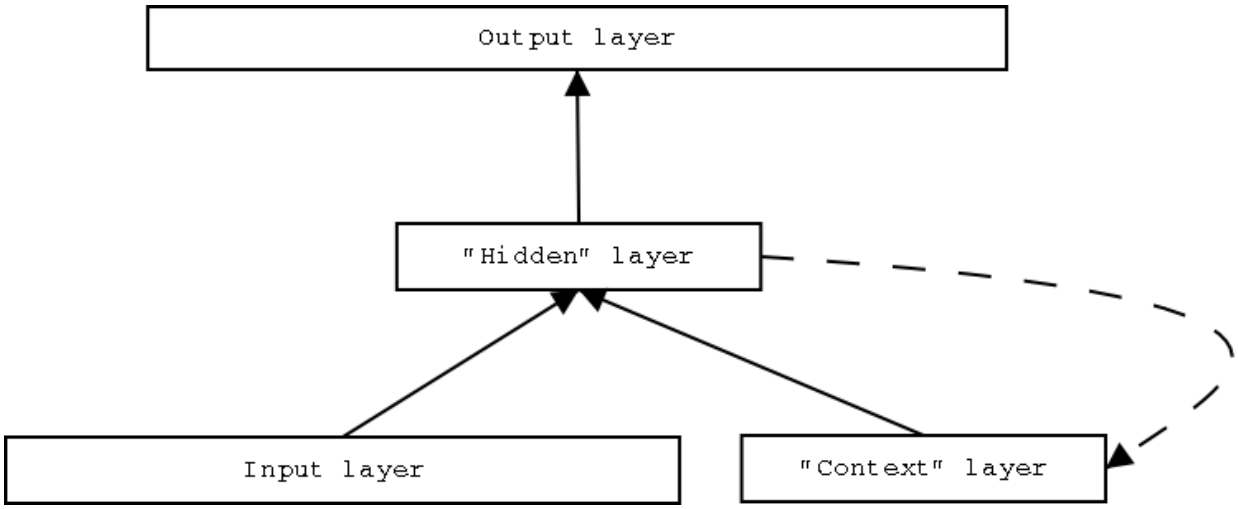

Figure 8.1: Modular architecture of a Simple Recurrent Network. The boxes correspond to layers of units. The solid arrows represent sets of trainable 'all-to-all' connections between the units in two layers. The dashed arrow stands a for fixed 'one-to-one' not trainable connection between two layers. These connections have the function of copying the activation of the hidden units into the context units at every time step.

Figure 8.1 shows the basic architecture of a Simple Recurrent Network. It consists of the traditional input, 'hidden', and output layers plus an additional 'context' layer. At every time step, the units in the context layer act as additional inputs, presenting the activation of the hidden units at the previous moment in time. In this way the network is able to keep a memory of the previous sequence of inputs that it has received. These networks are trained in sequence prediction: from a given sequence of inputs they are trained to predict a sequence of outputs. Those networks in which the input and output sequences are the same are trained to predict the next element in a sequence of items (Elman, 1990). .

Elman (1990; 1993) and Rohde and Plaut (1999) trained SRN's on predicting the next word in a sequence. They presented sentences generated from an artificial

grammar to their networks, one word at a time. At every time step, the networks were trained to predict the word that should follow in the sentence. Even for the very simple artificial grammars that they used, the task of word prediction is non-deterministic, networks can never achieve complete success in predicting the next word. Nevertheless, although in most cases it is not possible to be sure which word is to come next, SRN's learn which words are likely to follow, and which ones are not.

To explore the kinds of knowledge that are acquired by the SRN's in the word prediction task, the activation of the hidden units is recorded following presentation of the input words. Cluster analyses on these activation vectors show that SRN's develop excellent representations of the grammatical features of each of the words. Moreover, these representations are systematically different for each occurrence of the same word, reflecting the grammatical context in which the word token appeared. The representations acquired by the SRN's reported by Elman (1990; 1993) are purely syntactic. Rohde and Plaut (1999) included explicit distributional information in their artificial grammars used for sentence generation. Their intention was to simulate the distributional patterns that are present in realistic language. Importantly, they argue that this distributional information is strongly correlated with the semantic properties of words. From their reasoning, it follows that if networks can learn these patterns from artifical corpora, then similar networks should be able to capture information on the meaning of words in real texts. Note that SRN's are predicting the following word in the sequence from the words that precede it. SRN's share with HAL that information is calculated on the basis of the words in a sliding window. The main differences between these two approaches are that in the SRN the windows have a variable size which is dynamically adjusted in each sentence, that the windows are not centered on the term being represented, but on their rightmost element, and, most importantly, that the windows are not taken as 'bags' of words but rather as real sequences. The present paper extends the previous work with artificial grammars to more realistic corpora, and examines whether the representations formed in the hidden layer capture lexical semantics.

In the following section we will begin by replicating the results of Elman (1993) using both fully localistic input/output representations, and semi-distributed Error Correcting Output Codes. We will evaluate the prediction and representation performance of the SRN's, as compared to other kinds of models trained on the same task. In section 3 we present the results of extending our technique to realistic corpora of English and Dutch, and we provide an evaluation of the kinds of knowledge

that have been acquired.

Simulations on the Artificial Corpus

In this section, we begin by describing a small artificial corpus, we then describe the SRN that we used. Finally, we discuss the predictive and representational abilities of the SRN on this corpus. This is done by comparing its performance with two other techniques: an MBL model and traditional n-gram language models.

Data

Our artificial corpus is the one used by Elman (1993). This corpus consists of $10,000$ sentences of different lengths generated by an artificial grammar that mirrors a number of properties of natural language. The vocabulary consists of 23 types of different grammatical categories.

Both the verbs and the common nouns have a number of features that we want to capture. As the grammar generating the corpus did not implement any additional semantic or distributional information, we will use the features Number and Grammatical Category to evaluate the quality of the representations generated by the models. Words were represented in a completely localistic fashion, without any information about their phonological forms.

We divided the corpus in two parts, a test set of 1000 (10%) sentences, and a training set of $9,000$ (90%) sentences.

Description of the SRN Models

In his original experiments, Elman (1990; 1993) suggested that for a simple recurrent neural network to succeed in the task of predicting the next word in a sequence, its training regime should be manipulated to limit the amount of memory at the initial stages of training. However, Rohde and Plaut (1999) show that this is not necessary, and that an SRN can succeed in learning the task without any initial memory limitation.

Figure 8.2 shows the basic architecture of the neural networks used in this study. We used 80 hidden units and 80 context units. For the input and output layer, we used two different approaches.

1. The first approach made use of a localistic representation with one bit per

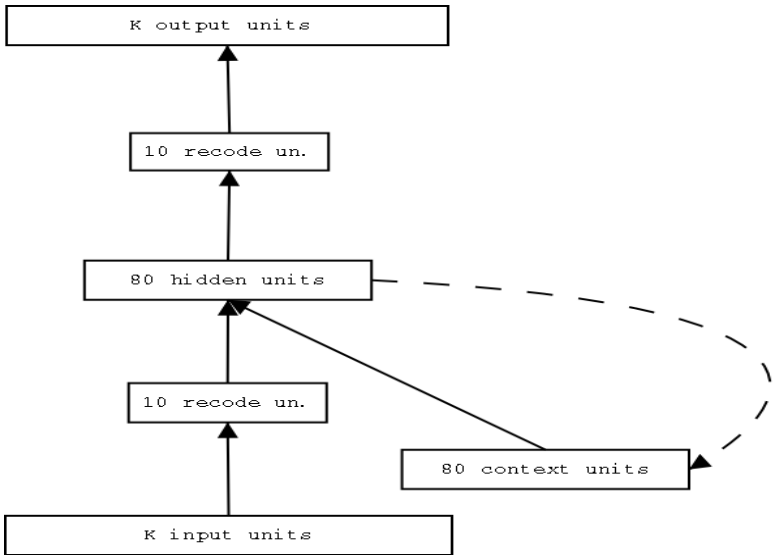

Figure 8.2: General architecture of the simple recurrent networks in this study. K is the size of the input/output representation, 24 for the localistic simulations and 15 for the ECOC codes.

word. One additional bit was added to represent the end of sentence, totaling up to 24 bits for the input and output layers of the network.

2. The second approach used Error Correcting Output Codes (ECOC; Dietterich and Bakiri, 1995) to encode both the input and the output of the network. ECOC codes, although formally distributed, are equivalent to traditional localistic representations in that the dissimilarities between any two codes follow a probabilistic distribution which does not include any information about word similarity. They present the advantage of facilitating learning while reducing the number of necessary bits. We have used a pre-designed ECOC representation consisting of 15 bits to encode 24 items.

Both networks were trained for 50 epochs using the modification of the momentum descent algorithm proposed in Rohde and Plaut (1999). We used a learning rate of 0.04, a momentum of 0.9 and weights were randomly initialized in the $[-1, 1]$ range. The simulations were run using the *Light Efficient Network Simulator* (Rohde, 1999). We found that after 10 epochs of training, the performance in word prediction stalled. However, we noticed important improvements in the representation quality so we decided to continue training up to 50 epochs. Above this level, performance did not show any significant improvement, neither in prediction nor in representation. .

Prediction Performance

To provide base-lines against which to evaluate the prediction performance of our SRN models, we compare them with the prediction performance of equivalent models built using a Memory Based Learning (MBL) Technique, and with more traditional n-gram Language Models.

Memory Based Learning (MBL), also referred to as Case Based Reasoning, is a family of machine learning techniques which rely on the raw storage of all available data without any abstraction procedure. Each time these systems are presented with a new instance of a problem to solve, they search through their instance base of previously solved cases, and they output the same answer that was given in the most similar (according to different similarity metrics) of the precious cases. The MBL models were built using TiMBL (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 1999). In general, there are two basic parameters that have to be considered in this technique: the number of nearest neighbors used for the decision (k), and the similarity metric used to determine the nearest neighbor set.

In our experiments we used a sliding window technique: The models have to predict the next word on the basis of the w preceding words. We built 9 models with window sizes ranging from 1 to 9 words. The words were represented by plain text labels. Examples were compared using the Modified Value Difference Metric (MVDM; Stanfill & Waltz, 1986; Cost & Salzberg, 1993) combined with Information Gain Ratio Feature Weighting (IGR; Quinlan, 1993) , the features being the words in the sliding window. The models considered the 11 nearest neighbors to make the classifications.

Our artificial corpus is composed of sentences that were randomly generated using a simple grammar, therefore there could not be any relevant information beyond the sentence boundary. For this reason, we filled all slots of the text window that exceeded the sentence domain with a $[NULL]$ token. Another token was added to represent the end of the sentence.

Per-word cross-entropy was calculated over the probability distributions of the words predicted by the system. The per-word cross-entropy was calculated according to the equation:

$$H(L,m) \simeq -\frac{1}{n}\sum_{j=1}^{n}\log_2 m(w_j|w_1,w_2,\ldots,w_{j-1}) \tag{8.1}$$

with $H(L,m)$ the cross-entropy between language L and language model m, and $m(w_j|w_1,\, w_2,\ldots,w_{j-1})$ the conditional probability in model m that word w_j follows,

given that the previous sequence is w_1, w_2, w_{j-1}. To calculate the cross-entropy, the test corpus was divided into 10 parts. Cross entropy was calculated for each of the 10 sub-corpora and then averaged.

We also built traditional n-gram language models using the *CMU-Cambridge Statistical Modeling Toolkit v2* (Clarkson & Rosenfeld, 1997). We built models with n values ranging from 2 to 5. Note that a bigram model is approximately equivalent to an MBL model with a window size of two. All models were trained on the same training corpus. For testing, the test set was divided in ten subsets, each containing 100 of the original 1000 sentences. The cross-entropy was calculated using Equation 8.1 across the 10 testing corpora and then averaged.

To evaluate the performance of the neural network using the 24-bit localistic representation, we took the activation of the bits representing each word (after normalization) to be a measure of the probability with which that word was expected to appear. With these probability estimates we applied Equation 8.1 to obtain the cross entropy for each of the 10 testing sets, and then averaged the result.

The calculation of the entropy for the neural network using the 15-bit ECOC was not as straightforward. Due to the distributed nature of the output, it is difficult to interpret it in terms of a probability distribution. Each bit in the output represents the summed probabilities of several words that share that active bit. To solve this problem, we estimated the cross-entropy.

We started by assuming that the activation of each bit of the representation is independent from the activation of all other bits. Although this is an obvious simplification, it provides us with a lower bound to the actual probabilities. This first estimate was then corrected by taking into account the number of bits used and the number of words being encoded, so that our final estimation is that shown in Equation 8.2 . In the equation B stands for the number of bits in the representation, N stands for the number of types coded, n is the number of tokens in the corpus, C is the previous context and b_i is the activation value of the i-th bit of the output. To make sure that the estimates provided by Equation 8.2 do not significantly differ from those obtained by Equation 8.1 , we performed a paired two-tailed t-test comparing the estimates provided by Equations 8.1 and 8.2 on the results of the MBL systems. For window sizes greater than two, our estimation was not statistically distinguishable ($t = 1.4099$,$df = 79$,$p = 0.1625$) from that obtained by the traditional method.

$$\widehat{H}(L,m) = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{B}\log_B(P(b_i|C)) - \frac{1}{\log_2 N} \tag{8.2}$$

Figure 8.3 compares the average cross-entropy of the n-gram models, the MBL models and the SRN using ECOC. Note that, as the SRN models do not use an explicit window size, its results are presented as a line extending across all window sizes. The results from the localistic SRN do not differ significantly from those of the ECOC, thus we do not include them in the graph for simplicity reasons. For greater window sizes, the MBL systems clearly outperform the n-gram models. The performance of both neural networks on word prediction is somewhere in between the prediction power of a bigram model and that of a trigram model.

Representation Performance

Following training, we extracted representations from the models. In the case of the TiMBL systems, the representations we used were based on the MVDM matrices. We averaged these matrices across features using the IGR weights, and then applied a Principal Component Analysis to the columns of the resulting matrix. Four principal components accounted for approximately 92% of the variance. We used the loadings of the words on these principal components as their vector representations.

In the case of the neural networks, we followed (Elman, 1990; 1993) and used the average activation of the hidden units of the network when a word was presented as the initial representation. We also applied Principal Component Analysis to these initial representations. We now found 4 principal components to account for 89% of the variation.

Hierarchical clustering was performed on each of these four representations. The results are shown in Figure 8.4. The general grammatical classification is almost perfect in all cases. However, the biggest TiMBL system, with a window size of 10, achieves a better classification in terms of number. It is also remarkable that TiMBL systems seem to outperform the networks with respect to number classification of common nouns. In general, we can say that the classification achieved by the case-based system with a window of 10 is perfect in most aspects, with the exceptions of proper nouns and transitive verbs. On the other hand, the network models also achieve excellent performance on the classification of grammatical categories, especially with respect to the upper branches of the tree which correctly group nouns and verbs. But the performance of the networks is less good for the more detailed Number classification.

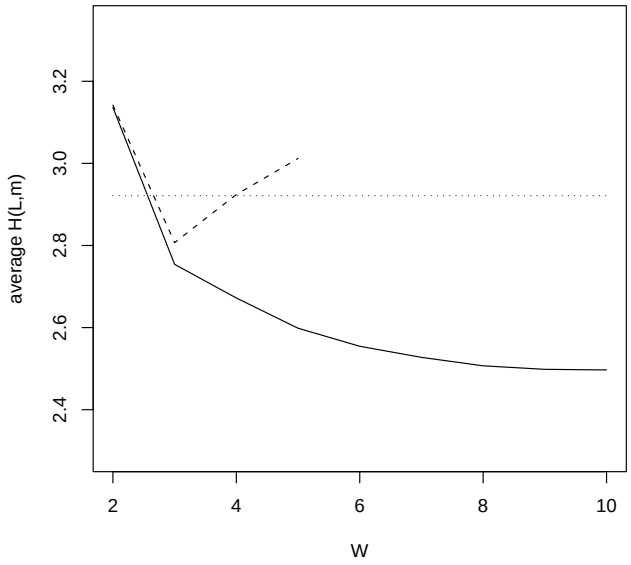

Figure 8.3: Comparison of the average cross-entropies produced by case-base reasoning systems, n-gram models and the SRN's. The solid line represents the performance of the MBL systems, the dashed line represents the n-gram models and the dotted line represents the SRN using ECOC. The horizontal axis represents the size of the sliding window, and the vertical axis represents the average cross-entropy value.

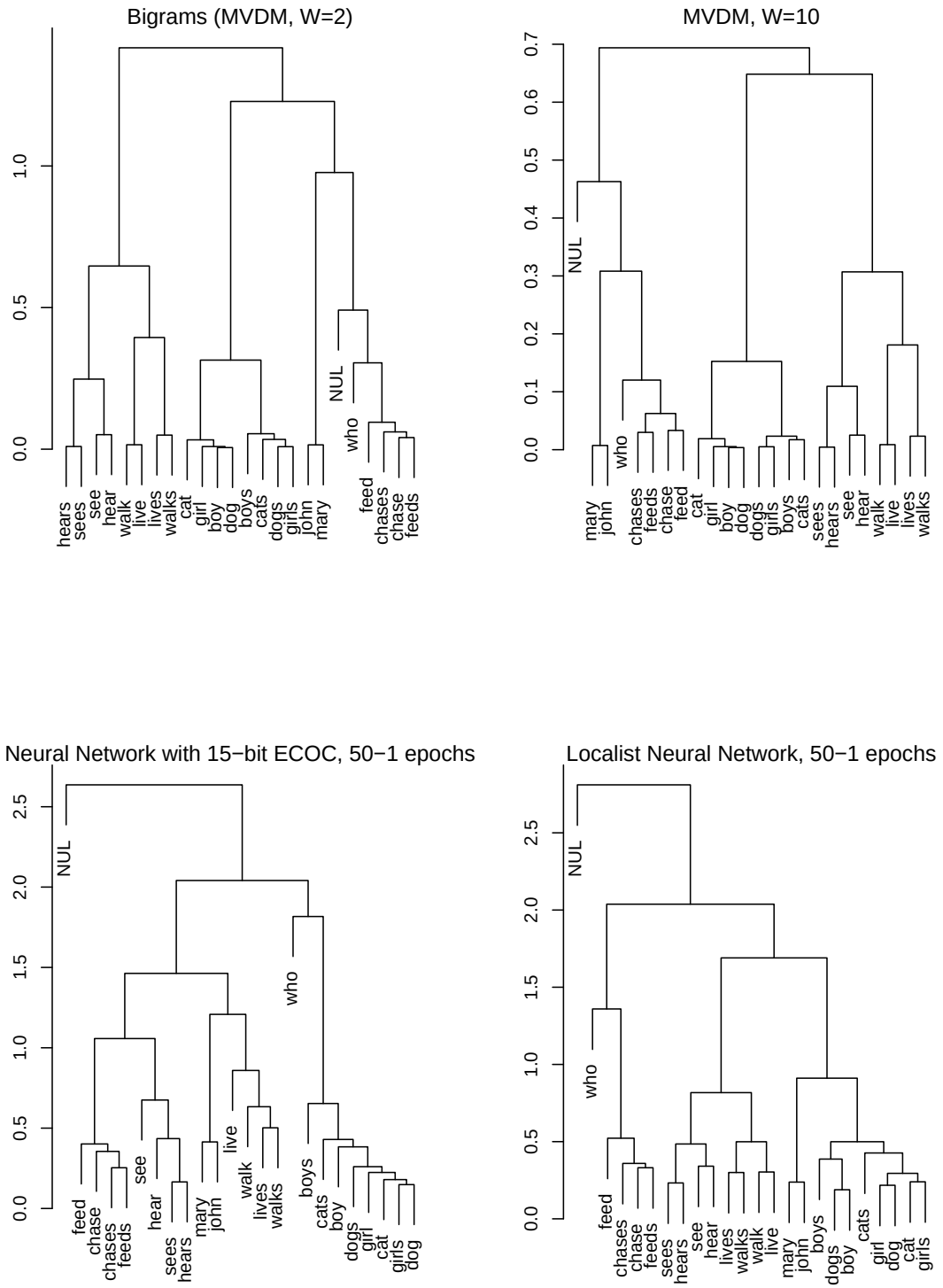

Figure 8.4: Clustering results from MBL systems with window sizes 2 and 10, and from the neural networks with and without the use of ECOC codes.

Using More Realistic Corpora

In the previous section, we showed that the MBL systems, when provided a sufficiently large window size, outperform SRN models both in prediction and representation performance. Unfortunately, this type of model is very difficult to scale up to realistic vocabulary sizes. It is also unclear how MBL models with limited window sizes deal with the information contained in the extra-sentential context of real texts. For these reasons we will continue with our exploration of the SRN models. We therefore applied the SRN technique to the following data sets:

- A collection of texts ($748,486$ tokens and $13,021$ different types), consisting of a collection of Jane Austen's novels extracted from *The Project Gutenberg*. Specifically, our corpus contained the novels: *Emma*, *Mansfield Park*, *Northanger Abbey*, *Pride and Prejudice*, and *Sense and Sensibility*.
- A corpus of Dutch newspaper articles, containing around $4,500,000$ tokens, divided into approximately $160,000$ types.

Some simple preprocessing was performed on the original ASCII files: they were all changed to lower-case, digits were substituted for a $[DIGIT]$ token, spaces between common punctuation marks were inserted, an end-of-paragraph token was added and all other non-alphabetical characters except common punctuation marks were removed. The texts were then divided into one-paragraph long 'examples' used in the training of the networks.

Network Design and Training

The networks employed were adaptations of that shown in Figure 8.2. The input and output layers consisted of 200 units each for the Austen Corpus, and 300 units in the case of the Dutch corpus. The hidden and context layers were kept the same size as for the small artificial corpus, while the number of recoding units was extended to 20.

Although in our initial experiments with the small corpus we found that the use of ECOC codes drastically reduced the training time of the networks, our experience with the large corpus showed that a semi-localistic initial representation converged faster. We trained several models, differing in the nature of their input representations and in the pattern of reset applied to the context units. The best performance was obtained using a semi-localistic, paragraph driven model.

A localistic representation with one-bit-per-word codings would require at least $13,021$ units to encode the vocabulary of the Austen Corpus. We therefore used a 200 bit coding with two active bits per word. In this way we are able to represent a maximum of $\binom{200}{2} = 19,900$ different types and we keep the input vectors very close to being orthogonal. To build this representation, we assigned two random active bits to each type in our vocabulary. This random selection of bits assures that the similarity introduced by having two, instead of one, bits per word behaves as if it were random noise, which actually improves learning in neural networks. It also guarantees that, in the worst case, any given word can have up to 396 other words that are not completely orthogonal to it (that is, they share 1 common bit). This is equivalent to having a maximum of approximately 2% random noise. In the Dutch corpus, we used 3 out of 300 active bits, which allows us to represent a maximum of $\binom{300}{3} = 4,455,100$ types.

The network was trained on word prediction for 20 epochs by modified momentum descent (Rohde & Plaut, 1999) on the Austen corpus. In the Dutch case, due to the larger size of the corpus, 3 epochs were sufficient to obtain our representations. We used a learning rate of 0.1, a momentum of 0.9 and a batch size of 1 paragraph. The order of presentation of the examples was randomized. Context units were reset to 0.5 at the end of every paragraph. In both cases, the networks were trained for an extra epoch using a learning rate of 0.01 in order to fine-tune the representations.

Overview of the Representations

For each corpus, we selected all the types that appeared with a frequency equal or greater than 10. The activation of the networks' hidden and recode units were recorded during test presentations of the full corpora after training. The vectors obtained for each token were averaged across the types. In this way we obtained, for each corpus, two different vectorial representations of the most frequent words in it, one corresponding to the hidden units and the other corresponding to the second recode layer.

We compared the results from both representations. The information of the recoding units turned out to be slightly superior. Henceforth we use these as our final representations.

As an overview of the representations obtained by our network, we present a list of words belonging to different grammatical categories, together with their nearest

neighbors in the representation using the *City-block* distance metric:

$$d(\overline{w}, \overline{v}) = \sum_i |w_i - v_i|. \tag{8.3}$$

Table 8.1: Nearest neighbors for several English words as obtained by the system from the Austen corpus. Words that share meaning and grammatical category with the target are marked with an asterisk. Words that share the grammatical features with the target, but do not appear to share any semantic features are marked with a plus symbol.

Word	**Nearest Neighbors**				
	1st	2nd	3rd	4th	5th
must	may*	should*	shall*	will*	might*
was	were*	are*	dispersed+	grandmama	is*
said	shows*	replied*	continued*	persuaded*	novels
say	sold+	imagine*	know*	directing	succeed+
in	beyond*	by*	of*	from*	within*
she	he*	who*	it*	this*	they*
his	your*	their*	her*	inconveniences	our*
small	public+	perfect+	tender*	unkindness	comforts
lucy	marianne*	elinor*	julia*	edward*	harriet*
mother	father*	harris*	goddard*	crawford*	weston*
house	case+	night+	system+	servant+	ball+

Tables 8.1 and 8.2 show some examples of the results obtained. The network tends to group nouns and adjectives together. It does not seem to be capturing much meaning in these categories. We believe that this indicates that the network is capturing mostly syntactic information and, in those cases where semantics can be inferred directly from syntactic patterns, some meaning as well. Interestingly, for nouns or adjectives with clear argument structure or with well-defined roles in the sentence, as in family terms and proper nouns, the performance is also good in terms of semantics. A good example of this is shown by the nearest neighbors of “Lucy” in the Austen corpus, which are feminine proper names. This is probably to the fact that, especially in a Jane Austen novel, a female name is likely to appear together with very specific kinds of verbs. A similar effect was found for place names in the Dutch newspapers. Nevertheless, although heavily guided by syntax, the network is also capturing some co-occurrence semantics. This can be seen in the nearest neighbors of “vrouw” (*woman*) and “wet”(*law*) in Dutch, which respectively include “kinderen”(*children*) and “vrouwen”(*women*), and “rechter”(*judge*), “ethische”(*ethical*), “buitenland”(*abroad*) and “belangstelling”(*“interest”*).

Table 8.2: Nearest neighbors for several Dutch words as obtained by the system from the Dutch corpus. Words that share meaning and grammatical category with the target are marked with an asterisk. Words that share the grammatical features with the target, but do not appear to share any semantic features are marked with a plus symbol.

Word	**Nearest Neighbors**				
	1st	2nd	3rd	4th	5th
moet	zou*	kan*	zouden*	zal*	wat
moeten	zouden*	kunnen*	mogen*	zullen*	zou*
zei	zegt*	vindt*	doe+	meent*	terwijl
zeggen	vinden*	bestaat	zien*	gezegd*	weten*
*in**	tussen*	voor*	naar*	tijdens*	bij*
ze	zij*	je*	daar+	nu+	we*
duitse	alle+	belgische*	politieke*	oude+	hogere+
duits	dodelijke	durft+	waarschijnlijk+	goed+	vermoedelijk+
vrouw	vrouwen*	kinderen*	cda+	namen+	woorden+
wet	rechter*	ethische*	buitenland*	belangstelling*	mens*
groep	partij*	dag+	kans+	verleden+	officiele*

Another interesting observation concerns the Dutch adjectives “Duits” and “Duitse” which are both forms of the adjective *German*, but in different inflectional contexts. Although the network hasn’t achieved much success in clustering semantically related adjectives, it tends to group forms with “-e” and unsuffixed forms together.

The performance of the model is excellent for those words that carry mainly grammatical information. The nearest neighbors of the Dutch verb “moeten” (*must*) were other modal verbs of similar meaning (“zouden” – *should*, “kunnen” – *can*, “mogen” – *may*, “zullen” – *will*) and always in the same inflectional variant, that is singular forms in the case of “moet” and plural/infinitive forms in the case of “moeten”. Similarly, neighbors of the preposition “in”, were all prepositions, and the pronoun “ze” (*she/they*) selects other pronouns, preferably those that are also 3rd person singular nominative. In this case it is remarkable that the nearest neighbor is “zij” which is an non-reduced version of the same pronoun.

We also checked whether the network made use of long distance relationships to capture its representations, or whether it was just taking into account very local contexts only. To do this, a modification in the training regime was introduced: every 6 time ticks the context units were reset to 0.5. In this way we restricted the networks’ memory to only the 6 previous time ticks. We found that networks trained in this way failed to capture as much information as those that had an unrestricted memory extending to a full paragraph. This is an indication that, in this task, small window approaches cannot capture all the information required.

Results for Dutch are much better than those obtained for English. This is probably due to the much bigger and varied corpus used for Dutch.

Evaluation of the Syntactic Knowledge Acquired by the SRN from the Dutch Corpus

In the previous subsection, we pointed out that our SRN's capture mainly syntactic information, although some semantic information is also captured when it can be inferred directly from the syntax. Now we provide a more precise statistical evaluation of the nature and amount of syntactic knowledge extracted. We will center only on the Dutch corpus, given that it was the largest one used in these experiments.

To evaluate the quality of the grammatical information contained in our vectors, we extracted the values of their grammatical features from the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995). We extracted from CELEX the data corresponding to the grammatical category of the words and information about inflectional details such as: gender in adjectives, number in nouns, and person / number / tense in verbs. In each of these cases, we checked how well these labels could be predicted from the information in the vectors. To provide a base-line against which to compare the prediction performance, we calculated how well this prediction could be performed using Maximum Likelihood Estimation (MLE) based solely on the frequencies of each value. Note that an MLE-based decision on this task will always predict the most frequent value for each feature (e.g. 'noun' in the case of grammatical category). This results in perfect prediction performance for the most frequent value, and null prediction power for the rest of the possible values of the feature. If the values of the features can be predicted from the vectors with a performance above that obtained by MLE, then it is clear that the vectors contain information about those features. Moreover, we shall see that these predictions, unlike MLE-based ones, should also be valid for the least frequent feature values.

To obtain predictions from our vectors, we turn again to Memory Based Learning (MBL) techniques using TiMBL. In each analysis, we provide TiMBL with the values in the vectors and train it to predict the values of the corresponding features. If the syntactic information is encoded in the numeric values in our vectors, an MBL system should benefit from that structure in order to make the predictions. The performance of MBL in predicting the value of a feature from the values of the elements in the vectors is a lower bound of the amount of information encoded. Given that the representations that we are dealing with are averaged across all tokens of a

particular type, we will not consider those words which can be ambiguous with respect to a particular feature (e.g., a word like “press”, which could be either a noun or a verb, is not considered in the evaluation of grammatical category). All the results that we report are based on ten-fold cross-validations, i.e., training using 90% of the data and testing on the remaining 10%, repeated ten times so that every item is used for training in nine of the tests and for testing in the remaining one. All systems we trained and tested using TiMBL with Information Gain Feature Weighting and considering the 7 nearest neighbors according to City-Block distance metric. To test the statistical significance of differences between the performance obtained from the vectors and that obtained from MLE, we report one-tailed t-tests on this differences (in each of the ten trials of the cross-validation). We use one-tailed tests because we are testing whether the performance is significantly better when predicting from the vectors, not just different.

Grammatical Category Information

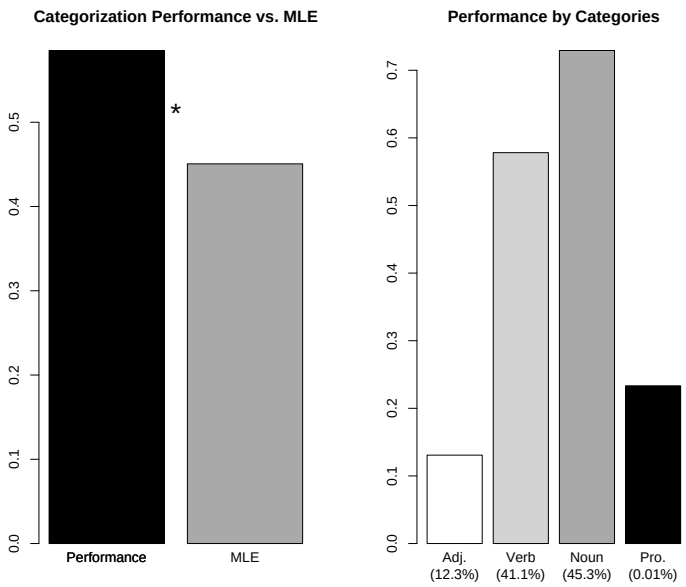

Figure 8.5: Performance in classification of vectors corresponding to unambiguous words with respect to their grammatical category.

Figure 8.5 shows the categorization performance of a TiMBL system trained on predicting the grammatical category of words from the vectors. The left part of the graph compares the performance obtained by the classifier with the performance

that would be achieved using MLE. According to a one-tailed t-test the results obtained in the classification are significantly better than those obtained by MLE ($t = 22.6387$,$df = 9$,$p = 0.0000$). The right part of the figure compares the classification performance for the most common categories. The percentages of words with each of the labels can be found on the horizontal axis. The effect of having multiple categories to choose from is very negative for the MLE method, which would predict all words to be nouns. Moreover, note that even for the extremely infrequent pronouns, it achieves a classification performance above 20%, although they only represent 0.01% of the words.

Inflectional Details: Person, number, gender and tense in adjectives, nouns and verbs

To evaluate the amount of more subtle inflectional information contained in our vectors, we applied the same technique to three different classification tasks:

Adjectival Inflection: We extracted all unambiguous adjectives from the CELEX database, and added a label to each vector indicating whether it had the inflectional '-e' (neuter) suffix.

Nominal Inflection: We extracted all unambiguous nouns from CELEX and added a label indicating the number (singular or plural) of these nouns.

Verbal Inflection: We extracted all unambiguous verbs from CELEX and added a label indicating whether the verb was a first person singular, second-third person singular, a plural or a past participle.

Figure 8.6 shows the categorization performance of a TiMBL system trained on predicting the gender inflection of adjectives from the vectors. The left part of the graph compares the performance obtained by the classifier with the performance of the MLE. According to a one-tailed t-test, the results obtained in the classification are significantly better than those obtained by MLE ($t = 4.011$,$df = 9$,$p = 0.0015$). The right part of the figure compares the classification performance for the adjectives with and without the inflectional '-e'. The percentages of words with each of the labels can be found on the horizontal axis. Interestingly, the classification performance is high also for the least frequent case, for which the MLE strategy has a performance of zero.

Figure 8.7 shows the categorization performance of a TiMBL system trained on predicting the number inflection of nouns from the vectors. The left part of the

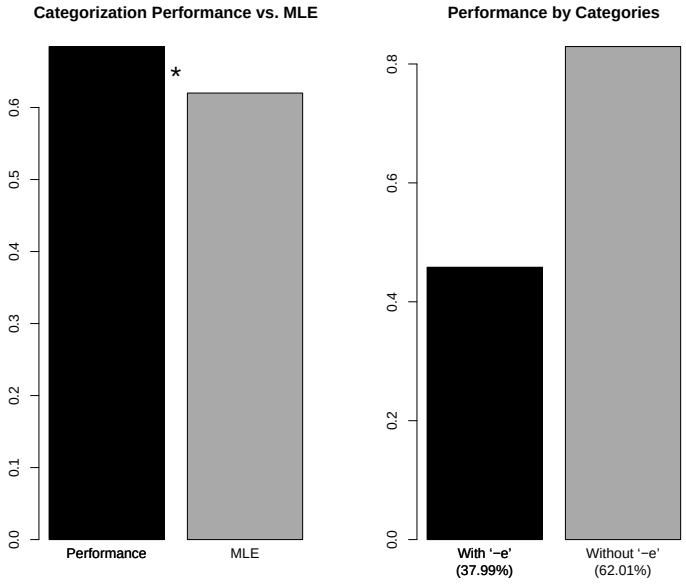

Figure 8.6: Performance in classification of vectors corresponding to adjectives with respect with the presence/absence of the neuter suffix '-e'.

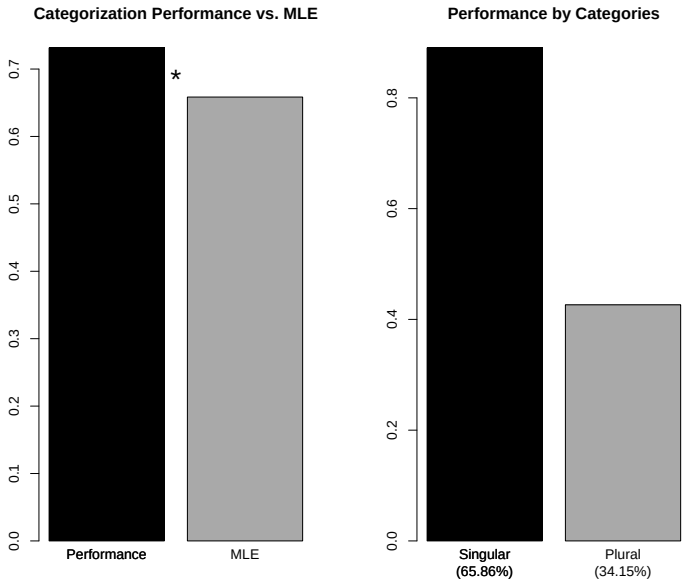

Figure 8.7: Performance in classification of vectors corresponding to nouns with respect to their number.

graph compares the performance obtained by the classifier with that that would be achieved using MLE. According to a one-tailed t-test the results obtained in the classification are significantly better than those obtained by MLE ($t = 11.4373$,$df = 9$,$p = 0.0000$). The right part of the figure compares the classification performance for both cases. The percentages of words with each of the labels can be found on the horizontal axis. Again we find a performance above 40% even for the minority case of plurals.

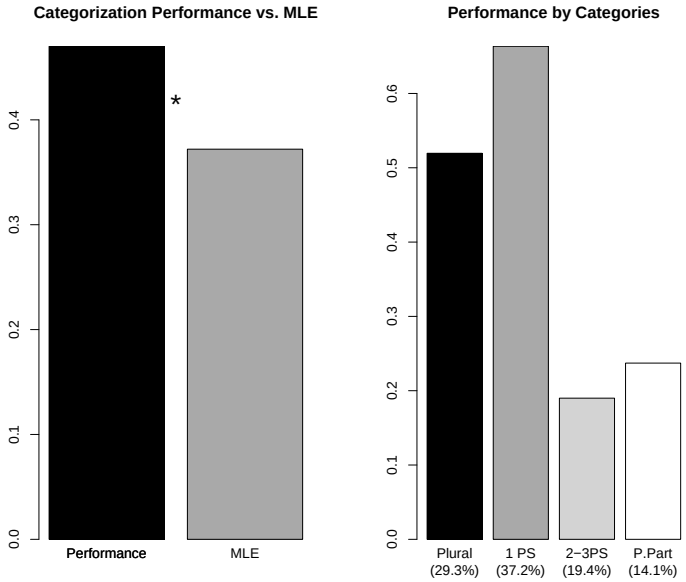

Figure 8.8: Performance in classification of vectors corresponding to verbs with respect to their person,number and tense inflection.

Figure 8.8 shows the categorization performance of a TiMBL system trained on predicting the person, number and tense inflection of verbs from the vectors. The left part of the graph compares the performance obtained by the classifier with what can be achieved using MLE. According to a one-tailed t-test, the results obtained in the classification are significantly better than those obtained by MLE ($t = 15.6836$,$df = 9$,$p = 0.0000$). The right part of the figure compares the classification performance for the four kinds of inflected verbs. Again, vectors contain sufficient information to achieve some success in predicting the minority forms as past participles and second-third plural forms.

General Discussion

Neural networks and MBL systems provide promising tools for word prediction comparable in power to standard n-gram language models. Interestingly, Figure 8.3 shows that n-gram models achieve their best prediction performance with $n = 3$, but that the TiMBL systems continue to improve their performance as the window-size is increased, without ever getting worse.

MBL systems, when given a sufficiently large window size, outperform neural networks with respect to the more fine-grained aspects of the representations they build. However, SRN's have the advantage of being easier to scale up to real-life corpora, because they do not require to store the full example sets, while still achieving a fairly good performance. Neural Networks provide natural, compact and manageable representations of around 20 elements, without the need to deal with huge matrices that then need to be reduced via Principal Component Analysis or Singular Value Decomposition.

The success-percentages in Figures 8.5, 8.6, 8.7 and 8.8 should not be taken as reliable estimators of the amount of knowledge about each feature in our vectors, but rather as very pessimistic lower bounds. They only tell us how much of this informations is salient enough to be easily captured by an MBL system. Performance for the least frequent categories may seem to be very low. However, Tables 8.1 and 8.2 show that this is in fact not the case at all, and precisely these low-frequent, closed categories that generally correspond to function words are those for which the information encoded in the vectors is most clear. However, as they are presented in the training sets together with the other categories, which are two orders of magnitude more frequent, their relevant properties tend to ignored by MBL systems.

The technique we propose in this paper provides a useful tool for extracting morpho-syntactic features from large corpora, requiring very little design and processing time. This is of great convenience for the semi-automatic development of lexical resources, where new text resources for new domains or even languages need to be added. Although the representations obtained by our systems focus on syntax, we have shown that features that are traditionally considered syntactic have strong semantic implications. This was specially evident in the case of verb semantics. We believe that a combination of the technique presented here, with some other traditional co-occurrence based technique such as HAL or LSA, can produce cheap, high quality representations that encode syntactic and semantic information in a distributed way. These representations can also be used to enhance traditional knowledge-based semantic type systems in a semi-automatic way. Another advan-

tage of these kind of models is that they provide a natural mechanism to represent context specific properties of certain tokens, both in syntactic and semantic domains.

There have been many previous studies on the use of neural networks to capture different aspects of syntax and semantics (Cotrell, 1989; Elman, 1993; Jain, 1989; Miikkulainen, 1993; StJohn & McClelland, 1990; Wiener, Pedersen, & Weigend, 1995; Schütze, 1995; Henderson & Lane, 1998; Morris, Cottrell, & Elman, 2000). However, most of these studies impose a great amount of structure on the networks, reflecting various assumptions about the structure of language. The results of our experiments suggest that much of the grammatical knowledge in language *can* be learned from the distributional information that language itself provides. Interestingly, all this information has been learned in a very constrained environment, the systems being deprived of some relevant information, such as morphology. As it is shown in the chapter by Sahlgren in this volume, the inclusion of morphological information enhances the performance of co-occurrence based systems. This is also supported by the well known fact that the morphology of a language conveys a great deal of semantic information about the words. Some authors even consider morphology as a plain interaction between form and meaning (Seidenberg & Gonnerman, 2000). The inclusion of form information for the words, together with word-order information has the potential to enhance the performance of co-occurrence based models. Plain “bag-like” co-occurrence, word order information and word form information, will have to be combined in one single system extracting semantic information simultaneously from all these sources. In addition, the inclusion of word form information is itself probably a prerequisite for languages with extremely rich morphological systems such as Finnish or Hungarian. Without word-form information, systems dealing with these languages will encounter severe problems relating to data sparseness. At the same time, much of the information that is conveyed by word order in languages like English or Dutch is embedded in the morphological structure of these languages. An unsupervised method for capturing semantics or syntax for morphologically rich languages will have to take word-form into account.

References

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX lexical database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Clarkson, P. and Rosenfeld, R.: 1997, Statistical language modeling using the CMU-Cambridge toolkit, *ESCA Eurospeech 1997*, ESCA.

Cost, S. and Salzberg, S.: 1993, A weighted nearest neighbor algorithm for learning with symbolic features, *Machine Learning*.

Cotrell, G. W.: 1989, A connectionist approach to word sense dissambiguation, *Research Notes in Artificial Intelligence*, Morgan Kaufmann, San Mateo.

Daelemans, W., Zavrel, J., Van der Sloot, K. and Van den Bosch, A.: 1999, TiMBL: Tilburg memory based learner reference guide 2.0, *Report 99-01*, Computational Linguistics Tilburg University.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: *to appear*, Morphological resonance in the mental lexicon, *Linguistics*.

Dietterich, T. G. and Bakiri, G.: 1995, Solving multiclass learning problems via error-correcting output codes, *Journal of Artificial Intelligence Research* **2**, 263–286.

Elman, J. L.: 1990, Finding structure in time, *Cognitive Science* **14**, 179–211.

Elman, J. L.: 1993, Learning and development in neural networks: The importance of starting small, *Cognition* **48**, 71–99.

Foltz, P. and Dumais, S.: 1992, Personalized information delivery: An analysis of information filtering methods., *Communications of the ACM* **35**(12), 51–60.

Foltz, P., Laham, D. and Landauer, T.: 1999, Automated essay scoring: Applications to educational technology, *Proceedings of EdMedia'99*.

Henderson, J. and Lane, P.: 1998, A connectionist architecture for learning to parse, *ACL 36 / COLING 17*, pp. 531–537.

Jain, A. N.: 1989, A connectionist architecture for sequential symbolic domains, *Technical report CMU-CS-89-187*, Carnegie Mellon University.

Jordan, M. I.: 1986, Serial order: A parallel distributed approach, *Institute for Cognitive Science Report 8604*, University of California, San Diego.

Kanerva, P., Kristofersson, K. and Holst, A.: 2000, Random indexing of text samples in latent semantic analysis, *Proceedings of the 22nd Annual Conference*

of the Cognitive Science Society, p. 1036.

Landauer, T. and Dumais, S.: 1997, A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge, *Psychological Review* **104**(2), 211–240.

Landauer, T. K., Foltz, P. W. and Laham, D.: 1998, Introduction to latent semantic analysis, *Discourse Processes* **25**, 259–284.

Landauer, T., Laham, D., Rehder, B. and Schreiner, M.: 1997, How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans., *Proceedings of 19th annual meeting of the Cognitive Science Society*, Mawhwah, NJ., pp. 412–417.

Livesay, K. and Burgess, C.: 1997, Mediated priming in high-dimensional meaning space: What is "mediated" in mediated priming?, *Proceedings of the Cognitive Science Society*, Lawrence Erlbaum Associates, Inc., Hillsdale, N.J., pp. 436–441.

Lund, K. and Burgess, C.: 1996, Producing high-dimensional semantic spaces from lexical co-occurrence, *Behaviour Research Methods, Instruments, and Computers* **28**(2), 203–208.

Lund, K., Burgess, C. and Atchley, R. A.: 1995, Semantic and associative priming in high-dimensional semantic space, *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, Erlbaum, Hillsdale, NJ., pp. 660–665.

McDonald, S. and Ramscar, M.: 2001, Testing the distributional hypothesis: The influence of context judgements of semantic similarity, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

Miikkulainen, R.: 1993, *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon and Memory*, MIT Press, Cambridge, MA.

Miller, G. A.: 1990, Wordnet: An on-line lexical database, *International Journal of Lexicography* **3**, 235–312.

Morris, W. C., Cottrell, G. W. and Elman, J. L.: in press, A connectionist simulation of the empirical acquisition of grammatical relations, *in* S. Wermter and R. Sun (eds), *Hybrid Neural Symbolic Integration*, Springer Verlag.

Quinlan, J. R.: 1993, *Programs for Machine Learning*, Morgan Kauffmann, San Mateo, CA.

Rohde, D. L. T.: 1999, LENS: The light, efficient network simulator, *Technical Report CMU-CS-99-164*, Carnegie Mellon University, Pittsburg, PA.

Rohde, D. L. T. and Plaut, D. C.: 1999, Language acquisition in the absence of explicit negative evidence: how important is starting small?, *Cognition* **72**(1), 67–109.

Schone, P. and Jurafsky, D.: 2001, Knowledge free induction of inflectional morphologies, *Proceedings of the North American Chapter of the Association for Computational Linguistics NAACL-2001*.

Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **37**, 118–139.

Schütze, H.: 1994, Towards connectionist lexical semantics, *in* S. D. Lima, R. L. Corrigan and G. K. Iverson (eds), *The Reality of Linguistic Rules*, Vol. 26 of *Studies in Language Companion Series*, John Benjamins Publishing Company, Amsterdam, PA., pp. 171–191.

Schütze, H.: 1995, Distributional part-of-speech tagging, *EACL 7*, pp. 251–258.

Seidenberg, M. S. and Gonnerman, L. M.: 2000, Explaining derivational morphology as the convergence of codes, *Trends in Cognitive Sciences* **4**(9), 353–361.

Stanfill, C. and Waltz, D.: 1986, Towards memory-based reasoning, *Communications of the ACM* **29**(12), 1213–1228.

StJohn, M. F. and McClelland, J. L.: 1990, Learning and applying contextual constraints in sentence comprehension, *Artificial Intelligence* **46**, 217–457.

Wiener, E., Pedersen, J. and Weigend, A.: 1995, A neural network approach to topic spotting, *Proceedings of SDAIR 95*, Las Vegas, NV., pp. 317–332.

Automatic Construction of Semantic Representations

CHAPTER 9

This chapter has been published as Fermín Moscoso del Prado Martín and Magnus Sahlgren (2002): An integration of Vector-Based Semantic Analysis and Simple Recurrent Networks for the automatic acquisition of lexical representations from unlabeled corpora, *in* A. Lenci, S. Montemagni and V. Pirrelli (eds.), *Proceedings of the LREC'2002 Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*, Paris:ELRA

Abstract

We present an integration of Simple Recurrent Networks to extract grammatical knowledge and Vector-Based Semantic Analysis to acquire semantic information from large corpora. Starting from a large, untagged sample of English text, we use Simple Recurrent Networks to extract morpho-syntactic vectors in an unsupervised way. These vectors are then used in place of random vectors to perform Vector-Space Semantic Analysis. In this way, we obtain rich lexical representations in the form of high-dimensional vectors that integrate morpho-syntactic and semantic information about words. We argue how these vectors can be used to account for the particularities of each different word token of a given word type. The amount of lexical knowledge acquired by the technique is evaluated both by statistical analyses comparing the information contained in the vectors with existing 'hand-crafted' lexical resources such as CELEX and WordNet, and by performance in language proficiency tests. We conclude by outlining the cognitive implications of this model and its potential use in the bootstrapping of lexical resources

Introduction

Collecting word-use statistics from large text corpora has proven to be a viable method for automatically acquiring knowledge about the structural properties of language. The perhaps most well-known example is the work of George Zipf, who, in his famous *Zipf's laws* (Zipf, 1949), demonstrated that there exist fundamental statistical regularities in language. Although the useability of statistics for extracting structural information has been widely recognized, there has been, and still is, much scepticism regarding the possibility of extracting semantic information from word-use statistics. We believe that part of the reason for this scepticism is the conception of meaning as something external to language — as something *out there* in the world, or as something *in here* in the mind of a language user. However, if we instead adopt what we may call a "Wittgensteinian" perspective, in which we do not demand any rigid definitions of word meanings, but rather characterize them in terms of their use and their "family resemblance" (Wittgenstein, 1953), we may argue that word-use statistics provide us with exactly the right kind of data to facilitate semantic knowledge acquisition. The idea, first explicitly stated in Harris (1968), is that the meaning of a word is related to its distributional pattern in language. This means that if two words frequently occur in similar context, we may assume that they have similar meanings. This assumption is known as "the Distributional Hypothesis," and it is the ultimate rationale for statistical approaches to semantic knowledge acquisition, such as Simple Recurrent Networks or Vector-Based Semantic Analysis.

Simple Recurrent Networks

Simple Recurrent Networks (SRN; Elman, 1990) are a class of Artificial Neural Networks consisting of the three traditional 'input', 'hidden' and 'output' layers of units, to which one additional layer of 'context' units is added. The basic architecture of an SRN is shown in Figure 9.1. The outputs of the 'context' units are connected to the inputs of the 'hidden' layer as if they formed and additional 'input' layer. However instead of receiving their activation from outside, the activations of the 'context' layer at time step n are a copy of the activations of the 'hidden' layer at time step $n-1$. This is achieved by adding simple, one-to-one 'copy-back' connections from the 'hidden' layer into the 'context' layer. In contrast to all the other connections in the network, these are special in that they are not trained (their weights are fixed at 1), and in that they perform a raw copy operation from a hidden unit into

a context unit, that is to say, they employ the identity function as the activation function. Networks of this kind combine the advantages of recurrent networks, their capability of maintaining a history of past events, with the simplicity of multilayer perceptrons as they can be trained by the backpropagation algorithm.

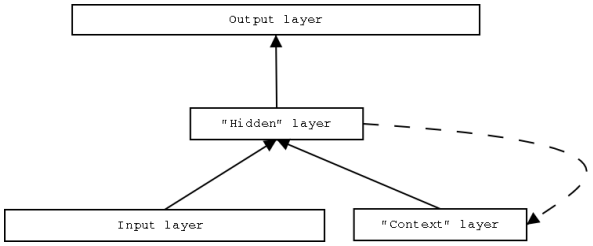

Figure 9.1: Modular architecture of a Simple Recurrent Network. The boxes correspond to layers of units. The solid arrows represent sets of trainable ‘all-to-all’ connections between the units in two layers. The dashed arrow stands a for fixed ‘one-to-one’ not trainable connection between two layers. These connections have the function of copying the activation of the hidden units into the context units at every time step.

Elman (1993) trained an SRN on predicting the next word in a sequence of words, using sentences generated by an artificial grammar, with a very limited vocabulary (24 words). He showed that a network of this class, when trained on a word prediction task and given the right training strategy (see Rohde and Plaut, 2001) for further discussion of this issue), acquired various grammatical properties such as verbal inflection , plural inflection of nouns, argumental structure of verbs or grammatical category. Moreover, the activations of the hidden units of the network provided detailed, token-specific characterizations of the morpho-syntactic properties of a word. Moscoso del Prado and Baayen (to appear) showed how this method can be extended to deal with the large vocabulary sizes of realistic corpora. They trained a network akin to those of (Rohde and Plaut, 1999; Elman, 1993) on a word prediction task, using moderately large corpora of written English and Dutch (approximately 700,000 tokens of English and 4,500,000 of Dutch). The hidden units again provided rich representations of the morpho-syntactic properties of the words, containing information ranging from grammatical category, to subtle inflectional details such as verbal inflection or adjective gender. Moreover, the network had also captured some semantic properties of words, namely semantic properties that can be inferred from syntactic properties such as argumental structure.

Vector-Based Semantic Analysis

While SRN's appear to be more sensitive towards syntactic features, vector-space models have been used for over a decade to acquire and represent semantic information about words, documents and other linguistic units. This is done by collecting co-occurrence information in a words-by-contexts matrix F where each row F_w represents a unique word and each column F_c represents a context, which can either be a multi-word segment such as a document or another word. Latent Semantic Analysis/Indexing (LSA/LSI; Deerwester et al., 1990; Landauer and Dumais, 1997) uses document-based co-occurrence statistics, while Hyperspace Analogue to Language (HAL; Lund et al., 1995) and Schütze (1992) use word-based statistics. The cells of the matrix indicate the (weighted and/or normalized) frequency of ocurrence in, or co-occurrence with, the co-occurrence context (i.e. documents or words). Vector-space models generally also use some form of dimension reduction to reduce the computational strains of dealing with the rather ungainly co-occurrence matrix. LSA uses Singular Value Decomposition (SVD) and HAL uses a "column variance method," which consists in discarding the columns with lowest variance. This reduces the dimensionality of the co-occurrence matrix to a fraction of its original size. linguistic units are thus represented in the final reduced matrix by semantic vectors of n dimensionality. LSA is reported to be optimal at $n = 300$ (Landauer & Dumais, 1997), HAL at $n = 200$ (Lund et al., 1995), while Schutze (1992) uses $n = 20$. A different approach to create the vector-space is Random Indexing (Kanerva et al., 2000; Karlgren & Sahlgren, 2001), which avoids the inefficient and inflexible dimensionality reduction phase by using high-dimensional sparse, random *index vectors* to accumulate a words-by-contexts matrix in which words are represented by high-dimensional (i.e. n is in the order of thousands) *context vectors*.

Vector-space methodology has been empirically validated in a number of experiments as a viable technique for the automatic extraction of semantic information from raw, unstructured text data. For example, Landauer and Dumais (1997) report a result on a standardized vocabulary test (TOEFL; Test of English as a Foreign Language) that is comparable to the average performance of foreign (non-English speaking) applicants to U.S. colleges (64.4% vs. 64.5% correct answers to the TOEFL). Sahlgren (to appear), showed that similar performance (64.5% – 67%) may be obtained by using distributed representations in the Random Indexing technique that eliminates the need for the computationally expensive SVD, and he also demonstrated that the performance may be further enhanced (72% correct

answers) by taking advantage of explicit linguistic information (morphology). Further empirical evidence can be found in, for example, Lund and Burgess (1996), who used semantic vectors to model reaction times from lexical priming studies, and from Landauer and Dumais (1997), who used LSA for evaluating the quality of content of student essays on given topics. Thus, it appears to be beyond doubt that the vector-space methodology really is able to form high-quality semantic representations by using such a simple souce of information as plain co-occurrence statistics. In the remainder of this paper, we will use the label *Vector-Based Semantic Analysis* to denote the practice of using co-occurrence information to construct vectors representing linguistic units in a high-dimensional semantic space.

In this study, we integrate two techniques to automatically obtain distributed lexical representations from corpora encoding morpho-syntactic and semantic information simultaneously. A hybrid technique such as the one that we describe here has several advantages. First, it requires a minimum of preexisting lexical resources, as it depends only on raw corpora. There is no need for taggers or parsers which, for many languages, may be unavailable.

Second, in contrast to other approaches that exploit word co-occurrences, our method keeps computational costs under control, as we avoid having to deal with huge co-occurrence matrices and we do not need to apply dimensional reduction techniques such as Singular Value Decomposition or Principal Component Analysis. The use of such dimensional reduction techniques imposes important limitations on the extension of existing resources, as the addition of a new item would require that a new reduced similarity space is calculated. In contrast, both SRN and the VBSA technique allow for the direct inclusion of new data. Another important advantage of our approach is that lexical representations become dynamic in nature: each token of a given type will have a slightly different representation.

We produce explicit measures of reliability that are directly associated to each distance calculated by our method. This is particularly useful for extending existing lexical resources such as computational thesauri.

In what follows, we introduce the corpus employed in the experiment, together with the SRN and VBSA techniques that we used. We then evaluate the grammatical knowledge encoded in the distributed representations obtained by the model. We subsequently evaluate the semantic knowledge contained in the system by means of scores on language proficiency tests (TOEFL), comparison with synonyms in WordNet, and a comparison of the properties of morphological variants. We conclude by discussing the possible application of this technique to bootstrap

lexical resources from untagged corpora and the cognitive implications of these results.

Experiment

Corpus

For the training of the SRN network, we used the texts corresponding to the first 20% of the British National Corpus; by first we mean that we selected the files following the order of directories, and we included the first two directories in the corpus. This corresponds to roughly 20 million tokens. To allow for comparison with the results from (Sahlgren, to appear), which were based on a 10 million word corpus, only the first half of this subset was used in the application of the VSBA technique.

Only a naive preprocessing stage was performed on the original SGML files. This included removing all SGML labels from the corpus, converting all words to lower case, substituting all numerical tokens for a $[num]$ token and separating hyphenated compound words into three different tokens ($firstword + [hyphen] + secondword$). All tokens containing non alphabetic characters different from the common punctuation marks were removed from the corpus. Finally, to reduce the vocabulary size, all tokens that were below a frequency threshold of two, were substituted by an $[unknown]$ token.

Design and Training of the SRN

The Simple Recurrent Network followed the basic design shown in Figure 9.1. We used a network with 300 units in the input and output layers, and 150 units in the hidden and context layers. To allow for representation of a very large number of tokens, we used the semi-localist approach described in Moscoso del Prado and Baayen (to appear) with a code of three random active units per word. On the one hand, this approach is close to a traditional style one-bit-per-word localistic representation in that the vectors of two different words will be nearly orthogonal. The small deviation from full orthogonality between representations has an effect similar to the introduction of a small amount of random noise, which actually speeds up the learning process. On other the hand, using semi-distributed input/output representations allows us to represent a huge number of types (a maximum of $\binom{300}{3} = 4,455,100$ types), while keeping the size of the network moderately small.

The sentences of the corpus were grouped into ‘examples’ of five consecutive sentences. At each time step, a word was presented to the input layer and the network would be trained to predict the following word in the output units. The corpus sentences were presented word by word in the order in which they appear. After every five sentences (a full ‘example’), the activation of the context units was reset to 0.5. Imposing limitations on the network’s memory on the initial stages of training is a pre-requisite for the networks to learn long distance syntactic relations (Elman, 1993; but see also, Rohde and Plaut, 1999). We implemented this ‘starting small’ strategy by introducing a small amount of random noise (0.15) in the output of the hidden units, and by gradually reducing to zero during training. At the same time that the random noise in the context units was being reduced, we also gradually reduced the learning rate, starting with a learning rate of 0.1 and finished training with a learning rate of 0.4. Throughout training, we used a momentum of 0.9.

Although the experiments in Elman (1993) used the traditional backpropagation algorithm, using the mean square error as the error measure to minimize, following Rohde and Plaut (1999) we substituted the training algorithm for a modified momentum descent using cross-entropy as our error measure,

$$\sum_i \left[t_i \log \left(\frac{t_i}{o_i} \right) + (1 - t_i) \log \left(\frac{1 - t_i}{1 - o_i} \right) \right] \tag{9.1}$$

Modified momentum descent enables stable learning with very aggressive learning rates as the ones we use. The network was trained on the whole corpus of 20 million for one epoch using the *Light Efficient Network Simulator* (LENS; Rohde, 1999).

Application of VBSA Technique

Once the SRN had been trained, we proceeded to apply the Vector Based Semantic Analysis technique. Sahlgren (to appear) used what he called ‘random labels’. These were sparse 1800 element vectors, in which, for a given word type, only a small set of randomly chosen elements would be active (± 1.0), while the rest would be inactive. Once these initial labels had been created, the corpus was processed in the following way. For each token in the corpus, the labels of the s immediately preceding or following tokens were added to the vector of the word (all vectors were initialized to a set of 0’s). The addition would be weighted giving more importance to the closer word in the window. Words outside a frequency range of $(3 - 14,000)$ are not included in these sums. This range excludes both the very frequent types, typi-

cally function words, and the least frequent types, about which there is not enough information to provide reliable counts. Optimal results are obtained with a window size ($s = 3$), that is, by taking into account the three preceeding and following words to a given token. In order to reduce sparsity, Sahlgren used a lemmatizer to unify tokens representing inflectional variants of the same root. Sahlgren had also observed that the inclusion of explicit syntactic information extracted by a parser did not improve the results, but led to lower performance. We believe that this can be partly due to the *static* character of the syntactic information that was used. We therefore use a *dynamic* coding of syntactic information, which is more sensitive to the subtle changes in grammatical properties of each different instance of a word.

In our study, we substituted the knowledge-free random labels of Sahlgren (to appear) by the dynamic context-sensitive representations of the individual tokens as coded in the patterns of activations of our SRN. Thus each type is represented by a slightly different vector for each different grammatical context in which it appears. To obtain these representation, we presented the text to the SRN and used the activation of the hidden units to provide the dynamic labels for VBSA

We then used a symmetric window of three words to the left and right of every word. We fed the text again through the neural network in test mode (no weight updating), and we summed the activation of the hidden units of the network for each of the words in the context window that fall within a frequency range of 8 and $30,000$ in the original corpus (the one that was used for the training of the neural network). In this way we excluded low frequency words about which the network might be extremely uncertain, and extremely high frequency function words. We used as weighting schema $w = 2^{1-d}$, were w is the weight for a certain position in the window, and d is the distance in tokens from that position to the center of the window. For instance, the label of the word following the target would be added with a weight $w = 2^{1-1} = 1$ and the label of the word occupying the leftmost position in the window would have a weight $w = 2^{1-3} = 0.25$. When a word in the window was out of the frequency range, its weight was set to 0.0. Punctuation marks were not included in window positions.

Table 9.1: Sample of 5 nearest neighbors to some words according to normalized cosine distance. Semantically unrelated words are marked by an asterisk

Word	Nearest neighbors
hall	centre, theatre, chapel, landscape*, library
half	period, quarter, phase, basis, breeze*
foreigners	others, people, doctors, outsiders, unnecessary*
legislation	orders, contracts, plans, losses, governments
positive	splendid, vital, poetic, similar*, bad
slightly	somewhat, distinctly, little, fake*, supposedly
subjects	issues, films, tasks, substances, materials
taxes	debts, rents, imports, investors, money
render	expose, reveal, extend, ignoring*, develop
re-	anti-, non-, pro-, ex-, pseudo-
omitted	ignored, despised, irrelevant, exploited*, theirs*
Bach	Newton, Webb, Fleming, Emma, Dante

Results

Overview of Semantics by Nearest Neighbors

We begin our analysis by inspecting the five nearest neighbors for a given word. Some examples can be found in Table 9.1Ṫo calculate the distances between words, we use normalized cosines (Schone & Jurafsky, 2001). Traditionally, high dimensional lexical vectors have been compared using metrics such as the cosine of the angle between the vectors or the classical Euclidean distance metric or the city-block distance metric. However, using a fixed metric on the components of the vectors induces undesirable effects pertaining to the centrality of representations. More frequent words tend to appear in a much wider range of contexts. When the vectors are calculated as an average of all the tokens of a given type, the vectors or more frequent words will tend to occupy more central positions in the representational space. They will tend to be nearer to all other words, thus introducing an amount of relativity in the distance values. In fact, we believe that this relativity actually reflects people's understanding of word meaning. For example, if we considered the most similar words to a frequent word such as "bird", we would find words as "pigeon" to be very related in meaning. A word such as "penguin" would be considered a more distantly related word. However, if we examined the nearest neighbors of "penguin", we would probably find "bird" among them, although the standard distance measure would still be high. A way to overcome this problem is

to place word distances inside a normal distribution, taking into account the distribution of distances of both words. Consider the classical cosine distance between two vectors $\mathbf{v}$ and $\mathbf{w}$:

$$d_{\cos}(\mathbf{v}, \mathbf{w}) = 1 - \frac{\mathbf{v} \cdot \mathbf{w}}{||\mathbf{v}|| \, ||\mathbf{w}||}. \tag{9.2}$$

For each vector $\mathbf{x} \in \{\mathbf{v}, \mathbf{w}\}$ we calculate the mean ($\mu_{\mathbf{x}}$) and standard deviation ($\sigma_{\mathbf{x}}$) of its cosine distance to 500 randomly chosen vectors of other words. This provides us with an estimate of the mean and standard deviation of the distances between $\mathbf{x}$ and all other words. We can now define the normalized cosine distance between two vectors $\mathbf{v}$ and $\mathbf{w}$ as:

$$d_{norm}(\mathbf{v}, \mathbf{w}) = \max_{\mathbf{x} \in \{\mathbf{v}, \mathbf{w}\}} \left(\frac{d_{\cos}(\mathbf{v}, \mathbf{w}) - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}} \right). \tag{9.3}$$

To speed up this process, the cosine distance means and standard deviation for all words were pre-calculated in advance and stored as part of the representation. The use of normalized cosine distance has the effect of allowing for direct comparisons of the distances between words. In our previous example the distance between “bird” and “penguin”, according to a non-normalized metric would suffer from the eccentricity of “penguin”; with the normalization, as the value of the distance would be normalized with respect to “penguin” (the maximum), it would render a value similar to the distance between “bird” and “pigeon”.

Grammatical Knowledge

Moscoso del Prado and Baayen (to appear) showed that the hidden unit representations of SRN’s similar to the one we used here contain information about morpho-syntactic characteristics of the words. In the present technique this information is implicitly available in the input labels for the VBSA technique. The VBSA component however, does not guarantee the preservation of such syntactic information. We therefore need to ascertain whether the grammatical knowledge contained in the SRN vectors is preserved after the application of VBSA.

Note that in Table 9.1, the nearest neighbors of a given word tend to have similar grammatical attributes. For example, plural nouns have other plural nouns as nearest neighbors, e.g., “foreigners” - “others”, “outsiders”, etc., and verbs tend to have other verbs as nearest neighbors, e.g., “render” - “expose”, “reveal”, etc. Although the nearest neighbors in Table 9.1 clearly suggest that morpho-syntactic information is coded in the representations, we need to ascertain how much morpho-

syntactic information is present and, more importantly, how easily it might be made more explicit. We do this using the techniques proposed by Moscoso del Prado and Baayen, (to appear), that is we employ a machine learning technique using our vectors as input and symbolic grammatical information extracted from the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995) as output. A machine learning system is trained to predict the labels from the vectors. The rationale behind this method is very straightforward: If there is a distributed coding of the morpho-syntactic features hidden inside our representation, a standard machine learning technique should be able to detect it.

We begin by assessing whether the grammatical category of a word can be extracted from its vector representation. We randomly selected 500 words that were classified by CELEX as being unambiguously nouns or verbs, that is, they did not have any other possible label. The nouns were sampled evenly between singular an plural nouns, and the verbs were sampled evenly between infinitive, third person singular and gerund forms. Using TiMBL (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 1999), we trained a memory based learning system on predicting whether a vector corresponded to a noun or a verb. We performed ten-fold cross-validation on the 500 vectors. The systems were trained using 7 nearest neighbors according to a city-block distance metric, the contribution of each component of the vectors weighted by Information Gain Feature Weighting (Quinlan, 1993). To provide a baseline against which to compare the results, we use a second set of files consisting of the same vectors but with random assignment of grammatical category labels to words. The average performance of the system of the Noun-Verb distinction was 68% (randomized average 56%). We compared the performance of the system with that of the randomized labels system using a paired two-tailed t-test on the result of each of the runs in the cross-validation, which revealed that the performance of the system was significantly higher than that of the randomized one ($t = 5.63,\ df = 9,\ p = 0.0003$).

We also tested for more subtle inflectional distinctions. We randomly selected 300 words that were unambiguously nouns according to CELEX, sampling evenly from singular and plural nouns. We repeated the test described in the previous paragraph, with the classification task this time being the differentiation between singular and plural. The average performance of the machine learning system was 65% (randomized average 48%). A paired two-tailed t-test comparing the results of the systems with the results of systems with the labels randomized revealed again a significant advantage for the non-random system ($t = 5.80,\ df = 9, p = 0.0003$). The

same test was performed on a group of 300 randomly chosen unambiguous verbs sampled evenly among infinitive, gerund and third person singular forms, with these labels being the ones the system should learn to predict from the vectors. Performance in differentiating these verbal inflections was of 55% on average while the average of randomized runs was 33%, and significantly above randomized performance accoriding to a paired two-tailed t-test ($t = 4.25$, $df = 9$, $p = 0.0021$).

Performance in TOEFL Synonyms Test

Previous studies (Sahlgren, to appear; Landauer and Dumais, 1997) evaluated knowledge about semantic similarity contained in co-occurrence vectors by assessing their performance in a vocabulary test from the Test of English as a Foreign Language (TOEFL). This is a standardized vocabulary test employed by, for instance, American universities, to assess foreign applicants' knowledge of English. In the synonym finding part of the test, participants are asked to select which word is a synonym of another given word, given a choice of four candidates that are generally very related in meaning to the target. In the present experiment, we used the selection of 80 test items described in Sahlgren (to appear), with the removal of seven test items which contained at least one word that was not present in our representation. This left us with 73 test items consisting of a target word and four possible synonyms. To perform the test, for each test item, we calculated the normalized cosine distance between the target word and each of the candidates, and chose as a synonym the candidate word that showed the smallest cosine distance to the target. The model's performance on the test was 51% of correct responses.

Reliability Scores

The results of this test can be improved once we have a measure of the certainty with which the system considers the chosen answer to be a synonym of the target. What we need is a reliability score, according to which, in cases where the chosen word is not close enough in meaning, i.e., its distance to the target is below a certain probabilistic threshold, the system would refrain from answering. In other words, the system would be allowed to give an answer such as: "I'm not sure about this one". Given that the values of the distances between words in our system, follow a normal distribution $\mathcal{N}(0, 1)$, it is quite straightforward to obtain an estimate of the probability of the distance between two words being smaller than a given value, by just using the Normal distribution function $F(x)$. However, while the *general* dis-

tribution of distances between any two given words follows $\mathcal{N}(0,1)$, the distribution of the distances from a *particular* word to the other words does not necessarily follow this distribution. In fact they generally do not do so. This difference in the distributions of distances of words is due to effects of prototypicality and probably also word frequency (McDonald & Shillcock, 2001) .

To obtain probability scores on how likely it is that a given word is at a certain distance from the target, we need to see the distance of this word *relative* to the distribution of distances from the target word to all other words in the representation. We therefore slightly modify 9.3, which takes the normalized distance between two words to be the maximum of the cosine distance normalized according to the distribution of distances to the first word, and the cosine distance normalized to the distribution of distances to the second word. We now define the cosine distance between two vectors $\mathbf{v}$ and $\mathbf{w}$ normalized relative to $\mathbf{v}$ as:

$$d^{\mathbf{v}}_{norm} = \frac{d_{\cos}(\mathbf{v},\mathbf{w}) - \mu_{\mathbf{v}}}{\sigma_{\mathbf{v}}}, \tag{9.4}$$

which provides us with distances that follow $\mathcal{N}(0,1)$ for each particular word represented by a vector $\mathbf{v}$.

Using 9.4, we calculated the distance between the target words in the synonym test and the word that the system had selected as most similar, counting only those answers for which the system outputs a probability value below 0.18. The performance on the test increases from 51% to 71%, but the number of items reduced to 45. If we choose probability values below 0.18, the percentage correct continues to rise, but the number of items in the test drops dramatically. Having such a reliability estimator is useful for real-world applications.

Performance for WordNet Synonyms

We can also use the WordNet (Miller, 1990) lexical database to further assess the amount of word similarity knowledge contained in our representations. We randomly selected synonym pairs from each of the four grammatical categories contained in WordNet: nouns, verbs, adjectives and adverbs. We calculated the normalized cosine distance for each of the synonym pairs. As expected, the median distances between synonymous words were clearly smaller than average distance. The median distances were -0.59 for verb synonyms, -0.53 for noun synonyms, -0.49 for adjective synonyms and -0.62 for adverbial synonyms. However, as we have already seen, our vectors contain a great deal of information about morpho-

syntactic properties. Hence the fact that synonyms share the same grammatical category could by itself explain the small distances obtained for WordNet synonyms. To check whether this is the case, each synonym pair from our set was coupled with a randomly chosen baseline word of the same grammatical category, and we calculated the distance between one of the synonyms and the baseline word. In this case, as we were interested in the distance of the word relative only to one of the words in the pair, we calculated distances using 9.4. We compared the series of distances obtained for the true WordNet synonym pairs with the baseline distances by means of two-tailed t-tests. We found that WordNet synonyms were clearly closer in all the cases: nouns ($t = -5.30$, $df = 197$, $p < 0.0001$), verbs ($t = -4.60$, $df = 190$, $p < 0.0001$), adjectives ($t = -3.09$, $df = 195$, $p = 0.0023$) and adverbs ($t = -4.06$, $df = 188$, $p < 0.0001$). This shows that true synonyms were significantly closer in distance space than baseline words.

Morphology as a Measure of Meaning

Morphologically related words tend to be related both in form and meaning. This is true both for inflectionally related words, and derivationally related words. As morphological relations tend to reflect regular correspondences to slight changes in the meaning and syntax, they can be used for assessing the amount of semantic knowledge that has been acquired by our system. In what follows, we investigate whether our system is able to recognize inflectional variants of the same word, and whether the vectors of words belonging to the same suffixation class cluster together.

Inflectional Morphology

We randomly selected 500 roots that were unambiguously nominal (they did not appear in the CELEX database under any other grammatical category) and for which both the singular and the plural form were present in our dataset. For each of the roots, we calculated the normalized cosine distance between the singular and plural forms. The median of the distance between singular and plural forms was -0.39, which already indicates that inflectional variants of the same noun are represented by similar vectors. As in the case of the WordNet synonyms, it could be argued that this below average distance is completely due to all these word pairs sharing the “noun” property. To ascertain that the observed effect on the distances was at least partly due to real similarities in meaning, each stem r_1 in our set

was paired with another stem r_2 also chosen from the original set of 500 nouns. We calculated the normalized cosine distance between the singular form of r_1 and the plural form of r_2. In this way we constructed a data set composed of word pairs plus their normalized cosine distance. A linear mixed effect model (Pinheiro & Bates, 2000) fit to the noun data with normalized cosine distance as dependent variable, the 'stem' (same v. other) as independent variable and the root of the present tense form as random effect, revealed a main effect for stem-sharing pairs ($F(1, 499) = 44.42$, $p < 0.0001$). The coefficient of the effect was -0.29 ($\hat{\sigma} = 0.043$). This indicates that the distances between pairs of nouns that share the same stem are in general smaller than the distance between pairs of words that do not share the same root but have the same number. Interestingly, according to a Pearson correlation, 65% of the variance in the distances is explained by the model.

In the same way, we randomly selected 500 unambiguously verbal roots for which we had the present tense, past tense, gerund and third person singular present tense in our representation. The median normalized cosine distance between the present tense and the other forms of the verb was -0.48, so verbs seem to be clustered together somewhat more tightly than nouns. We repeated the test described above by random pairing of stems, but now we calculated the distances between the present tense form of r_1 and the rest of the inflected forms of r_2. We fit a linear mixed effect model with the normalized cosine distance between the pairs as dependent variable, the pair of inflected forms, i.e., *present-past*, *present-gerund*, or *present-third person singular*, and the 'stem' (same versus different) as independent variables and the root of the first verb as random effect. We found significant, independent effects for type of inflectional pair ($F(1, 2495) = 289.06$, $p < 0.0001$) and stem-sharing ($F(1, 2495) = 109.76$, $p < 0.0001$). The interaction between both independent variables was not significant ($F < 1$). The coefficient for the effect of sharing a root was -0.18 ($\hat{\sigma} = 0.017$), which again indicates that words that share a root have smaller distances than words that do not. It is also interesting to observe that the coefficients for the pairs of inflected forms also provide us with information of how similarly these forms are used in natural language, or, phrased in another way, how similar their meanings are. So, the value of the coefficient for pairs of present tense (uninflected) and past tense forms was -0.48 ($\hat{\sigma} = 0.21$) and the coefficient for pairs composed of a present tense uninflected form and a past tense was -0.38 ($\hat{\sigma} = 0.21$), which suggests that the contexts in which an un-inflected form is used are more similar to the contexts where a past tense form is used than to the contexts of a gerund. The model explained 43% of the variance according to

a Pearson correlation.

Derivational Morphology

Derivational morphology also captures regular meaning changes, although these changes are often not as regular as the ones that are carried out by inflectional morphology. We tested whether our system captures derivational semantics using the Memory-Based Learning technique that we used for evaluating grammatical knowledge in the system (see section 9). Concentrating on morphological categories, i.e. on words that share the same outer affix. For instance “compositionality” belongs to the morphological category “-ity” and not to the category “-al”, although it also contains the suffix “-al”. Derivational suffixes generally effect both syntactic and semantic changes. To test whether our vectors reflect semantic regularities, we selected all words ending in the two derivational suffixes “-ist” and “-ness”. Both of these suffixes produce nouns, but while the first one generates nouns that are considered agents of actions, the second generates abstract ideas. These affixes generate words with the same grammatical category, but with different semantics. We trained a TiMBL system on predicting the morphological category of the vectors, that is, to predict “-ist” or “-ness”. The average performance of the system in predicting these labels in a ten-fold cross-validation was of 78% (compared to an average of 51% obtained when randomizing the affix labels). A paired two-sided t-test between the system performance at each run and the performance of a randomized system on the same run, revealed a significant improvement for the non random system ($t = 10.95,\ df = 9,\ p < 0.0001$).

Although performance was very good for these two nominal affixes, a similar comparison between the adjectival affixes “-able” and “-less”, did not render significant differences between randomized and non-randomized labels, indicating that the memory-based learning system was not able to discriminate these two affixes on the sole basis of their semantic vectors. This indicates that, although some of the semantic variance produced by derivational affixes can be captured, many subtler details are being overlooked.

General Discussion

The analyses that we have performed on the vectors indicate that a high amount of lexical information has been captured by the combination of an SRN with VBSA. On the one hand, the results reported in section 9 indicate that the morpho-syntactic information that is coded in the hidden units of a SRN is maintained after the application of VSBA. Moreover, it is clear that the coding of the morpho-syntactic features can be extracted using a standard machine-learning technique such as Memory-Based Learning. This, by itself can be of great use in the bootstrapping of language resources. Given a fairly small set of words that have received morpho-syntactic tags, it is possible to train a machine learning system to identify these labels from their vectors, and then apply this to the vectors of words that are yet to receive morpho-syntactic tagging. Importantly, our technique relies only on word-external order and co-occurrence information, but does not make use of word-internal form information. As it it is evident that word-form information such as presence of inflectional affixes is crucial for morpho-syntactic tagging, our technique can be used to provide a confirmation of possible inflectional candidates. For instance, suppose that two words such as “rabi” and “rabies” are found in a corpora, one would be inclined to classify them as singular and plural version of the same word, when in fact they are both singular forms. The inflectional information in our vectors could be used to disconfirm this hypothesis. In this same aspect, the fact that inflectional variants of the same root tend to be very related in meaning could be used as additional evidence to reject this pair as being inflectional variants.

On the other hand, the nearest neighbors, the TOEFL scores, the results on detecting inflectionally and derivationally related words, and the results on the WordNet synonyms, provide solid evidence that the vectors have succeeded in capturing a great deal of semantic information. Although it is clear to us that our technique needs further fine-tuning, the results are already surprising given the constraints that have been imposed on the system. For instance, the performance on the TOEFL test (51% without the use of the Z scores) is certainly lower than many results that have been reported in the literature. Sahlgren (tp appear), using the Random Indexing approach to VBSA with random vectors reports 72% correct responses on the same test items. However, he was using a tagged corpus where all inflectional variants had been unified under the same type. Without the use of stemming, the best performance he reports is of 68%. In the current approach we have used vectors of 150 elements, that is, less than 10% of the size of the vectors used by Sahlgren, and much smaller than the vectors needed to apply techniques such

Hyperspace Analog to Language (Lund et al., 1995; Lund & Burgess, 1996) or Latent Semantic Analysis (Landauer & Dumais, 1997) which need to deal with huge co-occurrence matrices. Given the computational requirements of using such huge vectors, we consider that our method provides a good alternative. Our result of 51% on the TOEFL test is clearly above chance performance (25%) and not that far from the results obtained by average foreign applicants to U.S. universities (64.5%). Interestingly, Landauer and Dumais (1997) reported a 64.4% performance on these test items using LSA, but this was only after the application of a dimensional reduction technique (SVD) to their original document co-occurrence vectors. Before the application of SVD, they report a performance of 36% on the plain normalized vectors. Of course, a technique such as SVD could be subsequently applied to the vectors obtained by our method, probably leading to some improvement in our results. However, given that our vectors already have a moderate size, and especially, given that, in their current state, one does not need to re-compute them to add information contained in new corpora, we do not favor the use of such techniques.

Regarding the evaluation of the system against synonym pairs extracted from the WordNet database, although the vectors represent synonyms as being more related than average, it still seems that most of the similarity in these cases was due to morpho-syntactic properties (the average difference in distances between the synonym and baseline conditions was always smaller than 0.1). We believe this is due to several factors. WordNet synonym sets (*synsets*) contain an extremely rich amount of information, that may be too rich for the purposes of evaluating our current vectors. First, many WordNet synonyms correspond to plain spelling variants of the same word in British and American English, e.g., “analyze”-“analyse”. Our whole training corpus was composed of British English, so the representation of words in American spelling is probably not very accurate. Second, and more importantly, given that the synsets encoded in WordNet reflect in many cases rare or even metaphoric uses of words, we think that the evaluation based on the average type representations provided by our system are not the most appropriate to detect these relations. Possibly, evaluating these synonyms against the vectors corresponding to the particular tokens referring to those senses might be more appropriate. An indication of this is also given by the TOEFL scores, which reflect that the meaning differences can still be detected in many cases. This is important because the synonyms pairs chosen in the TOEFL test, generally reflect the more standard senses of the words involved.

Another important issue is the difference between meaning *relatedness* and

meaning *similarity*. These are two different concepts that appear to be somewhat confounded. While our representations reflect in many cases similarity relations, e.g, synonymy, they also appear to capture many relatedness and general world knowledge relations, for instance, the three nearest neighbors of “student” are “university” “pub” and “study”, none of which is similar in meaning to “student”, but all of them bearing a strong relationship to it. Sahlgren (to appear) argues that using a small window to compute the co-occurrences (3 elements to each side, as compared to the 10 elements used by Burgess & Lund, 1998), has the effect of concentrating on similarity relations instead of relatedness, which would need much larger contexts such as the full documents used in LSA. The motivation to use very small context windows was to provide an estimation of the syntactic context of words. However, since syntactic information is already made more explicit by our SRN this may not be necessary in our case, and using larger window sizes might actually improve our performance both in similarity and relatedness. A further improvement that should be added to our vectors should come from the inclusion of word internal information. In a pilot experiment we have used the VBSA technique using (automatically constructed) distributed representations of the formal properties of words instead of the random labels. Performance on the TOEFL test were in the same range that was reported here (49%). This suggest that a combination of the technique described here with the formal vectors could probably provide much more precise semantic representations, exploiting both word internal and internal sources of information. This is also in line with the improvement of results found by (Sahlgren, to appear) when using a stemming technique. The use of formal vectors provides an interesting alternative, as it would supply implicit stemming information to the system.

In this paper, we have presented a representation that encodes jointly morpho-syntactic and semantic aspects of words. We have also provided evidence on how morphology is an important cue to meaning, and vice-versa, meaning is also an important cue to morphology. This corroborates previous results from (Schone & Jurafsky, 2001). The idea of integrating formal, syntactic and semantic knowledge about words in one single representation is currently gaining strength within the psycholinguistic community (e.g., Gaskell & Marslen-Wilsonm, 1997; Plaut & Booth, 2000). Some authors are considering morphology as the “convergence of codes”, that is, as a set of quasi-regular correspondences between form and meaning, that would probably be linked at a joint representation level (Seidenberg & Gonnerman, 2000). Clear evidence of this strong link has also been put forward

by Ramscar (2002) showing that the choice of regular or non-regular past tense inflection of a nonce verb is strongly influenced by the context in which the nonce verb appears. If the word appears in a context which entails a meaning similar to that of an irregular verb that is also similar in form to the nonce word, e.g. “frink” - “drink”, participants form its past tense in the same manner as the irregular form, e.g., “frank” from “drank”. If it appears in a context alike to a similar regular verb, e.g, “wink”, participants inflect in regularly, e.g. “frinked” from “winked”. Crucially, the meaning of this form is totally determined by context. This in line with the results of McDonald and Ramscar (2001), which show how the meaning of a nonce word is modulated by the context in which it appears. In this respect, our vectors constitute a first approach to such kind of representation: they include contextual and syntactic information. A further step will be the inclusion of word form information in this system, which is left for future research. Our lexical representations are formed by accumulation of predictions. On the one hand, several authors are currently investigating the strong role played by anticipation and prediction in human cognitive processing. On the other hand, some current models of human lexical processing include the notion of accumulation, generally by recurrent loops in the semantic representations (e.g., Plaut & Booth, 2000).

References

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX lexical database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Burgess, C. and Lund, K.: 1998, The dynamics of meaning in memory, *in* E. Dietrich and A. B. Markman (eds), *Cognitive dynamics: Conceptual change in humans and machines*, Lawrence Erlbaum Associates.

Daelemans, W., Zavrel, J., Van der Sloot, K. and Van den Bosch, A.: 1999, TiMBL: Tilburg memory based learner reference guide 2.0, *Report 99-01*, Computational Linguistics Tilburg University.

Elman, J. L.: 1990, Finding structure in time, *Cognitive Science* **14**, 179–211.

Elman, J. L.: 1993, Learning and development in neural networks: The importance of starting small, *Cognition* **48**, 71–99.

Gaskell, M. G. and Marslen-Wilson, W.: 1997, Integrating form and meaning: A distributed model of speech perception, *Language and Cognitive Processes* **12**, 613–656.

Landauer, T. and Dumais, S.: 1997, A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge, *Psychological Review* **104**(2), 211–240.

Lund, K. and Burgess, C.: 1996, Producing high-dimensional semantic spaces from lexical co-occurrence, *Behaviour Research Methods, Instruments, and Computers* **28**(2), 203–208.

Lund, K., Burgess, C. and Atchley, R. A.: 1995, Semantic and associative priming in high-dimensional semantic space, *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, Erlbaum, Hillsdale, NJ., pp. 660–665.

McDonald, S. and Ramscar, M.: 2001, Testing the distributional hypothesis: The influence of context judgements of semantic similarity, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

McDonald, S. and Shillcock, R.: 2001, Rethinking the word frequency effect: The neglected role of distributional information in lexical processing, *Language and Speech* **44**, 295–323.

Miller, G. A.: 1990, Wordnet: An on-line lexical database, *International Journal of Lexicography* **3**, 235–312.

Moscoso del Prado Martín, F. and Baayen, R. H.: to appear, Unsupervised ex-

traction of high-dimensional lexical representations from corpora using simple recurrent networks, *in* A. Lenci, S. Montemagni and V. Pirrelli (eds), *The Acquisition and Representation of Word Meaning*, Linguistica Computazionale, Pisa, Italy.

Pinheiro, J. C. and Bates, D. M.: 2000, *Mixed-effects models in S and S-PLUS*, Statistics and Computing, Springer, New York.

Plaut, D. C. and Booth, J. R.: 2000, Individual and developmental differences in semantic priming: Empirical and computational support for a single mechanism account of lexical processing, *Psychological Review* **107**, 786–823.

Quinlan, J. R.: 1993, *Programs for Machine Learning*, Morgan Kauffmann, San Mateo, CA.

Ramscar, M.: 2002, The role of meaning in inflection: Why the past tense doesn't require a rule, *Cognitive Psychology* **45**, 45–94.

Rohde, D. L. T.: 1999, LENS: The light, efficient network simulator, *Technical Report CMU-CS-99-164*, Carnegie Mellon University, Pittsburg, PA.

Rohde, D. L. T. and Plaut, D. C.: 1999, Language acquisition in the absence of explicit negative evidence: how important is starting small?, *Cognition* **72**(1), 67–109.

Sahlgren, M.: to appear, Vector-based semantic analysis: Representing word meanings based on random labels, *in* A. Lenci, S. Montemagni and V. Pirrelli (eds), *The Acquisition and Representation of Word Meaning*, Kluwer, Dordrecht, The Netherlands.

Schone, P. and Jurafsky, D.: 2001, Knowledge free induction of inflectional morphologies, *Proceedings of the North American Chapter of the Association for Computational Linguistics NAACL-2001*.

Schütze, H.: 1992, Dimensions of meaning, *Proceedings of Supercomputing '92*, pp. 787–796.

Seidenberg, M. S. and Gonnerman, L. M.: 2000, Explaining derivational morphology as the convergence of codes, *Trends in Cognitive Sciences* **4**(9), 353–361.

Wittgenstein, L.: 1953, Philosophical investigations.

Zipf, G. K.: 1949, *Human Behavior and the Principle of the Least Effort. An Introduction to Human Ecology*, Hafner, New York.

Modelling Paradigmatic Effects in Visual Word Recognition

CHAPTER 10

An extended version of this chapter will be submitted as Fermín Moscoso del Prado Martín and R. Harald Baayen: BC–DC: A broad-coverage distributed connectionist model of visual word recognition.

Abstract

In this study we describe a distributed connectionist model of visual word recognition. The purpose of this model is to explore how the paradigmatic entropy effects described by Moscoso del Prado Martín, Kostić, and Baayen (2003) can arise in a probabilistic model of lexical processing. We present a model that learns to produce at its output the vectorized semantic representation of a word, the vectorized orthographic representation of which is presented at the input of the model. After training, we compare the outputs of the model with the lexical decision latencies for large sets of English monomorphemic nouns and verbs. Finally, we show that a network with these characteristics exhibits paradigmatic entropy effects similar to those observed by participants in visual lexical decision.

Introduction

Moscoso del Prado Martín, Kostić, and Baayen (2003) introduced a quantitative measure of the support that the recognition of a word receives from the inflectional paradigms to which that word belongs, its inflectional entropy. They define the inflectional entropy of an inflectional paradigm $\mathcal{P}(b) = \{w_1, w_2, \ldots, w_n\}$, were the w_i are the different inflectional variants of the base form b, as:

$$H_i(b) = -\sum_{w_i \in \mathcal{P}(b)} P(w_i|b) \log_2 P(w_i|b). \tag{10.1}$$

In this equation, $P(w_i|b)$ represents the probability of the surface (inflected) form being w_i given that the base form is known to be b. If $F(w_i)$ is the frequency of the inflected form w_i, and $F(b) = \sum_{w_i \in \mathcal{P}(b)} F(w_i)$ is the frequency of the base form b, the sum of the frequencies of all its inflectional variants, then the probability of the surface form w_i occurring, given that the base form is b, is $P(w_i|b) = F(w_i)/F(b)$.

Moscoso del Prado Martín and colleagues showed that the inflectional entropy of a word is a co-predictor of response latencies in visual lexical decision in Dutch. They presented evidence that the information residual of word, defined in terms of the inflectional entropy of a word, its surface frequency, and its derivational entropy (a measure akin to inflectional entropy calculated over derivational paradigms) predicts response latencies better than any combination of surface frequency, base frequency, cumulative root frequency, and morphological family size.

The measures introduced by Moscoso del Prado Martín and colleagues provide a simple way to quantify the support that a word receives from its morphological paradigms. These measures are calculated over a tree-like structure in which inflected forms are linked to their base forms, which, if morphologically complex, are themselves linked to the simpler words from which they are derived. The predictive value for RTs of measures calculated from such tree structures would arise naturally in decompositional models of morphological processing in which a word is processed through probabilistic activation of its constituent stems and affixes.

At first sight, the predictivity of inflectional entropy calculated on the basis of morphological trees might seem problematic for models of morphological processing that do not make use of discrete representations of the stems or base forms of complex words. In distributed connectionist models of lexical processing (Gaskell & Marslen-Wilson, 1997; Plaut & Booth, 2000, Plaut & Gonnerman, 2000; Seidenberg & Gonnerman, 2000), systematic correspondences between similarities in form and similarities in meaning lead to morphologically related words generat-

ing similar patterns of activation, that capture their morphological relations without explicit activation of a shared 'stem' unit.

These models have the advantage of being able to capture various graded effects arising from systematic pairings between form and meaning. Bergen (2003) reports that groups of words that are systematically related in form and meaning, but not morphologically related (e.g., the cluster of English words all relating to LIGHT and starting with the letter sequence 'gl' such as *glitter*, *glow*, *glimmer*, *glisten*, ...) prime each other in way a similar to the priming effects that have been observed for morphologically related words. Bergen shows that this effect is not solely due to orthographic or semantic similarity between the prime-target pairs. Boudelaa and Marslen-Wilson (2001) report a similar effect for Arabic words that share groups of two consonants. However, these groups of two consonants by themselves do not constitute a full morphological unit in the Arabic lexicon in the same sense that the three-consonantal stems do (Bentin & Frost, 2001). These findings suggest that the mental lexicon is sensitive to systematic correspondences between form and meaning, even when these do not come in the form of decomposable units. Non-decompositional theories of lexical processing such as distributed connectionist models and Bybee's network model (Bybee, 1985) are better suited to account for these effects, as they do not depend on the decomposition of a complex word into discrete morphemes.

The question addressed in this study is whether non-decompositional distributed connectionist models can account for the observed predictivity of paradigmatic entropy, a measure that is calculated on the basis of the probability distributions of discrete morphological forms in a hierarchical tree representing the decompositional dependencies between the members of the paradigm.

Another issue that has caused considerable discussion in the psycholinguistic literature is that of past-tense formation. On the one hand, a large group of authors have argued for a dual-route system in which irregulars would be stored in an associative memory system in a non-decompositional fashion, while regulars would be processed by application of a symbolic rule (e.g., Clahsen, 1999; Pinker, 1997, 1999). The proponents of the dual-route processing model base their arguments on differences found in the processing of regular and irregular past-tense forms in behavioural studies (e.g., Clahsen, 1999) and neuro-psychological double dissociations (e.g., Miozzo, in press), and brain-imaging studies (e.g., Ullman, Bergida, & O'Craven, 1997; Indefrey, Brown, Hagoort, Sach, & Seitz, 1997). On the other hand, another group of authors have proposed a single-route ap-

proach, by which both regulars and irregular verbs would be processed by the same basic mechanism (e.g., MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1993; Plunkett & Juola, 1999; Rumelhart & McClelland, 1986).

A factor that has not been given enough consideration in the past-tense debate (although already suggested to play a role by MacWhinney and Leinbach, 1991) is to what extent semantic information might interact with regularity. Ramscar (2002) provided experimental evidence that the semantic context in which a pseudo-verb is presented influences people's choices of the past-tense form for that pseudo-verb. Ramscar noted that this fact is problematic for dual route theories. More recently, Baayen and Moscoso del Prado Martín (2003) have added a new dimension to the debate by showing that regulars and irregulars tend to be clustered in meaning, and crucially, that some of the processing differences between regular and irregular verbs that were found for past-tense forms appear also in the processing of the (completely regular) present-tense forms of those same verbs. This constitutes a challenge for the dual route mechanism, in that according to that theory there should be no such differences. Baayen and Moscoso del Prado Martín also report that one of the main differences between regular and irregular verbs is that, in general, irregular verbs have a higher inflectional entropy than regular verbs.

This raises the question of whether a single-route model of lexical processing might mirror the experimental processing differences observed for the uninflected stems of regular and irregular verbs, as a function of the differences in their meanings.

In what follows, we first describe a distributed connectionist model that was trained to produce the semantic representation of a word from a representation of its orthography. As orthographic representations, we used the Accumulation of Expectations (AoE) vectors for English orthographic forms described by Moscoso del Prado Martín, Schreuder, and Baayen (2003). As a representation of a word's meaning, we used the semantic vectors developed by Moscoso del Prado Martín and Sahlgren (2002), which provide semantic representations for a large set of the inflected forms of English words.

Next, we investigate whether the responses of the model to a large set of words from the Balota, Cortese, and Pilotti (1999) database reflect the pattern of response latencies shown by the participants in that study. We then examine whether the participants in the Balota et al. database show the inflectional and derivational entropy effects observed for Dutch by Moscoso del Prado Martín, Kostić, and Baayen (2003), and whether the network shows similar paradigmatic effects. We then pro-

ceed to investigate whether our model captures the differences in the processing of present-tense forms of regular and irregular verbs. Finally, we address the possible confounds that may arise between paradigmatic entropy measures, and other purely formal measures such as neighborhood size. We conclude by outlining the implications of our model for current theories of visual lexical processing.

Technical Specifications of the Model

Network Architecture

We built a three-layered backpropagation network (Rumelhart, Hinton, & Williams, 1986) whose general architecture is shown in Figure 10.1. The network consisted of 40 orthographic input units, 120 “hidden” units, and 150 semantic output units. The units in the input layer had all-to-all connections to the units in the hidden layer, which themselves had all-to-all connections to the units in the output layer.

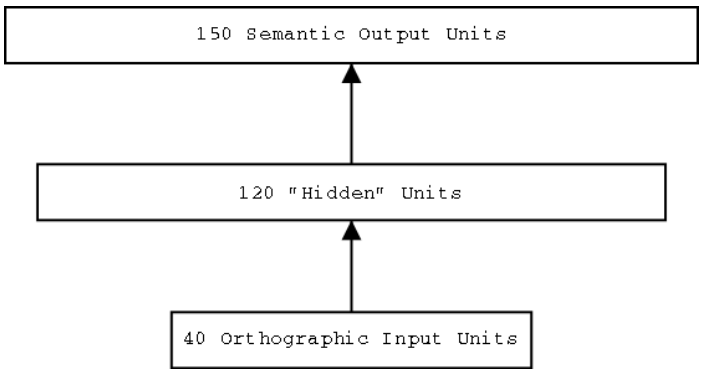

Figure 10.1: General architecture of the model. The lines represent trainable all-to-all connections between the units in two layers.

Training Data

The training set consisted of 48,260 English words. These corresponded to those English words that appear with a frequency higher than 10 in the first 20 million words of the British National corpus and were also listed in the English part of the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995). For each of these words, we constructed orthographic vectors using the AoE technique described by Moscoso del Prado Martín, Schreuder, and Baayen (2003), and we associated them with the semantic vectors for English words described by Moscoso

del Prado Martín and Sahlgren (2002).

Network Training

The network was presented with a word's orthographic vector at its input layer, and was trained to produce the corresponding semantic vector at its output layer. We trained the network with $64 \cdot 10^6$ words that were chosen randomly from the $48,260$ words in the example set, each word being chosen a number of times proportional to its frequency of occurrence in the corpus from which the semantic vectors were built. Training was done by backpropagation, using the modified momentum descent algorithm (Rohde, 1999) with the cosine distance as the error measure. We used a momentum of 0.9 and an initial learning rate of 0.1. Five times during training (each $12.8 \cdot 10^6$ words), the learning rate was divided by two. After training, the network showed an average cosine error of 0.0370 on the words present in the training set.

Results

Once the network had been trained, we evaluated its performance on recognizing a large set of words from the Balota et al. (1999) study. In English, nouns and verbs differ with respect to the number of inflectional variants that they may have. While most English nouns have only two inflected forms (singular and plural), most English verbs have at least three inflectional variants (present-tense, past-tense/past-participle, and gerund), with a maximum of five different forms. This causes the distribution of inflectional entropies of nouns to be different from the distribution of inflectional entropies for verbs, with the latter having a higher inflectional entropy on average. This difference in the entropy distributions might also give rise to differences in the effects of inflectional entropy for nouns and verbs. We take this into consideration by analyzing separately the 1,295 monomorphemic English nouns and 795 monomorphemic English verbs in our dataset.

Nouns

We selected from the Balota et al. dataset those monomorphemic nouns that were also present in our training set. As in many other connectionist studies (e.g., Shillcock, Ellison, & Monaghan, 2000), we use the distance between the model's output

for a word and its correct value as an analog of the reaction time measures for human participants. In other words, we view RTs as reflecting the processing load of mapping a word token in the input stream onto its associated semantic representation, and we are interested in ascertaining whether the cosine distances of our model, which quantify the complexity of this mapping, correlate with the RTs. Both RTs and cosine distances are lognormally distributed as revealed by quantile-quantile plots, and we therefore used their logarithm transform in all analyses. The Pearson correlation between the logarithm of the average reaction time produced by the group of young participants in the Balota et al. database with the logarithm of the cosine distance for the monomorphemic nouns was $0.55 (p < 0.0001)$. This correlation is illustrated in Figure 10.2. Note that the non-parametric regression line indicates that, in general, an increase in the average log reaction time of the young participants, corresponds linearly to the linear increase in the network's log cosine distance.

In order to ascertain that the participants in the English lexical decision study were showing inflectional entropy effects similar to those reported for Dutch by Moscoso del Prado Martín, Kostić, and Baayen (2003), we fitted a linear regression model with the logarithmic average RT for young participants as the dependent variable, and the logarithm of a word's surface frequency and inflectional entropy (calculated according to Equation 10.1) as independent variables. A sequential analysis of variance revealed significant main effects of surface frequency ($F(1, 1293) = 755.396, p < 0.0001$) and inflectional entropy ($F(1, 1293) = 34.78, p < 0.0001$, after having partialled out the effect of surface frequency), with no significant interaction ($F < 1$). The coefficients for the effects of both independent variables were negative ($\beta = -0.0365, t = -26.85, p < 0.0001$ for surface frequency, and $\beta = -0.0262, t = -5.90, p < 0.0001$ for inflectional entropy), indicating that both of these effects were facilitatory: Words with a high frequency or a high inflectional entropy were recognized faster. Introducing base frequency (i.e., the summed frequency of all the inflectional variants of a word) as an additional predictor after partialling out the effects of surface frequency and inflectional entropy resulted in a marginally significant main effect ($F(1, 1288) = 2.80, p < 0.0947$).

We examined whether the network was showing the surface frequency and inflectional entropy effects similar to those observed for the participants by means of a linear model fitting the model's logarithmic cosine distance of a word as a function of the logarithm of that word's frequency of occurrence during training (its surface frequency) and its inflectional entropy (calculated according to Equa-

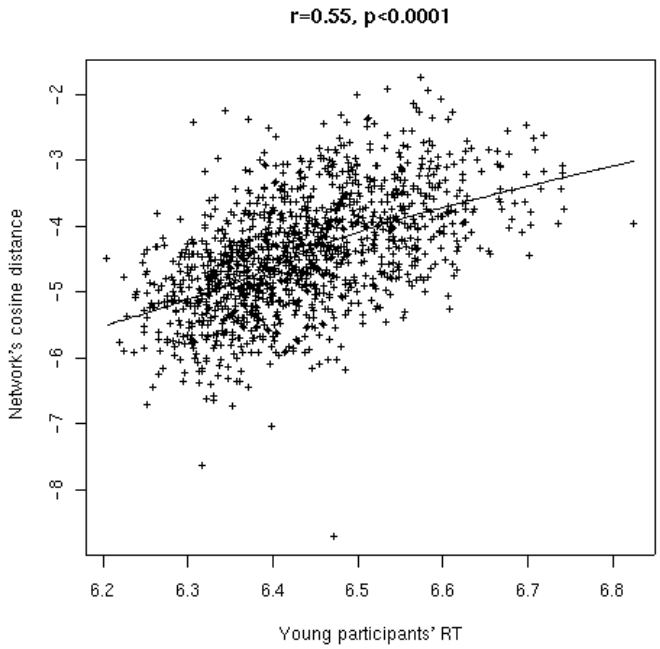

Figure 10.2: Comparison (in bilogarithmic scale) between the young participants' average reaction time to English monomorphemic nouns (horizontal axis) provided by Balota et al. (1999), with the model's cosine distance for those same nouns (vertical axis). The line represents a non-parametric regression (Cleveland, 1979).

tion 10.1 on the basis of only those inflectional variants of a word that appeared in the training set). This revealed significant main effects of surface frequency ($F(1, 1293) = 2958.69, p < 0.0001$) and inflectional entropy ($F(1, 1293) = 54.92, p < 0.0001$, after partialling out the effect of frequency) without any significant interaction ($F < 1$). As in the case of the participants, both of these effects had negative coefficients ($\beta = -0.4390, t = -53.52, p < 0.0001$ for word frequency, and $\beta = -0.1982, t = -7.41, p < 0.0001$ for inflectional entropy), indicating that words with a high frequency or a high inflectional entropy produce smaller cosine distances. Adding base frequency (considering only those inflectional variants of a word that appeared in the training set) into the regression after having partialled out the effects of surface frequency and inflectional entropy did not result in a significant main effect ($F < 1$).

Verbs

We selected from the Balota et al. (1999) dataset those monomorphemic verbs in first person singular form that were also present in our training set. The Pear-

son correlation between the logarithm of the average RT of the young participants and the network's log cosine distance for those verbs was $r = 0.54$ ($p < 0.0001$). Figure 10.3 illustrates this correlation. Again, we observe a roughly linear relation between the participants' log RTs for verbs and the network's log cosine distance for those verbs.

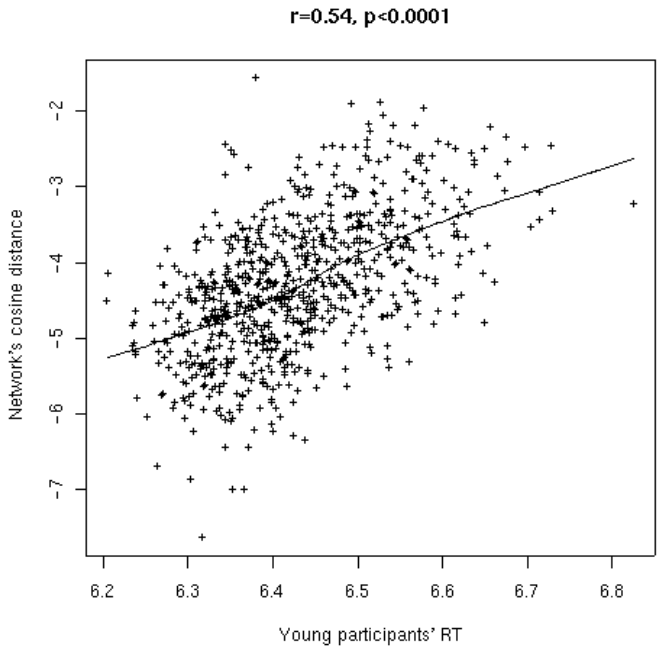

Figure 10.3: Comparison (in bilogarithmic scale) between the Balota et al. (1999) young participants' average reaction time to English monomorphemic verbs (horizontal axis), with the model's cosine distance for those same verbs (vertical axis). The line represents a non-parametric regression (Cleveland, 1979).

A linear regression model fitted to the young participants' average RT for the monomorphemic verbs, with surface frequency and inflectional entropy as the independent variables, revealed significant main effects of surface frequency ($F(1, 793) = 453.73, p < 0.0001$), and inflectional entropy ($F(1, 793) = 47.08, p < 0.0001$ after partialling out the effect of surface frequency), without any significant interaction ($F < 1$). Both main effects had facilitatory coefficients (surface frequency: $\beta = -0.0309, t = -20.09, p < 0.0001$; inflectional entropy: $\beta = -0.0385, t = -6.86, p < 0.0001$). As before, there was no significant main effect of base frequency ($F < 1$) after partialling out the effects of surface frequency and inflectional entropy, and inflectional entropy still showed a significant main effect when we introduced it in the regression after base frequency ($F(1, 792) = 31.67, p < 0.0001$). Again, this confirms for English verbs the effect of inflectional entropy reported by Moscoso del

Prado Martín and colleagues for Dutch words.

We fitted a linear regression model with the same independent variables to the logarithm of the network's cosine distances. The same effects emerged: a main effect of surface frequency ($F(1,793) = 1609.57, p < 0.0001$), a main effect of inflectional entropy ($F(1,793) = 29.64, p < 0.0001$ after partialling out the effect of surface frequency), and only a marginally significant effect of base frequency ($F(1,792) = 2.83, p = 0.0929$) after partialling out surface frequency and inflectional entropy. The marginal effect of base frequency disappears ($F < 1$) when base frequency is included into the regression before inflectional entropy. Inflectional entropy remains significant when entered into the model after base frequency ($F(1,792) = 29.94, p < 0.0001$). As in the case of the reaction times, the effects of surface frequency ($\beta = -0.4082, t = -38.91, p < 0.0001$) and inflectional entropy ($\beta = -0.2084, t = -5.45, p < 0.0001$) were both facilitatory.

Regular and Irregular Verbs

To investigate the issue of the differential processing of the present-tense forms of regular and irregular verbs, we included as additional independent variable the verb's regularity (i.e., *regular* vs. *irregular*).

The analysis of the network's cosine distance revealed a marginally significant main effect of regularity ($F(1,791) = 3.62, p = 0.0574$, after partialling out the effects of surface frequency and inflectional entropy) and a significant frequency by regularity interaction ($F(1,791) = 8.62, p = 0.0034$, after partialling out the main effects). The interaction had a negative coefficient for the regular verbs ($\beta = -0.9, t = -2.93, p = 0.0034$), indicating that, after partialling out the other variables, regular verbs are more sensitive to the frequency effect. A similar pattern emerged in the analysis of the RTs. We did not find a main effect of regularity ($F < 1$, after partialling out frequency and inflectional entropy), and we observed the same frequency by regularity interaction ($F(1,791) = 6.03, p = 0.0143$, after partialling out the main effects) with a negative coefficient ($\beta = -0.01, t = -2.46, p = 0.0143$).

These results confirm those of Baayen and Moscoso del Prado Martín (2003) in that there is indeed and effect of regularity on the response latencies for the present-tense forms of verbs. However, neither in the participants nor in the network can this effect be completely attributed to the difference in inflectional entropy between regulars and irregulars, as shown by the observed frequency by regularity interactions. Recall here that the surface frequency and inflectional entropy counts that were used for the network were exact, in the sense they were calculated on

the actual frequency distributions to which the network was exposed. Therefore, the crucial interaction in the network cannot be attributed to regularity being a mere correction to the other two counts.

Neighborhood Size

At this point, it is necessary to consider a possible confound in our data. Although we are arguing that the inflectional entropy effect that we are observing arises due to the network creating morphological generalizations over form-meaning regularities in its training set, it could be argued that we are only observing an effect of form similarity. Inflectionally related forms, independently of their relationship in meaning, tend to be (by definition) very similar in form. This could lead the model to be affected by the raw number of orthographically similar words in its training set. Such effects have been widely reported in the literature on visual word recognition. Coltheart, Davelaar, Jonasson, and Besner (1977) reported that words with large orthographic neighborhoods are recognized slower in a visual lexical decision task, where lexical neighborhood is defined as the number of words in the lexicon that differ in only one letter from the target. In contrast, the effects of inflectional entropy that we observed both in the network and in the experimental data were facilitatory. Words with large inflectional entropies were recognized faster by the participants and elicited lower distances in the network. However, a number of studies, using word naming tasks, have reported facilitatory effects of orthographic neighborhoods (e.g., Andrews, 1989; 1992). As our network is never actually performing a pure visual lexical decision task, it might be argued that we are observing effects more similar to those found in the word naming paradigm, with our inflectional entropy effect being unrelated to the true effects of inflectional entropy observed for the participants in visual lexical decision. This would imply that our inflectional entropy effect found in the network would be more related to the facilitatory neighborhood size effect in naming, and thus reflect only properties of orthographic form processing.

In order to address this potential confound, we fitted a new linear regression model with the log cosine distance to all nouns and verbs in the previous analyses as the dependent variable, and log surface frequency, inflectional entropy and log neighborhood size (as calculated from the CELEX database) as independent variables. A sequential analysis of variance revealed significant main effects of surface frequency ($F(1, 2001) = 4309.65, p < 0.0001$), inflectional entropy ($F(1, 2001) = 38.20, p < 0.0001$ after partialling out the effect of frequency), and

neighborhood size ($F(1, 2001) = 12.83, p = 0.0005$, after partialling out the effect of frequency and inflectional entropy), and a significant interaction between frequency and neighborhood size ($F(1, 2001) = 6.21, p = 0.0128$, after partialling out the main effects). As in the previous analyses, we found negative, facilitatory, coefficients for both frequency ($\beta = -0.47, t = -30.27, p < 0.0001$) and inflectional entropy ($\beta = -0.13, t = -6.22, p < 0.0001$). The main effect of neighborhood size did not have a significant coefficient ($\beta = -0.06, t = -1.25, p = 0.2100$) and the frequency by neighborhood size interaction had a positive, inhibitory coefficient ($\beta = 0.02, t = 2.49, p = 0.0128$).

A similar analysis with the log RT as dependent variable and the same independent variables as above, revealed significant main effects of frequency ($F(1, 2001) = 1167.79, p < 0.0001$) and inflectional entropy ($F(1, 2001) = 88.88, p < 0.0001$), with the main effect of neighborhood size not reaching significance ($F(1, 2001) = 3.05, p = 0.0809$) and a significant interaction between surface frequency and neighborhood size ($F(1, 2001) = 13.92, p < 0.0002$, after partialling out all the main effects). Of the significant effects, the coefficients of the main effects of frequency ($\beta = -0.04, t = -17.56, p < 0.0001$) and inflectional entropy ($\beta = -0.03, t = -9.28, p < 0.0001$) were facilitatory, while the frequency by neighborhood size interaction was inhibitory ($\beta = 0.005, t = 3.73, p = 0.0002$).

These analyses indicate that inflectional entropy is facilitatory in nature both for the response latencies and for the model's cosine distances. By contrast, neighborhood size emerges as an inhibitory interaction with frequency, again for both the distances and the RTs, in line with the inhibitory neighborhood effects of neighborhood size in visual lexical decision reported by Coltheart et al. (1977). We therefore conclude that our model captures essential aspects of the participants' sensitivity to frequency, form similarity, and inflectional similarity in visual lexical decision. We also conclude that inflectional entropy and neighborhood size effects in our network reflect different aspects of processing, with inflectional entropy capturing the form-meaning correlations present in the inflectional paradigms of the training set, and with neighborhood size reflecting pure form similarities.

Derivational Entropy

After having ascertained that both our model and the participants showed comparable effects of inflectional entropy, we turn to investigate the effects of derivational entropy. Once more, we do this by means of two linear regression models, one with the RTs and the other with the cosine distances as the dependent variable. We

calculated the derivational entropy according to the definition provided by Moscoso del Prado Martín, Kostić, and Baayen (2003), using only those words that appeared in our networks' training corpus.

A linear regression fit to the logarithm of the RT, with log surface frequency, inflectional entropy, and derivational entropy as independent variables, revealed main effects of frequency ($F(1, 2000) = 1165.75, p < 0.0001$), inflectional entropy ($F(1, 2000) = 88.72, p < 0.0001$, after partialling out the effect of frequency), and significant interactions of frequency by derivational entropy ($F(1, 2000) = 6.88, p = 0.0088$, after partialling out the main effects), and of inflectional by derivational entropy ($F(1, 2000) = 5.22, p = 0.0225$). The effect of derivational entropy did not reach significance after partialling out the effects of word frequency and inflectional entropy ($F(1, 2000) = 2.35, p = 0.1255$). The marginal coefficients of the main effects of frequency ($\beta = -0.04, t = -27.54, p < 0.0001$), inflectional entropy ($\beta = -0.03, t = -8.88, p < 0.0001$), and derivational entropy ($\beta = -0.06, t = -3.69, p = 0.0002$), were all negative, while the coefficients of the frequency by derivational entropy interaction ($\beta = 0.01, t = 2.52, p = 0.0119$) and the inflectional by derivational entropy ($\beta = 0.02, t = 2.28, p = 0.0225$) were both positive.

A similar linear regression model with log cosine distance as the dependent variable, and the same independent variables as before, revealed main effects of frequency ($F(1, 2000) = 1083.91, p < 0.0001$), inflectional entropy ($F(1, 2000) = 38.18, p < 0.0001$, after partialling out the effect of frequency), and derivational entropy ($F(1, 2000) = 6.16, p = 0.0132$, after controlling for frequency and inflectional entropy, and significant interactions of frequency by derivational entropy ($F(1, 2000) = 5.49, p = 0.0192$, after partialling out the main effects), and of inflectional by derivational entropy ($F(1, 2000) = 6.89, p = 0.0087$). The marginal coefficients of the main effects of frequency ($\beta = -0.44, t = -52.55, p < 0.0001$) and inflectional entropy ($\beta = -0.09, t = -3.42, p = 0.0007$) were again negative, with the coefficient of the interaction between frequency and inflectional entropy being positive ($\beta = 0.04, t = 4.46, p = 0.0140$), and the interaction between both entropies negative ($\beta = -0.12, t = -2.65, p = 0.0087$). The marginal coefficient for the effect of derivational entropy was negative, but not significant ($\beta = -0.13, t = -1.35, p = 0.1768$).

In both regressions, we found significant interactions of derivational entropy by frequency, and derivational by inflectional entropy. Both for the participants and for the network, the marginal coefficient of the effect of derivational entropy was negative, but not significant in the case of the network. In the case of the participants, this effect does not reach significance in a sequential analysis of variance, after

having partialled out the effects of surface frequency and inflectional entropy. Recall here that the derivational entropy was calculated from the distribution of words in which the model was trained. Therefore, for the model, the derivational entropy is an exact measure, while it is only an approximation of its value for the participants. We think that using a more accurate measure of derivational entropy for the participants should also reveal a significant main effect, similar to that reported for Dutch by Moscoso del Prado Martín and colleagues. An additional problem is the difference in signs between marginal coefficients for the inflectional by derivational entropy interactions in the models of RTs and cosine distances. However, visual examination of the scatterplots for both models revealed that there is a lot of non-linearity in these interactions, making the coefficients for the linear effect unreliable. The emergence of this derivational entropy effect is a clear indication of the form-meaning interactions present in the model. Moscoso del Prado Martín and Sahlgren (2002), indicated that the semantic vectors that we have used in this simulation contained a detailed representation of inflectional relations between words, thus it could be argued that the effect inflectional entropy arises solely due to the semantic neighborhoods. However, Moscoso del Prado Martín and Sahlgren also reported that, using these semantic vectors only, one could not detect many derivational relations. This is a clear sign that the model must be exploiting the correspondences between form and meaning in order for the effect of derivational entropy to arise. However, further research is needed on these issues to clarify the consequences of using semantic vectors that are more sensitive to derivational relations, and more accurate calculations of derivational entropy for the participants.

Age of Acquisition

Another issue that has caused a considerable amount of debate in the literature is the effect of Age of Acquisition (AoA; Carroll & White, 1973). Words that are acquired early in development are recognized faster than words that are acquired later in life, independently of their frequency. Morrison and Ellis (1995) argued that connectionist networks would not be able to show effects of AoA given that they suffer from ‘catastrophic forgetting’, by which patterns that are acquired later in training ‘overwrite’ the representation of patterns that appeared earlier during training. In this respect, Ellis and Lambon Ralph (2000) proved that, when a network is trained on a set of early and a set of late patterns, it does show AoA effects. Smith, Cottrell, and Anderson (2001) reported effects of AoA on a network’s error. Interestingly, instead of manipulating the moment during training at which a pattern was

presented to the network for the first time, they measured the moment in training at which a given pattern is learned, without any manipulations on order of pattern presentation. This finding suggests that the AoA effect might arise from the inherent difficulty of learning (and processing) a particular pattern, independently of any developmental considerations. In simulations using artificial datasets, Anderson and Cottrell (2001) tested the hypothesis that the AoA effect reflects the patterns of similarity between the items in the dataset, that is, words which are similar (in form, meaning, or both) to many others, are easier to learn and faster to process. If the hypothesis put forward by Anderson and Cottrell is true, given that our model is trained on a realistic sample of the English language, it should show AoA effects in a similar way to participants, even though the order of presentation of the words during training is completely arbitrary.

In order to compare the effects of AoA in our network with those shown by human participants, we obtained AoA ratings for 521 words from the MRC Psycholinguistic Database (Coltheart, 1981) and we combined them with the reaction times for young participants to those same items from the Balota et al. (1999) dataset, and the cosine distances obtained by our network for those words.

We fitted a linear regression model to the young participants' average RT, with surface frequency, inflectional entropy, derivational entropy, and AoA as the independent variables. A sequential analysis of variance revealed significant main effects of surface frequency ($F(1, 517) = 241.67, p < 0.0001$), inflectional entropy ($F(1, 517) = 42.85, p < 0.0001$ after partialling out the effect of surface frequency), and AoA ($F(1, 517) = 80.94, p < 0.0001$ after partialling out the effects of surface frequency, and inflectional entropy). Additionally, we observed a significant interaction between frequency and AoA ($F(1, 517) = 7.79, p = 0.0055$). After partialling out the remaining main effects, we did not observe any additional effect of derivational entropy in this dataset ($F < 1$). The main effect of AoA had an inhibitory coefficient ($\beta = 6.128 \cdot 10^{-4}, t = 5.36, p < 0.0001$) while the interaction had a facilitatory coefficient ($\beta = -5.416 \cdot 10^{-5}, t = -2.79, p = 0.0055$).

A linear regression model fitted with the same independent variables to the logarithm of the network's cosine distances revealed the same effects as above: a main effect of surface frequency ($F(1, 517) = 438.12, p < 0.0001$), a main effect of inflectional entropy ($F(1, 517) = 10.96, p = 0.0010$ after partialling out the effect of surface frequency), and a main effect of AoA ($F(1, 517) = 6.88, p = 0.0090$) after partialling out surface frequency and inflectional entropy. Once more, there was a significant frequency by AoA interaction ($F(1, 517) = 56.99, p < 0.0001$) and no

significant effect of derivational entropy ($F < 1$) after partialling out the remaining variables. As in the case of the reaction time analyses, the main effect of AoA had a positive coefficient in the regression ($\beta = 1.326 \cdot 10^{-4}, t = 7.99, p < 0.0001$) and the frequency by AoA interaction had a negative coefficient in the regression ($\beta = -2.128 \cdot 10^{-5}, t = -7.55, p < 0.0001$).

These results support the hypothesis advanced by Anderson and Cottrell (2001) that the AoA effect reflects, at least in part, the position of a word in morpho-semantic space, independently of development of neural plasticity.

General Discussion

In this study, we have presented a broad coverage distributed connectionist model of visual word recognition. The model was trained to map distributed orthographic representations onto distributed semantic representations. After training, we compared the model's cosine distances with the response latencies of participants performing visual lexical decision for large sets of English monomorphemic nouns and verbs. We found that, in both cases, the model produced output patterns that were remarkably similar to the pattern of responses of actual participants.

The model that we have introduced constitutes a considerable departure from previously implemented distributed connectionist models of lexical processing in that it has a much broader coverage and in that it avoids the traditional restrictions on word length and morphological complexity. In this study, we have used a vocabulary of 48,260 different words to train our model. This represents a realistic sample of the lexicon, containing the full range of morphological phenomena present in English. In principle, a model of these characteristics could be exposed to even larger vocabularies, approximating the number of different words to which an average adult is exposed.

The key to this broad coverage lies in the use of truly distributed representations of word forms and meanings, as provided by the AoE representational paradigm (Moscoso del Prado Martín, Schreuder, & Baayen, 2003), and the realistic context-based semantic vectors of Moscoso del Prado Martín and Sahlgren (2002). A corpus-based co-occurrence approach to semantic representation is based on realistic assumption of co-occurrence being one of the sources of information that humans use to determine the meaning of a word (e.g., Boroditsky & Ramscar, 2003; McDonald & Ramscar, 2001). Additionally, it overcomes the bottleneck for realistic models caused by having to rely on hand-crafted semantic representations.

The coding scheme used for word forms has the advantage that it obviates the need for slot-based templates that require manual preprocessing of the words form. The use of slot-based structures for the coding of word forms, and hand-crafted representations to code meaning has been criticized for assuming a great amount of hard-wired symbolic information about orthographic and semantic structure (e.g., Pinker & Ullman, 2002b). Moreover, slot-based representations require arbitrary decisions on alignment for coding the similarities and dissimilarities of onsets, word-centers, and codas, as in the onsets of the Dutch words *sap*, *stap*, and *tap*.

Our model also illustrates how the differences found in the processing of the present-tense forms of regular and irregular verbs arise naturally in a single-route model of lexical processing. The fact that this model was never actually trained on the past-tense formation task confirms the results of Baayen and Moscoso del Prado Martín (2003) in that there are indeed important differences between both the orthographic and semantic properties of regular and irregular verbs. It is not unlikely that these differences underlie many of the double dissociations and processing differences that have been found between these two kinds of verbs. Additionally, it is not clear how the dual-route models could account for the differences in processing the present tense, especially since their proponents explicitly deny any possible influence of verbal semantics in the selection of a verb's past-tense form (e.g., Pinker & Ullman, 2002a).

The response patterns produced by the model account for approximately 30% of the variance in the RTs produced by the human participants. This is remarkable given that many factors that are known to affect visual lexical processing are not taken into account by our system. In particular, as mentioned above, our contextual semantic representations do not fully capture the type of semantic relations that are present in derivational morphology, and therefore our model does not completely mirror the effects caused by derivational paradigms. Additionally, other variables that are known to correlate with visual lexical decision latencies, such as concreteness or imageability, are absent. Such effects can only be captured by a model that also includes sensory-motor information in its semantic representations (cf., Pulvermüller, 2002).

Additionally, our model also mirrored participants' behavior with respect to the Age of Acquisition effect. Our model produced significantly lower error scores for words that are acquired early by people, according to Age of Acquisition norms. Crucially, our network's training regime did not follow any developmental considerations. This supports the proposal of Anderson and Cottrell (2001) that Age of

Acquisition reflects the similarity structure of words in the lexicon, instead of decreases in neural plasticity during development.

With surface frequency, inflectional entropy, derivational entropy, and neighborhood size, we are able to account for approximately two thirds of the variance present in the model's cosine distances. This suggests that further research is necessary to understand the source of the remaining one third of the model's variance. We think that there are more psychologically relevant factors that are captured by the model. In particular, different types of effects that are claimed to arise at form recognition levels such as word length or bigram frequency need to be investigated. We leave these for further research.

Crucially, although the model did not receive any explicit symbolic representation of the morphological relations between the words in its training set, it developed sensitivity to morphological structure, as indicated by the effects of inflectional and derivational entropy that we observed. In particular, the effects of derivational entropy, and the analyses including neighborhood size, showed that the model is sensitive to effects that cannot be attributed to just form or just meaning similarity on its own. Instead, the effects emerge from the systematic form-meaning associations shared by the morphological variants of a word. In conclusion, we have shown that the paradigmatic entropy effects described by Moscoso del Prado Martín, Kostić, and Baayen (2003) do not constitute a problem for distributed connectionist models of lexical processing. In fact, we believe that such effects are a fundamental property of neural processing systems (e.g., Deco & Obradović, 1996).

References

Anderson, K. L. and Cottrell, G. W.: 2001, Age of acquisition in connectionist networks, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum, Mahwah, pp. 27–32.

Andrews, S.: 1989, Frequency and neighborhood size effects on lexical access: Activation or search?, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15**, 802–814.

Andrews, S.: 1992, Frequency and neighborhood size effects on lexical access: Similarity or orthographic redundancy?, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**(2), 234–254.

Baayen, R. H., Dijkstra, T. and Schreuder, R.: 1997, Singulars and plurals in Dutch: Evidence for a parallel dual route model, *Journal of Memory and Language* **37**, 94–117.

Baayen, R. H. and Moscoso del Prado Martín, F.: 2003, Questioning the unquestionable: Semantic density and past-tense formation in three Germanic languages, *Manuscript submitted for publication, Max Planck Institute for Psycholinguistics*.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX lexical database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Balota, D. A., Cortese, M. J. and Pilotti: 1999, Item-level analyses of lexical decission performance: Results from a mega-study, *Abstracts of the 40th Annual Meeting of the Psychonomics Society*, Los Angeles, CA, p. 44.

Bentin, S. and Frost, R.: 2001, Linguistic theory and psychological reality: a reply to Boudelaa and Marslen-Wilson, *Cognition* **81**(1), 113–118.

Bergen, B. K.: 2003, The psychological reality of phonaesthemes, *Manuscript submitted for publication, University of Hawai'i at Manoa*.

Boroditsky, L. and Ramscar, M.: 2003, Guilt by association: Gleaning meaning from contextual co-occurrence, *Manuscript, Massachusetts Institute of Technology*.

Boudelaa, S. and Marslen-Wilson, W. D.: 2001, Morphological units in the Arabic mental lexicon, *Cognition* **81**(1), 65–92.

Bybee, J. L.: 1985, *Morphology: A study of the relation between meaning and form*, Benjamins, Amsterdam.

Carroll, J. B. and White, M. N.: 1973, Word frequency and age of acquisition as

determiners of picture-naming latency, *Quaterly Journal of Experimental Psychology* **25**, 85–95.

Clahsen, H.: 1999, Lexical entries and rules of language: a multi-disciplinary study of German inflection, *Behavioral and Brain Sciences* **22**, 991–1060.

Cleveland, W. S.: 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

Coltheart, M.: 1981, The MRC Psycholinguistic Database, *Quarterly Journal of Experimental Psychology* **33A**, 497–505.

Coltheart, M., Davelaar, E., Jonasson, J. T. and Besner, D.: 1977, Access to the internal lexicon, *in* S. Dornick (ed.), *Attention and performance*, Vol. VI, Erlbaum, Hillsdale, New Jersey, pp. 535–556.

Deco, G. and Obradović, D.: 1996, *An Information-Theoretic Approach to Neural Computing*, Springer Verlag, New York.

Ellis, A. W. and Lambon Ralph, M. A.: 2000, Age of acquisition effects in adult lexical processing reflects loss of plasticity in maturing systems: Insights from connectionist networks, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26**, 1103–1123.

Gaskell, M. G. and Marslen-Wilson, W.: 1997, Integrating form and meaning: A distributed model of speech perception, *Language and Cognitive Processes* **12**, 613–656.

Indefrey, P., Brown, M. C., Hagoort, P., Sach, S. and Seitz, R. J.: 1997, A PET study of cerebral activation patterns induced by verb inflection, *Neuroimage* **5**, 548.

MacWhinney, B. and Leinbach, J.: 1991, Implementations are not conceptualizations: revising the verb learning model, *Cognition* **40**, 121–157.

McDonald, S. and Ramscar, M.: 2001, Testing the distributional hypothesis: The influence of context judgements of semantic similarity, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

Miozzo, M.: in press, On the processing of regular and irregular forms of verbs and nouns: evidence from neuropsychology, *Cognition*.

Morrison, C. M. and Ellis, A. W.: 1995, Roles of word frequency and age of acquisition in word naming and lexical decision, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**, 116–133.

Moscoso del Prado Martín, F., Kostić, A. and Baayen, R. H.: 2003, Putting the bits together: An information theoretical perspective on morphological processing,

Manuscript submitted for publication, Max Planck Institute for Psycholinguistics.

Moscoso del Prado Martín, F. and Sahlgren, M.: 2002, An integration of vector-based semantic analysis and simple recurrent networks for the automatic acquisition of lexical representations from unlabeled corpora, *in* A. Lenci, S. Montemagni and V. Pirrelli (eds), *Proceedings of the LREC'2002 Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*, European Linguistic Resources Association, Paris.

Moscoso del Prado Martín, F., Schreuder, R. and Baayen, R. H.: 2003, Using the structure found in time: Building real-scale orthographic and phonetic representations by Accumulation of Expectations, *Manuscript submitted for publication, Max Planck Institute for Psycholinguistics*.

Pinker, S.: 1997, Words and rules in the human brain, *Nature* **387**, 547–548.

Pinker, S.: 1999, *Words and Rules: The Ingredients of Language*, Weidenfeld and Nicolson, London.

Pinker, S. and Ullman, M.: 2002a, Combination and structure, not gradedness, is the issue: Reply to McClelland and Patterson, *Trends in the Cognitive Sciences* **6**(11), 472–474.

Pinker, S. and Ullman, M.: 2002b, The past and future of the past tense, *Trends in the Cognitive Sciences* **6**(11), 456–462.

Plaut, D. C. and Booth, J. R.: 2000, Individual and developmental differences in semantic priming: Empirical and computational support for a single mechanism account of lexical processing, *Psychological Review* **107**, 786–823.

Plaut, D. C. and Gonnerman, L. M.: 2000, Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing?, *Language and Cognitive Processes* **15**(4/5), 445–485.

Plunkett, K. and Juola, P.: 1999, A connectionist model of English past tense and plural morphology, *Cognitive Science* **23**(4), 463–490.

Plunkett, K. and Marchman, V.: 1993, From rote learning to system building: acquiring verb morphology in children and connectionist nets, *Cognition* **48**, 21–69.

Pulvermüller, F.: 2002, *The Neuroscience of Language*, Cambridge University Press, Cambridge, U.K.

Ramscar, M.: 2002, The role of meaning in inflection: Why the past tense doesn't require a rule, *Cognitive Psychology* **45**, 45–94.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: 1986, Learning internal repre-

sentations by error propagation, *in* D. E. Rumelhart and J. L. McClelland (eds), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, The MIT Press, Cambridge, Mass., pp. 318–364.

Rumelhart, D. E. and McClelland, J. L.: 1986, On learning the past tenses of English verbs, *in* J. L. McClelland and D. E. Rumelhart (eds), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models*, The MIT Press, Cambridge, Mass., pp. 216–271.

Seidenberg, M. S. and Gonnerman, L. M.: 2000, Explaining derivational morphology as the convergence of codes, *Trends in Cognitive Sciences* **4**(9), 353–361.

Shillcock, R., Ellison, T. M. and Monaghan, P.: 2000, Eye-fixation behaviour, lexical storage and visual word recognition in a split processing model, *Psychological Review* **107**, 824–851.

Smith, M. A., Cottrell, G. W. and Anderson, K. L.: 2001, The early word catches the weights, *in* T. K. Leen, T. G. Dietterich and V. Tresp (eds), *Proceedings of NIPS 13*, MIT Press, Cambridge:MA, pp. 52–58.

Ullman, M., Bergida, R. and O'Craven, K. M.: 1997, Distinct fMRI activation patterns for regular and irrgular past tense, *NeuroImage* **5**, 549.

Summary and Conclusions

CHAPTER 11

This dissertation addressed the consequences of morphological structure for lexical processing. Our primary goal was to understand how the systematic relations that exist between the orthographic or phonetic forms of words and their meanings influence the processing and storage of words in the mental lexicon, by focussing on the family size effect in derivational morphology. (The family size effect is the phenomenon that the number of complex words in which a given word occurs as a constituent is a predictor of processing speed.) However, the approach developed in this dissertation provides a parsimonious account of morphological processing in general, including not only derivation and compounding, but also inflection. In what follows, we first provide a summary of the contents and implications of each of the preceding chapters. We continue with two sections, the first of which draws the general conclusions of our studies, while the second one outlines possible lines for future research.

Summary

The dissertation consists of an experimental part and a modeling part. The experimental research is reported in chapters 2–4, and addresses the family size effect across different languages as well as in the bilingual mental lexicon. Chapters 5–10 develop computational models for understanding the morphological aspects of visual word recognition and the role of the family size effect.

In Chapters 2 and 3, we reported three parallel visual lexical decision experiments in Hebrew, Finnish, and Dutch, all three using translation equivalents of the same words in each of the languages. Chapter 2 investigated whether the family size effect is present in Hebrew and, if so, whether it shows similar properties to those shown by the morphological family size in Dutch. Hebrew is a Semitic language with a radically different morphological structure and relatively small mor-

phological families. Nevertheless, a visual lexical decision experiment using Hebrew words documented, for the first time, that the family size effect is also present in Hebrew, and that its properties are similar to those reported for the family size effect in Dutch, English, and German. However, detailed analyses of the results of this experiment revealed some subtle differences between the effects found for Hebrew and those previously reported for Dutch: The direction of the morphological family size effect for Hebrew is strongly modulated by semantic relatedness, to the extreme that semantically distant members of a paradigm inhibit, instead of facilitate, the recognition of targets in that morphological paradigm. This result provides further support for previous experimental evidence suggesting that the family size effect arises at the semantic level of lexical processing (cf., De Jong, 2002).

The experiment reported in Chapter 3 is the first empirical study to reveal a morphological family size effect for Finnish, a Finno-Ugric language. This language is characterized by huge morphological paradigms. One might think that such large paradigms, with thousands of members, would give rise to a ceiling effect for processing latencies. Nevertheless, the family size appeared to predict response latencies in Finnish. However, the extreme productivity of Finnish morphology comes with a new restriction on the family size effect for complex words. For derived words and compounds, only their own morphological descendants (the members of a target's paradigm that are contain that target itself as a constituent) contribute to the family size effect. This contrasts with Dutch and Hebrew, where the full derivational paradigm was responsible for the morphological family size effect. Interestingly, a closer examination of the family size effect found in the parallel Dutch experiment revealed a similar but weaker gradedness of the family size effect. Although all members of a Dutch morphological paradigm participate in the effect, its own descendants have a significantly stronger contribution to the effect than its other family members, which are more semantically distant, but still share a stem with the target. Thus, the difference between Finnish and Dutch emerges as a matter of degree, and not a matter of principle.

All three experiments used translation equivalents of the same words, allowing us to investigate the cross-linguistic predictive power of morphological family size counts. We observed that family size counts for Dutch words predicted the Hebrew visual lexical decision latencies to their translation equivalents. Conversely, family size counts for Hebrew words predicted the response latencies for the corresponding Dutch words. This bi-directional cross-language predictivity was also present between Dutch and Finnish, and persists (for Dutch and Finnish, and for

Dutch and Hebrew as well) even after first partialling out frequency, word length, and family size within a language. No such predictivity, however, was observed between Hebrew and Finnish. This suggests that there is a large degree of overlap in the conceptual organization of the mental lexicon across speakers of languages with not too dissimilar degrees of morphological productivity, such as Hebrew and Dutch, or Dutch and Finnish. The breakdown of the cross-linguistic predictivity of the family size count across Hebrew and Finnish suggests that a large difference in morphological productivity goes hand in hand with a reduction in the degree of overlap in conceptual organization in the mental lexicon.

In Chapter 4 we broadened the scope of our research on the family size effect by addressing its possible role in the Dutch-English bilingual mental lexicon. The bilingual lexicon poses a special challenge for research research on the family size effect as previous studies have shown that the words in the lexicons of both languages may well be integrated in a single lexical store in memory. By investigating the processing of interlingual homographs, words such as *angel* that share the same spelling but that have different meanings and generally different pronunciations, the contribution of the family members of interlingual homographs to the comprehension process can be clarified. A re-analysis of two experiments reported by Schulpen (2003) and an additional visual lexical decision experiment revealed unambiguous effects of family size counts from both languages in the recognition of the interlingual homographs. For participants performing Dutch lexical decision, the Dutch family size count shows a negative correlation with response latencies, while the English family size count shows a positive correlation with RTs. For participants performing English lexical decision, the effects reverse with the English family size count being negatively correlated and the Dutch count having a positive correlation. This result points to the parallel activation of the readings of the homograph and their derivational paradigms in both languages. Crucially, these results hold irrespective of whether the participants performed the experiment in their first language, and of whether they were aware of the presence of words from the other language. Note that these results are consistent with the inhibitory and facilitatory effects of the family size counts for the different semantic fields of homonymic Hebrew roots.

Chapter 5 is the first of the modeling chapters of this dissertation. It introduced an information-theoretical characterization of the effects of morphological paradigms. We quantified the cognitive cost of recognizing a word in terms of the morphological paradigms in which it participates. For this purpose we developed a new measure:

the information residual of a word. This measure summarizes the effects of a large set of variables that had been reported to affect lexical processing, providing a single, parsimonious, account of their effects. The information residual was calculated over pre-defined tree-like symbolic structures that estimate the probabilistic support provided by morphological paradigms to the recognition of their members. We provided reanalyses of three previously published datasets on morphological family size effects, showing that the information residual measure outperforms any combination of traditional measures of morphological complexity. This concise mathematical formulation allows us to analyze in detail the consequences for lexical processing of the internal organization of morphological paradigms. Innovative is that we have succeeded to provide a unified account of both inflectional paradigmatics (the only paradigmatics studied systematically in linguistics) and derivational paradigmatics (the paradigmatics under investigation in the family size research).

Chapters 6-10 report the development of a distributed connectionist model for visual word recognition in which morphology emerges from the interaction of distributed representations of word form (Chapter 6) and word meaning (Chapters 8,9). These representations are integrated in the model presented in Chapter 10.

In Chapter 6, we addressed the problem of creating realistic representations of word form. We developed the Accumulation of Expectations (AoE) technique for representing the orthographic and phonetic forms of a language, using a completely distributed corpus-based approach. This technique was based on the use of the representations that are developed in the hidden layer of a Simple Recurrent Network trained on a letter or phoneme prediction task (Elman, 1990). The AoE technique creates abstract fully distributed vector represenations for word forms, without restrictions on word length and morphological complexity, and without building explicit linguistic structure into the vectors. This method contrasts with current practice in connectionist modeling, in which the representations are restricted with respect to their complexity and which incorporate explicit linguistic structure (such as onset/nucleus/coda and syllabification information). To ensure that such a representational scheme could be used by a distributed connectionist model, we tested the performance of such orthographic vectors on a realistic problem (Chapter 7). We addressed in a single model the problems of Dutch past-tense formation, Dutch final devoicing, and phoneme to grapheme mapping for the complete Dutch verbal system. We showed that a distributed connectionist model developed using AoE representations is able to capture the graded paradigmatic effects that are present in Dutch phonological perception as described by Ernestus

and Baayen (2001; 2003).

Chapters 8 and 9 addressed the development of realistic representations of word meaning. Current practice in the connectionist literature is to construct hypothetical semantic vectors, or to use hand-crafted feature-based semantic vectors for small sets of words. We explored the possibility of creating realistic empirical semantic vectors by means of a Simple Recurrent Network trained on word prediction. We attempted to use the network's hidden layer to capture the semantic information carried by word co-occurrence patterns. Although the technique developed in this chapter captured aspects of word category, valency, and inflection, it did not perform as desired for capturing similarities in word meanings. This was probably due to the model's training regime, word prediction, which turned out to be highly sensitive to morpho-syntactic properties but not so much to lexical meaning. A better technique for creating semantic vectors was developed in Chapter 9, building on the morpho-syntactic vectors developed in Chapter 8, combined with a variant of the Accumulation of Expectations technique used in Chapter 6. By explicitly averaging over several words in a context window, this approach ensures that the semantics carried by lexical co-occurrence is propertly captured. As this technique represents word meanings as vectors in a multi-dimensional space, special techniques were developed to assess similarity (and dissimilarity) of word meanings as well as morphosyntactic properties of words in this multi-dimensional space.

Finally, in Chapter 10, we combined the form and meaning vector representations developed in the previous chapters to build a broad-coverage model of visual word recognition. The model is a simple three-layered backpropagation network (Rumelhart, Hinton, & Williams, 1986) which was trained on producing at its output the semantic vector corresponding to an orthographic vector presented at its input, over a very large vocabulary of 48,000 words. (The largest vocabulary size used in a connectionist model known to us is in the order of 10,000 words.) After training, the distance between the network's output and the corresponding semantic prototype emerged as a good predictor of response latencies in visual lexical decision. Our model shows effects of word frequency, inflectional entropy, derivational entropy, number of orthographic neighbors, and verb regularity, similar to those observed for participants performing visual lexical decision. Crucially, the effects of the inflectional and derivational entropy measures calculated on the basis of symbolic structures (Chapter 5) emerge automatically in this distributed system. The convergence between the symbolic information-theoretical account and the distributed connectionist model indicates that the representation of morphological paradigms

arises naturally from the statistical regularities in the mappings between form and meaning.

Conclusions

The results of the experimental studies on Hebrew, Finnish, Dutch, and the bilingual lexicon reported in Chapters 2–4 highlight the importance of the paradigmatic structure in a language for lexical processing. Crucially, our studies show that the gradedness of the semantic relations between the members of a derivational paradigm shapes the family size effect. The huge productivity of Finnish morphology has as a consequence that a great many family members often have only faint semantic relations to a target word. Thus, as observed for complex target words, speakers of Finnish need not be affected by all family members, but only by those that are in its immediate semantic vicinity. In the smaller derivational paradigms of Dutch, semantic relations tend to be stronger, but even here the strength of a family member's contribution to the family size effect emerged to be modulated by its degree of semantic relatedness to the target. The high semantic heterogeneity of the derivational paradigms of homonymic Hebrew roots revealed a more radical form of modulation by semantic distance, with inhibition instead of facilitation for Hebrew semantically distant family members, just as in the bilingual lexicon the families of the two languages may play an opposite role. Our studies also show that the family size effect is a useful tool for investigating the conceptual similarities across different language families.

We have proposed two complementary models for visual word recognition. The information-theoretical model developed in Chapter 5 provides a direct quantitative estimate of the consequences of linguistic structure for the the processes driving word recognition. It is remarkable that this very simple approach accounts for roughly up to two thirds of the variance in Dutch visual lexical decision latencies. However, despite its advantages of interpretability and predictivity, this approach has two disadvantages that are part and parcel of its dependence on formal symbolic hierarchical representations of morphological constituent structure. The first disadvantage is that such representations do not lend themselves well for representing gradedness in morphological relatedness. The second disadvantage is that the constituent structure is taken for granted and is not accounted for by the model itself. In other words, the model does not provide insight into how such structured representations might emerge.

These two problems do not arise in the distributed connectionist model developed in Chapters 6–10, which is the first real-scale connectionist approximation of visual word recognition. The model represents human lexical processing load as distances in multi-dimensional vector space. It demonstrates the possibility of learning tree-like paradigmatic structures in terms of distributed patterns of activation in multi-dimensional similarity space, without requiring any especialized mechanisms for dealing with compositionality, without special training regimes targeted at compositionality, and without solving the problem by using sufficiently structured input representations. The representations that our model develops are completely corpus-based, and rely only on minimal assumptions about the cognitive system such as the projection of expectations (e.g., Tabor, Juliano, & Tanenhaus, 1996) and sensitivity to word co-occurrence (Boroditsky & Ramscar, 2003; McDonald & Ramscar, 2001). In this approach, morphology emerges 'as a convergence of codes' (Seidenberg & Gonnerman, 2000). Although a backpropagation network clearly does not directly model the way in which information processing occurs in biological neurons, we believe it provides a reasonable approximation of the interactions between distributed patterns of activation representing the meanings and forms of words in the brain. The disadvantage of this model is that it is extremely difficult to understand what factors underlie the distributed representation of a particular word and its associated processing cost. Interestingly, McKay (2003) argues that information-theoretical measures such as those used in Chapter 5 are the adequate tools for describing processing load in neural networks. From this perspective, our information-theoretical model can be viewed as the symbolic interpretation of the subsymbolic network.

Topics for Further Research

There are several lines of research that follow naturally from our results:

Physiological constraints: Both the form and the semantic representation techniques that have been developed in this dissertation are in clear need of further refinement. On the one hand, the technique for orthographic representations needs to include a greater level of physiological detail such as the vertical split in the human fovea (e.g., Shillcock, Ellison, & Monaghan, 2000). On the other hand, the technique for semantic representations also needs to include physiological considerations, such as the sensory-motor information known to co-determine semantic processing (see, e.g., Pulvermüller, 2002).

In principle, this could be done by introducing sensory-motor features for selected seed words in the network that creates the semantic representations. This might lead to bootstrapping during the development of contextual semantic information, and might account for effects such as those of concreteness and imageability. Finally, the learning and information processing mechanisms of the neural network itself should be modified to reflect how learning and processing occurs in real neurons (see again Pulvermüller, 2002).

Developmental considerations: The modelling of cognitive development has a long-standing tradition in distributed connectionism (e.g., Elman et al., 1996). Developmental considerations need to be included in both our models. For the distributed connectionist model, a straightforward first step is to train the model on an evolving corpus of child language. Additionally, one could incorporate limitations on memory and representational capacity in the early stages of training, as suggested by Elman (1993). In the information-theoretical model the solution is less clear, but would probably involve creating a separate model for each stage of development, considering only the vocabulary and word frequency distributions at that particular stage.

From isolated words to words in context: Both the distributed connectionist approach and the information-theoretical model that we have developed need to be further constrained by context in order to account for context-sensitive paradigmatic effects in morphological processing (e.g., De Jong, 2002; Kostić, 1995, 2003). A way to achieve this in a distributed framework is to exchange a feed-forward network for a recurrent network. In the symbolic entropy-based approach, this can be done by using an estimation of the information content of the contexts themselves (McDonald & Shillcock, 2001), and by making use of context sensitive restrictions on paradigm membership (Kostić, 2003).

Cross-language and cross-modality issues: The cross-linguistic differences in paradigmatic effects that were described in the experimental part of this dissertation need to be accounted for by both models. A first step will be to actually build the corresponding equivalent models for Hebrew and Finnish, and to compare the predictions of these models for those languages with the experimental results. Another line of research is to extend this technique to account for both visual and auditory word recognition in the same system. This might be accomplished by adding a phonological input layer to the model, connected directly to the model's hidden layer. The real challenge here is to

move from pre-coded phonological and phonetic representations to digitized speech.

References

Boroditsky, L. and Ramscar, M.: 2003, Guilt by association: Gleaning meaning from contextual co-occurrence, *Manuscript, Massachusetts Institute of Technology*.

De Jong, N. H.: 2002, *Morphological Families in the Mental Lexicon*, MPI Series in Psycholinguistics, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

Elman, J. L.: 1990, Finding structure in time, *Cognitive Science* **14**, 179–211.

Elman, J. L.: 1993, Learning and development in neural networks: The importance of starting small, *Cognition* **48**, 71–99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D. and Plunkett, K.: 1996, *Rethinking Innateness: A Connectionist Perspective on Development*, The MIT Press/Bradford Books, Cambridge, MA.

Ernestus, M. and Baayen, H.: 2001, Choosing between the Dutch past-tense suffixes *-te* and *-de*, *in* T. van der Wouden and H. de Hoop (eds), *Linguistics in the Netherlands 2001*, Benjamins, Amsterdam, pp. 81–93.

Ernestus, M. and Baayen, R. H.: 2003, Predicting the unpredictable: Interpreting neutralized segments in Dutch, *Language* **79(1)**, 5–38.

Kostić, A.: 1995, Informational load constraints on processing inflected morphology, *in* L. B. Feldman (ed.), *Morphological Aspects of Language Processing*, Lawrence Erlbaum Inc. Publishers, New Jersey.

Kostić, A.: 2003, The effects of the amount of information on processing of inflected morphology, *Manuscript submitted for publication, University of Belgrade*.

McDonald, S. and Ramscar, M.: 2001, Testing the distributional hypothesis: The influence of context judgements of semantic similarity, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

McDonald, S. and Shillcock, R.: 2001, Rethinking the word frequency effect: The neglected role of distributional information in lexical processing, *Language and Speech* **44**, 295–323.

McKay, D. J.: 2003, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge, U.K.

Pulvermüller, F.: 2002, *The Neuroscience of Language*, Cambridge University Press, Cambridge, U.K.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: 1986, Learning internal representations by error propagation, *in* D. E. Rumelhart and J. L. McClelland (eds),

Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations, The MIT Press, Cambridge, Mass., pp. 318–364.

Schulpen, B. J. H.: 2003, *Explorations in bilingual word recognition: Cross-modal, cross-sectional, and cross-language effects*, PhD thesis, University of Nijmegen, Nijmegen, The Netherlands.

Seidenberg, M. S. and Gonnerman, L. M.: 2000, Explaining derivational morphology as the convergence of codes, *Trends in Cognitive Sciences* **4**(9), 353–361.

Shillcock, R., Ellison, T. M. and Monaghan, P.: 2000, Eye-fixation behaviour, lexical storage and visual word recognition in a split processing model, *Psychological Review* **107**, 824–851.

Tabor, W., Juliano, C. and Tanenhaus, M. K.: 1997, Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing, *Language and Cognitive Processes* **12**(2/3), 211–271.

Samenvatting en Conclusies

Dit proefschrift beschrijft onderzoek naar de consequenties van morfologische structuur voor lexicale verwerking. Het primaire doel van dit onderzoek was inzicht te verkrijgen in hoe de systematische relaties tussen (orthografische en fonologische) vorm van woorden enerzijds en hun betekenis anderzijds de verwerking en opslag van woorden in het mentale lexicon beïnvloeden. De focus van dit onderzoek was het familiegrootte-effect in de derivationele morfologie. (Het familiegrootte-effect is het verschijnsel dat woorden die in veel andere woorden als constituent voorkomen sneller worden verwerkt.) Echter, het hier beschreven onderzoek laat zien hoe morfologische verwerking in het algemeen verantwoord kan worden, zowel de derivationele morfologie als de inflectionele morfologie. In wat volgt vatten we eerst de voorafgaande hoofdstukken samen, waarna we een overzicht geven van de conclusies die we op grond van de verkregen resultaten hebben getrokken.

Samenvatting

Dit proefschrift doet in hoofdstuk 2–4 verslag van een reeks experimenten die het familiegrootte-effect in verschillende talen, alsmede in het Nederlands-Engelse bilinguale lexicon onderzoekt. De hoofdstukken 5–10 beschrijven de resultaten van computationele studies die tot doel hadden de rol van morfologie in visuele lexicale verwerking te preciseren, en in het bijzonder de rol van de morfologische familiegrootte.

De hoofdstukken 2 en 3 rapporteren drie parallel geconstrueerde experimenten die gebruik maken van de lexicale decisietaak. Deze drie experimenten werden uitgevoerd in het Hebreeuws, in het Fins, en in het Nederlands, en maakten gebruik van woorden die elkaars vertaalequivalenten waren. Hoofdstuk 2 is de eerste studie naar het familiegrootte-effect in het Hebreeuws. Het Hebreeuws is een semitische taal, met een radikaal ander soort morfologie dan het Nederlands. Ondanks de relatief kleine morfologische families in deze taal bleek het familiegrootte-effect

zich ook in het Hebreeuws voor te doen, in grote lijnen vergelijkbaar met het Nederlands, het Engels, en het Duits. Een opmerkelijk verschil dat het familiegrootte-effect in het Hebreeuws karakteriseert is de sterke rol van betekenisverwantschap. Voor Hebreeuwse woorden met een homonieme wortel bleek dat de familiegrootte van de semantisch sterk gelijkende familieleden de verwerking versnelde, net als in het Nederlands en het Engels, terwijl de semantisch niet gelijkende familieleden de verwerking juist vertraagden. Dit resultaat bevestigt de eerder door De Jong (2002) verdedigde interpretatie van het familiegrootte-effect als een semantisch effect.

Hoofdstuk 3 is de eerste studie die het familiegrootte-effect voor een tweede niet indo-europesche taal attesteert, het Fins, een finno-ugrische taal. Het Fins kenmerkt zich door buitengewoon grote morfologische families. Men zou kunnen denken dat families met duizenden leden tot een plafond-effect zouden leiden, maar dit bleek niet het geval. Ook in het Fins bleek de familiegrootte de snelheid van de reactietijden mede te bepalen. Tegelijk kon worden vastgesteld dat het Finse familiegrootte-effect ook een eigen karakteristieke eigenschap heeft. De verwerking van een geleed woord in het Fins blijkt alleen beïinvloed te worden door de familieleden die van dit woord zelf afstammen, dat wil zeggen, door de woorden die dit gelede woord als constituent bevatten. Woorden die slechts een deel van de morfologische structuur delen dragen in het Fins niet aan het familiegrootte-effect bij. Bij nader onderzoek van het parallelle nederlandse experiment bleek daar een mildere vorm van het zelfde verschijnsel op te treden. Hoewel in het Nederlands alle familieleden aan het effect bijdragen, nemen de directe afstammelingen van een geleed woord het leeuwendeel van het effect voor hun rekening. Het verschil tussen het Nederlands en het Fins kenmerkt zich hiermee als een gradueel verschil.

Door gebruik te maken van woorden die elkaars vertaalequivalenten waren, konden we ook nagaan in hoeverre de morfologische familiegrootte in de ene taal voorspellende waarde heeft voor de reactietijden op de vertaalequivalenten in een andere, typologisch niet verwante taal. Het bleek dat de Hebreeuwse familiegrootte een goede voorspeller is van reactietijden op de Nederlandse vertaalequivalenten. Omgekeerd bleek de Nederlandse familiegrootte een goede voorspeller van de reactietijden op de Hebreeuwse vertaalequivalenten. Een vergelijkbare wederzijdse voorspelbaarheid kon vastgesteld worden voor het Fins en het Nederlands. De voorspellende waarde van de familiegrootte kon zelfs vastgesteld worden nadat woordfrequentie, woordlengte, en de familiegrootte in de taal van het experiment waren uitgepartialiseerd. Interessant is dat deze predictiviteit zich niet voor-

doet tussen het Fins en het Hebreeuws. De morfologische productiviteit is in deze talen blijkbaar zo verschillend dat de conceptuele organisatie in het Finse en Hebreeuwse mentale lexicon teveel verschillen om nog van de ene taal naar de andere te kunnen voorspellen.

Hoofdstuk 4 verbreedt de empirische evidentie voor het familiegrootte-effect naar het nederlands-engelse bilinguale lexicon. Het bilinguale lexicon vormt voor het onderzoek naar het familiegrootte-effect een bijzondere uitdaging omdat uit eerder onderzoek is gebleken dat de woorden van beide talen vergaand in het bilinguale mentale lexicon zijn geïntegreerd. Door interlinguale homografen te onderzoeken, woorden zoals *angel* die in het Engels en het Nederlands hetzelfde worden geschreven maar een verschillende betekenis hebben, hebben we kunnen vaststellen dat zowel de Nederlandse als de Engelse morfologische families gelijktijdig in het bilinguale mentale lexicon worden geactiveerd. Als de voertaal van het experiment het Nederlands is, leidt de nederlandse familiegrootte tot kortere reactietijden en de engelse familiegrootte tot langere reactietijden. Is het Engels de voertaal, dan is de nederlandse familiegrootte juist inhiberend en de engelse familiegrootte faciliterend. Dit alles gebeurt zonder dat proefpersonen zich bewust zijn van de aanwezigheid van interlinguale homografen in het experiment. Deze resulaten zijn consistent met de familiegrootte-effecten voor Hebreeuwse ’homografe’ wortels: een faciliterend effect voor semantisch verwante woorden, en een inhiberend effect voor de woorden uit andere betekenisvelden.

Hoofdstuk 5 doet verslag van de eerste computationele studie in dit proefschrift. Het beschrijft een informatie-theoretische quantificatie van het effect van de complexiteit van morfologische paradigma’s op de verwerking van de woorden in deze paradigma’s. Een nieuwe maat, het informatie-residu, uitgedrukt in bits, bleek de reactietijden in een serie visuele lexicale decisie-experimenten even goed en soms zelfs beter te voorspellen dan een reeks gecombineerde traditionele maten. Het informatie-residu wordt berekend over de distributies van het probabilistische gewicht van de deelbomen in de boomstructuur die het morfologische paradigma van een woord karakteriseert. Met behulp van het informatie-residu wordt een accurate gedetailleerde mathematische beschrijving gegeven van de consequenties van de interne structuur van het morfologische paradigma voor de lexicale verwerking. Innovatief is dat we met deze nieuwe maat erin geslaagd zijn om de inflectionele paradigmatiek (de enige paradigmatiek die in de taalkunde systematisch wordt onderzocht) samen met de derivationele paradigmatiek (de paradigmatiek waar het onderzoek naar het familiegrootte-effect zich op richt) op dezelfde principiële

manier te analyseren en te verantwoorden.

De hoofdstukken 6–10 beschrijven de ontwikkeling van een gedistribueerd connectionistisch model voor de visuele verwerking van gelede (en ongelede) woorden. Het doel van deze studies was na te gaan of het mogelijk is een realistisch model te ontwikkelen dat de effecten van morfologie op de verwerking verantwoord als het resultaat van de interactie van vorm en betekenis — zonder specifiek morfologische concepten in het model in te bouwen. De hoofdstukken 6 en 7 beschijven de ontwikkeling van de vorm representaties, de hoofdstukken 8 en 9 rapporteren de ontwikkeling van de betekenisrepresentaties. Hoofdstuk 10 brengt de vormrepresentaties en de betekenisrepresentaties samen in het eigenlijke model.

Hoofdstuk 6 introduceert de ’Accumulation of Expectations’ (AoE) techniek voor de representatie van de orthografische en fonologische vorm van woorden. Deze techniek is gebaseerd op de gedistribueerde representaties die ontstaan in de ’hidden layer’ van een eenvoudig recurrent netwerk getraind op het voorspellen van de volgende letter of het volgende foneem. Met deze techniek worden abstracte, volledig gedistribueerde vector-representaties voor woordvormen verkregen. Cruciaal is dat deze representaties, anders dan gebruikelijk in de connectionistische literatuur, niet gebonden zijn door restricties op woordlengte, noch door restricties op morfologische complexiteit, en dat er geen linguistische structuur (informatie over onset, nucleus, en coda, bijvoorbeeld) a priori in deze representaties is ingebracht.

Hoofdstuk 7 toetst de bruikbaarheid van de AoE techniek aan de hand van empirisch data. Met één en hetzelfde model, getraind op het afbeelden van fonemen op grafemen, blijkt het mogelijk om de vorming van de verleden tijd van werkwoorden in het Nederlands, inclusief de verstemlozing van finale obstruenten, te verantwoorden. De voorspellingen van dit eenvoudige model zijn vergelijkbaar, zowel quantitatief als qualitatief, met de voorspellingen van de modellen en de keuzes van proefpersonen in de studies van Ernestus en Baayen (2001; 2003).

De hoofdstukken 8 en 9 beschrijven de zoektocht naar adequate realistische semantische representaties. De connectionistische literatuur biedt hier weinig houvast: ofwel men gebruikt hypothetische semantische vectoren, ofwel men neemt zijn toevlucht tot een kleine verzameling handmatig samengestelde vectoren. In Hoofdstuk 8, beschrijven wij een eerste poging om empirische semantische vectoren te verkrijgen met behulp van een eenvoudig recurrent netwerk getraind op het voorspellen van het volgende woord in lopende tekst. De idee was dat de

’hidden layer’ een bruikbare representatie van woordbetekenissen zou afleiden uit hoe woorden in elkaars omgeving voorkomen. Hoewel de aldus verkregen representaties wel informatie hadden geëxtraheerd over woordsoort, valentie, en inflectie, bleken zij nauwelijks woordbetekenis te representeren. Dit is vermoedelijk het gevolg van het gevolgde trainingsregime, dat tezeer gericht is op het voorspellen van het volgende woord. De informatie die deze taak optimaliseert is woordsoort, inflectie, en valentie, maar niet woordbetekenis.

Een betere techniek is te vinden in hoofdstuk 9. Deze techniek past een variant van de AoE techniek van hoofdstuk 6 toe op de in hoofdstuk 8 ontwikkelde representaties. Door te middelen over een aantal woorden in een contextvenster lukt het om daadwerkelijk de contextueel bemiddelde aspecten van woordbetekenis te representeren. Aangezien het hier gaat om abstracte representaties in een multidimensionele vectorruimte, besteedt hoofdtuk 9 ruim aandacht aan het ontwikkelen van technieken om verschillen en overeenkomsten in betekenis tussen woorden te traceren.

Hoofdstuk 10, ten slotte, brengt de vormrepresentaties en de betekenisrepresentaties samen in een model voor visuele verwerking van gelede (en ongelede) woorden. Het model is een ’backpropagation’ netwerk met drie lagen (Rumelhart, Hinton, & Williams, 1986) dat werd getraind om als output de semantische vector te produceren bij aanbieding van de vormvector. Het model werd getraind op 48.000 verschillende woordvormen, alle woordvormen met een frequentie groter dan 5 die in een corpus van 20 miljoen woorden voorkomen. Het gaat hier dus om een model met een realistische empirische basis, en niet om een op een specifieke taak en op een specifieke verzameling woorden getraind model. Na training bleek de afstand tussen het semantische prototype en de door het model geproduceerde semantische vector een goede voorspeller te zijn van reactietijden in visuele lexicale decisie. De afstanden in het model bleken op dezelfde manier afhankelijk van variabelen als woordfrequentie, woordlengte, het aantal orthografische buren, inflectionele regelmaat, inflectionele entropie, en derivationele entropie als het geval is voor reactietijden. Uit de gevoeligheid van het model voor de inflectionele en derivationele entropie (beiden componenten van het informatie-residu) blijkt dat morfologische paradigmatiek in het netwerk is gerepresenteerd, zij het impliciet. Dit resultaat ondersteunt onze uitgangshypothese dat morfologische structuur ontstaat uit de statistische regelmatigheden in de afbeelding van vorm op betekenis.

Conclusies

De resultaten verkregen in het experimentele gedeelte van dit proefschrift wijzen op het belang van de paradigmatische structuren in het mentale lexicon voor het woordherkenningsproces. In de derivationele paradigma's speelt met name van semantische overeenkomst tussen de woorden in het paradigma een cruciale rol. De grote productiviteit van de finse morfologie heeft tot gevolg dat de semantische verwantschap tussen woorden in het paradigma zo zwak kan worden dat ze niet meetbaar aan het familiegrootte-effect bijdragen. In de kleinere derivationele paradigma's van het Nederlands is er in zulke gevallen sprake van een duidelijk geringere bijdrage aan het effect. De hoge graad van semantische heterogeniteit in de derivationele paradigma's van Hebreeuwse homonieme wortels leidt tot een omslag in het familiegrootte-effect: terwijl semantisch verwante woorden leiden tot facilitatie, leiden de familieleden in andere betekenisvelden tot inhibitie. Een soortgelijk verschijnsel doet zich voor bij interlinguale homografen in het bilinguale lexicon, met facilitatie voor de familie in de voertaal van het experiment, en inhibitie van de familie in de andere taal. Tenslotte is het familiegrootte-effect een uitstekend instrument gebleken om de mate van overeenkomst in conceptuele organisatie in genetisch niet verwante talen in kaart te brengen.

In het modelleergedeelte van dit proefschrift zijn twee complementaire modellen voor visuele woordherkenning ontwikkeld, een symbolisch, informatie-theoretisch model, en een gedistribueerd connectionistisch model. Het informatie-theoretische model biedt een quantitatieve schatter, het informatie-residu, voor de gevolgen van paradigmatische complexiteit voor de verwerking. Opvallend is dat deze eenvoudige benadering het mogelijk maakt om grofweg tweederde van de variantie in een aantal lexicale decisie-experimenten te verklaren. Het grote voordeel van dit model is de interpreteerbaarheid van het informatie-residu. Dit voordeel gaat echter gepaard met twee nadelen die gegeven zijn met de symbolische aard van de constituent-representaties waarover het informatie-residu wordt berekend. Standaard constituent-representaties lenen zich niet goed voor de representatie van de subtiele gradaties in geleedheid die zo kenmerkend zijn voor derivationele morfologie. Een tweede nadeel is dat het model uitgaat van constituent-representaties, en geen inzicht verschaft in hoe dergelijke representaties zouden kunnen ontstaan.

Deze nadelen kleven niet aan het connectionistische model ontwikkeld in de laatste hoofdstukken van dit proefschrift. Dit model is het eerste connectionistische model dat getraind is op een realistische hoeveelheid data (20 miljoen woorden). Het model quantificeert de menselijke woordverwerkingskosten met behulp van af-

standen in een semantische multi-dimensionele vectorruimte. Het model laat zien dat een netwerk in staat is paradigmatische structuren in de taal te representeren in deze multidimensionele vectorruimte, zonder dat daarvoor een specifiek trainingsregime toegesneden op morfologische compositionaliteit nodig is, en zonder morfologische kennis in de inputrepresentaties in te brengen. De representaties die in dit model ontwikkeld worden zijn volledig corpus-gebaseerd, en veronderstellen slechts dat het cognitieve systeem verwachtingen opbouwt over wat gaat komen (vergelijk Tabor, Juliano, & Tanenhaus, 1996), en dat het registreert hoe woorden in elkaars omgeving voorkomen (Boroditsky & Ramscar, 2003; McDonald & Ramscar, 2001). In deze benadering verschijnt morfologie uit de interactie van vorm en betekenis (Seidenberg & Gonnerman, 2000). Al is het duidelijk dat een backpropagation netwerk geen goede voorstelling geeft van hoe informatie verwerkt wordt door feitelijke neuronen, toch biedt ons model een redelijke benadering, nog altijd abstract en vereenvoudigd, maar wel gebruik makend van mechanismen die dichter staan bij de architectuur van het brein, van hoe gedistribueerde activatiepatronen vorm en betekenis zouden kunnen representeren in de hersenen. Maar ook aan dit model kleeft een nadeel, het nadeel dat het buitengewoon lastig is te achterhalen door wat voor factoren een gedistribueerde semantische representatie en de geassocieerde verwerkingskosten worden bepaald. Interessant is dat McKay (2003) betoogt dat informatie-theoretische maten zoals het informatie-residu bij uitstek geeigend zijn om de verwerkingskosten in neurale netwerken te beschrijven. Vanuit deze optiek is ons informatie-theoretische model te beschouwen als de symbolische interpretatie van het subsymbolische netwerkmodel.

References

Boroditsky, L. and Ramscar, M.: 2003, Guilt by association: Gleaning meaning from contextual co-occurrence, *Manuscript, Massachusetts Institute of Technology*.

De Jong, N. H.: 2002, *Morphological Families in the Mental Lexicon*, MPI Series in Psycholinguistics, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

Elman, J. L.: 1990, Finding structure in time, *Cognitive Science* **14**, 179–211.

Ernestus, M. and Baayen, H.: 2001, Choosing between the Dutch past-tense suffixes *-te* and *-de*, *in* T. van der Wouden and H. de Hoop (eds), *Linguistics in the Netherlands 2001*, Benjamins, Amsterdam, pp. 81–93.

Ernestus, M. and Baayen, R. H.: 2003, Predicting the unpredictable: Interpreting neutralized segments in Dutch, *Language* **79(1)**, 5–38.

McDonald, S. and Ramscar, M.: 2001, Testing the distributional hypothesis: The influence of context judgements of semantic similarity, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

McKay, D. J.: 2003, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge, U.K.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: 1986, Learning internal representations by error propagation, *in* D. E. Rumelhart and J. L. McClelland (eds), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, The MIT Press, Cambridge, Mass., pp. 318–364.

Seidenberg, M. S. and Gonnerman, L. M.: 2000, Explaining derivational morphology as the convergence of codes, *Trends in Cognitive Sciences* **4**(9), 353–361.

Tabor, W., Juliano, C. and Tanenhaus, M. K.: 1997, Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing, *Language and Cognitive Processes* **12**(2/3), 211–271.

Resumen y Conclusiones

En esta tesis investigamos las consecuencias de la estructura morfológica de un lenguaje para el procesamiento léxico. Nuestro principal objetivo es detallar la influencia que tienen las sistematicidades existentes entre la forma (ortográfica o fonética) de una palabra y su significado, en su procesamiento y almacenamiento en el léxico mental, prestando especial atención al efecto de tamaño de familia en la morfología derivacional. Aunque nos hemos concentrado especialmente en la morfología derivacional, la perspectiva ofrecida en esta tesis ofrece también un tratamiento uniforme del procesamiento morfológico en general, incluyendo no sólo la morfología derivacional y la de las palabras compuestas, si no también la morfología inflexional.

Resumen

La tesis se divide en dos partes: una experimental y otra de modelado computacional. Del Capítulo 2 al Capítulo 4 presentamos las investigaciones experimentales, que conciernen al efecto de tamaño de familia morfológica a través de varias lenguas y en el léxico mental de los bilingües. Los capítulos del 5 al 10 desarrollan modelos computacionales para comprender los aspectos morfológicos del reconocimiento visual de palabras y el papel del efecto de tamaño de familia morfológica.

En los capítulos 2 y 3, describimos tres experimentos de decisión léxica visual en hebreo, finés, y neerlandés, empleando traducciones de las mismas palabras a cada lengua. El Capítulo 2 investiga si el efecto de tamaño de familia aparece en hebreo y, dado el caso, si muestra propiedades similares a las previamente descritas en neerlandés. El experimento de decisión léxica en hebreo demuestra por vez primera dicho efecto en hebreo, con propiedades muy similares a las descritas en inglés, neerlandés, o alemán. Sin embargo, un análisis más detallado de los resultados experimentales revela algunas diferencias sutiles entre el

efecto que describimos en hebreo, y los efectos descritos con anterioridad en neerlandés y otras lenguas: La similitud semántica modula fuertemente la dirección del efecto de tamaño de familia en hebreo, hasta el extremo de que los miembros de un paradigma morfológico cuya relación semántica con una palabra dada es muy débil, en lugar de facilitar el reconocimiento de dicha palabra, lo inhiben. Esto constituye evidencia adicional para situar el efecto de tamaño de familia en el nivel semántico del procesamiento léxico, de acuerdo con los datos presentados por De Jong (2002).

El experimento descrito en el Capítulo 3 constituye el primer estudio empírico que documenta efectos de tamaño de familia en lengua finesa. Aparentemente, las enormes familias morfológicas del finés no dan lugar a un “efecto techo”. Sin embargo, la extremada productividad de la morfología finesa añade una restricción adicional al efecto de tamaño de familia en palabras morfológicamente complejas: En el caso de palabras derivadas y compuestas, tan sólo contribuyen al efecto los descendientes morfológicos de una palabra. Por el contrario, tanto en hebreo como en neerlandés, el paradigma derivacional completo de la raíz de una palabra contribuye al efecto, incluyendo también aquellos miembros que no se derivan de la palabra dada. Un examen más detallado del efecto de tamaño de familia en el experimento paralelo en neerlandés reveló que, de hecho, en neerlandés también se diferencian las contribuciones al efecto de las palabras derivadas, y las contribuciones de palabras que, aun compartiendo la raíz, no se derivan de la palabra. Por tanto, la diferencia que observamos entre el finés y el neerlandés es más una diferencia de magnitud que una de principios.

El hecho de que los tres experimentos usaran traducciones de las mismas palabras nos permite investigar la predictividad croslingüística del tamaño de familia mofológica. Mediante regresiones multinivel observamos que el tamaño de familia morfológica de una palabra en neerlandés predice las latencias de respuesta a su traducción al hebreo, del mismo modo en que la familia morfológica de una palabra en hebreo predice los tiempos de respuesta a su traducción al neerlandés. Esta predictividad croslingüística bidireccional también aparece entre palabras neerlandesas y finesas, y persiste (tanto entre neerlandés y finés como entre neerlandés y hebreo) incluso después de haber descontado los efectos de frecuencia, longitud de palabra, y tamaño de familia dentro de un mismo idioma. Sin embargo, no observamos predictividad alguna entre tamaño de familia y latencias de respuesta entre finés y hebreo, lo cual apunta a la existencia de un elevado nivel de solapamiento entre las organizaciones conceptuales del léxico mental de los

hablantes de lenguas cuyos niveles de productividad morfológica no se diferencian en exceso, como en el caso del hebreo y el neerlandés. La desaparición de la predictividad croslingüística del tamaño de familia morfológica entre el hebreo y el finés sugiere que las grandes diferencias en productividad morfológica entre dos lenguas van directamente asociadas a una reducción en el nivel de solapamiento en las representaciones conceptuales en el léxico mental de sus hablantes.

En el Capítulo 4 ampliamos el ámbito de nuestras investigaciones sobre del efecto de tamaño de familia morfológica, investigando el papel que puede jugar en el léxico mental de bilingües neerlandés–inglés. En el léxico mental bilingüe surge la pregunta de cómo contribuyen los tamaños de familia morfológica de un homógrafo interlingüe al proceso de comprensión visual. Aunque los homógrafos interlingües presentan idénticas ortografías en dos lenguas, sus significados y, en mayor o menor medida también sus pronunciaciones, difieren entre ambas lenguas. Un re-análisis de dos experimentos descritos por Schulpen (2003), y un experimento adicional de decisión léxica visual revelaron que el tamaño de familia morfológica en ambas lenguas afecta al reconocimiento de homógrafos interlingües por parte de las personas bilingües. Para los participantes que realizaron decision léxica visual en neerlandés, el tamaño de su familia morfolgica neerlandesa del homógrafo muestra una correlación negativa con las latencias de respuesta, mientras que el tamaño de su familia morfológica inglesa se correlaciona positivamente con las mismas. Por el contrario, en el caso de los participantes que realizaron el experimento de decisión léxica visual en lengua inglesa, los signos de dichas correlaciones se invierten, mostrando una correlación negativa entre las latencias y el tamaño de la familia morfológica inglesa, y positiva entre las latencias y el tamaño de la familia morfológica neerlandesa. Estos resultados señalan hacia la activación paralela de ambas interpretaciones de un homógrafo, junto cons sus paradigmas derivacionales en ambos idiomas, y se dan independientemente de la lengua en que los participantes realizan el experimento, y de si son conscientes de la presencia en el experimento de palabras en otra lengua. Es importante recalcar la coherencia de estos resultados con los efectos inhibidores y facilitadores de los diferentes campos semánticos de una raíz hebrea que describimos en el Capítulo 2.

El Capítulo 5 es el primero de los dedicados en esta tesis al modelado computacional. Introducimos una caracterización de los efectos de los paradigmas morfológicos basada en la teoría de la información, que cuantifica el coste cognitivo del reconocimiento de una palabra en función de los paradigmas morfológicos

en los que ésta se encuadra. Para ello desarrollamos una nueva medida de la complejidad de una palabra: su *residuo de información*. Esta medida captura los efectos de un gran número de otras variables que a las que se habían atribuido previamente efectos en el reconocieminto visual de una palabra, presentando una única descripción uniforme de sus effectos. El residuo de información se calcula en base a estructuras simbólicas predefinidas que estiman el apoyo probabilístico que los paradigmas morfológicos proporcionan a sus miembros. En nuestro estudio, el re-análisis de tres experimentos de decisión léxica visual previamente publicados concernientes a efectos de tamaño de familia morfológica, demostró que nuestro residuo de informacion supera a cualquier combinación de las medidas tradicionales en cuestión de la varianza explicada por cada modelo en una regresión lineal. Esta formulación matemática nos permite analizar en detalle las consecuencias para el procesamiento léxico de la organización interna de los paradigmas morfológicos, situando las relaciones inflexionales, derivacionales, y de palabras compuestas dentro de un marco unificado.

Los capítulos del 6 a 10 se concentran en el desarrollo de un modelo conexionista distribuido del reconocimiento visual de palabras, en el cual la morfología emerge de la interacción entre representaciones distribuidas de la forma (Capítulo 6) y el significado (capítulos 8 y 9) de las palabras. Estas representaciones se integran en el modelo descrito en el Capítulo 10.

En el Capítulo 6 tratamos con el problema que supone la creación de representaciones realistas de la forma de las palabras. Desarrollamos la técnica de *Acumulación de Expectaciones* (A.EE.) para representar las formas ortográficas y fonéticas de las palabras de un idioma, mediante un método computacional basado en el empleo de corpus y completamente distribuida. Nuestra técnica se basa en el uso de la representaciones que se desarrollan en la capa oculta de una Red de Recurrencia Simple entrenada en la prediccion del siguiente fonema o letra en una secuencia (Elman, 1990). El método A.EE. proporciona representaciones vectoriales abstractas de la forma de una palabra, sin restricciones sobre la longitud de la palabra o su complejidad morfológica, y sin introducir artificialmente estructura lingüística explícita en los vectores. Esta técnica contrasta con la práctica habitual en los modelos de conexionistas, que emplea representaciones que incorporan explícitamente estructura lingüística (tales como patrones del tipo CC-CVVCCC, etc) y restricciones en la complejidad de las palabras representables. Para asegurarnos de que nuestro sistema de representación puede emplearse en una red conexionista distribuida, comprobamos el rendimiento de una red que

modela un problema realista, empleando en su salida los vectores ortográficos A.EE. como representación. En un único modelo tratamos los problemas de la formación del pasado, la perdida del rasgo de voz al final de palabra, y la conversión de phonemas en grafemas en lengua neerlandesa. Demostramos que un modelo conexionista distribuido que emplea la técnica A.EE. es capaz de capturar los efectos paradigmaticos gradados que aparecen en la percepción fonológica en neerlandés, tal y como los describen Ernestus y Baayen (2001; 2003).

Los capítulos 8 y 9 desarrollan una técnica para la construcción de representaciones realistas del significado de las palabras. La práctica totalidad de los modelos conexionistas utiliza o bien vectores semánticos completamente artificiales, o representaciones basadas en estructuras de rasgos, construidas a mano para un pequeño grupo de palabras. En nuestros estudios, exploramos las posibilidades de crear representaciones empíricas y realistas del significado de las palabras mediante el uso de una Red de Recurrencia Simple entrenada en predicción de la siguiente palabra. Inicialmente, en el Capítulo 8, intentamos usar las representaciones formadas en la capa oculta de la red para capturar la información semántica contenida en los patrones de co-ocurrencia entre palabras. Aunque la técnica desarrollada en este capítulo consiguió capturar información sobre la categoría gramatical, valencia, e inflexión de las palabras, no mostró el rendimiento deseado en la captura de información sobre el significado de las palabras propiamente dicho. Esto se debe al régimen de entrenamiento de la red, predicción de la siguiente palabra en una sequencia, que demostró ser muy sensible a las propiedades morfo-sintácticas de las palabras, pero no a su significado léxico. El Capítulo 9 presenta un refinamiento sobre las representaciones morfo-sintácticas, que permite capturar también información sobre el significado de las palabras. La técnica se basa en aplicar una variante del método de A.EE. sobre los vectores morfo-sintácticos del capítulo anterior. Mediante la introducción explícita de una suma ponderada de las representaciones mofor-sintácticas en una ventana contextual que se desliza por el texto, esta técnica asegura la captura de los patrones de co-ocurrencia. De este modo obtenemos representaciones semánticas en un espacio multidimensional. Para la evaluación de las propiedades y contenido de las representaciones de estas características, desarrollamos también un conjunto de técnicas que se presentan en ambos capítulos.

Finalmente, en el Capítulo 10, combinamos las representaciones de la forma y el significado de las palabras desarrolladas en los capítulos anteriores, dando lugar a un modelo de amplia cobertura del reconocimiento visual de palabras. El

modelo es una simple red de retropropagación de tres capas (Rumelhart, Hinton, y Williams, 1986) que entrenamos para producir en su salida el vector semántico asociado a un vector ortográfico presentado en su entrada, sobre un vocabulario muy amplio, de aproximadamente 48.000 palabras, cuando el mayor vocabulario empleado anteriormente en un modelo conexionista del que tengamos noticia es de unas 10.000 palabras, sin llegar siquiera a tener en cuenta el significado de las mismas. Tras entrenar el modelo, la distancia entre la salida de la red y el vector semántico correspondiente (el error de la red), demostró ser un buen predictor de las latencias de respuesta humanas en decisión léxica visual. Nuestro modelo muestra efectos de frecuencia, entropía inflexional, entropía derivacional, número de vecinos ortográficos, regularidad verbal, y edad de acquisición, de manera similar a los efectos que aparecen en las latencias de los participantes en experimentos. El resultado crucial para nuestras investigaciones es que el modelo desarrolló automáticamente efectos de entropía derivacional e inflexional similares a los que calculamos en el Capítulo 5 sobre la base de estructuras simbólicas predefinidas, sin necesidad de incorporar ningún mecanismo especialmente dedicado al procesamiento morfológico. La convergencia entre el modelo simbólico basado en teoría de la información descrito en el Capítulo 5 y el modelo conexionista distribuido indica que la representación de los paradigmas morfológicos emerge automáticamente de las regularidades estadísticas presentes en las correspondencias entre forma y significado.

Conclusiones

Los resultados de los estudios experimentales sobre hebreo, finés, neerlandés, y bilingües que describimos en los capítulos del 2 al 4 subrayan la importancia de las estructuras paradigmáticas de una lengua para el procesamiento léxico. Lo crucial de estos resultados es el modo en que la gradación de las relaciones semánticas entre los miembros de un paradigma regula el efecto de tamaño de familia morfológica. La inmensa productividad de la morfología finesa tiene como consecuencia que un gran número de miembros de la misma familia morfológica están muy vagamente relacionados entre sí. Por tanto, tal y como observamos en el caso de estímulos morfológicamente complejos, los hablantes de finés no se ven necesariamente influidos por la familia morfológica completa de una palabra, sino sólamente por aquellos miembros que se encuentran en su vecindad semántica inmediata. En el caso de la menos productiva morfología del neer-

landés, las relaciones semánticas entre miembros de un paradigma tienden a ser más fuertes, pero, incluso en este caso, el nivel de similitud semántica con la palabra presentada, pondera la contribución de cada miembro al efecto de tamaño de familia. La elevada hetereogeneidad semántica de los paradigmas derivacionales de las raíces homonícas hebreas reveló una modulación más radical del efecto de tamaño de familia por factores semánticos, presentando inhibición en lugar de facilitación para los miembros de un paradigma semánticamente distantes de la palabra-estímulo, del mismo modo en que, en el léxico bilingüe, las familias morfológicas de un homógrafo interlingüe en cada idioma tienen efectos opuestos en su reconocimiento. Nuestras investigaciones también muestran que el efecto de tamaño de familia morfológica es una herramienta muy útil para investigar similitudes y diferencias en las estructuras conceptuales de varios idiomas.

Hemos propuesto dos modelos complementarios del reconocimiento léxico. El modelo basado en teoría de la información desarrollado en el Capítulo 5 nos proporciona una estimación cuantitativa directa de las consecuencias de la estructura lingüística para los procesos que gobiernan el reconocimiento léxico. Es de destacar que esta sencilla técnica permite explicar hasta las dos terceras partes de la varianza en experimentos de decisión léxica visual en lengua neerlandesa. Sin embargo, a pesar de las ventajas que presenta en cuestiones de interpretabilidad y predictividad, esta técnica presenta dos inconvenientes derivados directamente del empleo de representaciones jerárquicas predefinidas de la estructura morfológica. El primero de estos inconvenientes es que dichas representaciones jerárquicas son incapaces de reflejar la gradación continua de las relaciones morfológicas. El segundo inconveniente es que el modelo no explica la aparición de las propias estructuras jerárquicas de constituyentes.

Estos dos problemas no se dan en el modelo conexionista distribuido que desarrollamos en los capítulos del 6 al 10, que constituye la primera aproximación conexionista al reconocimiento visual de palabras en una escala realista. Dicho modelo representa el coste cognitivo del procesamiento léxico en términos de distancias en un espacio multidimensional. Este modelo demuestra que es posible aprender y representar estructuras paradigmáticas arbóreas mediante patrones de activación distribuidos en un espacio multidimensional, prescindiendo de regímenes de entrenamiento específicamente orientados a tratar la composicionalidad morfológica, y sin soluccionar el problema de antemano mediante el uso de representaciones estructuradas de la entrada. Nuestro modelo desarrolla representaciones completamente basadas en corpus que sólamente requieren presuposiciones mínimas

sobre el sistema cognitivo, tales como la proyección de expectaciones (ej., Tabor, Juliano, y Tanenhaus, 1996) y sensibilidad a los patrones de co-ocurrencia (Boroditsky y Ramscar, 2003; McDonald y Ramscar, 2001). En este modelo, la morfología emerge automáticamente como la 'convergencia de códigos', en el sentido propuesto por Seidenberg y Gonnerman (2000). Aunque es evidente que las redes de retropropagación no constituyen un modelo real del procesamiento de información en neuronas biológicas, dichas redes proporcionan una aproximación razonable de las interacciones entre los patrones de activación distribuidos que represantan las formas y significados de las palabras en el cerebro. La desventaja de este método es que resulta extremadamente difícil comprender qué factores influyen en el coste de procesamiento de una palabra en particular. Es de recalcar que McKay (2003) defiende que medidas basadas en la teoría de la información, similares a las que presentamos en el Capítulo 5, son las herramientas adecuadas para caracterizar la carga de procesamiento de una red de neuronas artificiales. Desde esta perspectiva, nuestro modelo basado en teoría de la información puede considerarse como la interpretación simbólica del procesamiento en la red subsimbólica.

References

Boroditsky, L. and Ramscar, M.: 2003, Guilt by association: Gleaning meaning from contextual co-occurrence, *Manuscript, Massachusetts Institute of Technology*.

De Jong, N. H.: 2002, *Morphological Families in the Mental Lexicon*, MPI Series in Psycholinguistics, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

Elman, J. L.: 1990, Finding structure in time, *Cognitive Science* **14**, 179–211.

Ernestus, M. and Baayen, H.: 2001, Choosing between the Dutch past-tense suffixes *-te* and *-de*, *in* T. van der Wouden and H. de Hoop (eds), *Linguistics in the Netherlands 2001*, Benjamins, Amsterdam, pp. 81–93.

Ernestus, M. and Baayen, R. H.: 2003, Predicting the unpredictable: Interpreting neutralized segments in Dutch, *Language* **79(1)**, 5–38.

McDonald, S. and Ramscar, M.: 2001, Testing the distributional hypothesis: The influence of context judgements of semantic similarity, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

McKay, D. J.: 2003, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge, U.K.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: 1986, Learning internal representations by error propagation, *in* D. E. Rumelhart and J. L. McClelland (eds), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, The MIT Press, Cambridge, Mass., pp. 318–364.

Seidenberg, M. S. and Gonnerman, L. M.: 2000, Explaining derivational morphology as the convergence of codes, *Trends in Cognitive Sciences* **4**(9), 353–361.

Tabor, W., Juliano, C. and Tanenhaus, M. K.: 1997, Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing, *Language and Cognitive Processes* **12**(2/3), 211–271.

Curriculum Vitae

Fermín Moscoso del Prado Martín was born in Ferrol, Spain, on April 24, 1974. He studied Computer Engineering at the *Universidad Politécnica de Madrid* (UPM), Spain, and received his *Licenciado* (M.Eng.) degree in 1998. During 1997 and 1998 he was a research assistant in the Universal Networking Language Project (UNL) in the Department of Artificial Intelligence of the UPM, where he also completed his master's thesis in December 1998. From January 1999 until July 1999 he was a visiting lecturer in the Department of Mathematics and the Faculty of Engineering of the Private University of Santa Cruz de la Sierra, Bolivia. From September 1999 until June 2000, he was a research student and assistant lecturer in the Departments of Psychology and Computer Science of the University of Hertfordshire, U.K. In July 2000 he returned to the UNL Project where he worked as a a research consultant until September 2000. In October 2000, he joined the Max Planck Institute for Psycholinguistics, where in August 2003, he completed his doctoral dissertation within the PIONIER project 'The balance of storage and computation in the mental lexicon', funded by the Dutch Research Council (NWO), the Faculty of Arts of the Universtity of Nijmegen, and the Max Planck Institute for Psycholinguistics. Currently, since September 2003, he is a postdoctoral research associate in the Medical Reseach Council – Cognition and Brain Sciences Unit at Cambridge, U.K.

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda van Turennout*

2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels Schiller*

3. Lexical access in the production of ellipsis and pronouns. *Bernadette Schmitt*

4. The open-/closed-class distinction in spoken-word recognition. *Alette Haveman*

5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*

6. Gesture and speech production. *Jan-Peter de Ruiter*

7. Comparative intonational phonology: English and German. *Esther Grabe*

8. Finiteness in adult and child German. *Ingeborg Lasser*

9. Language input for word discovery. *Joost van de Weijer*

10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*

11. Producing past and plural inflections. *Dirk Janssen*

12. Valence and transitivity in Saliba, an Austronesian language of Papua New Guinea. *Anna Margetts*

13. From speech to words. *Arie van der Lugt*

14. Simple and complex verbs in Jaminjung: A study of event categorization in an Australian language. *Eva Schultze-Berndt*

15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*

16. Language-specific listening: The case of phonetic sequences. *Andrea Weber*

17. Moving eyes and naming objects. *Femke van der Meulen*

18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*

19. Morphology in speech comprehension. *Kerstin Mauth*

20. Morphological families in the mental lexicon. *Nivja H. de Jong*

21. Fixed expressions and the production of idioms. *Simone A. Sprenger*

22. The grammatical encoding of postural semantics in Goemai (a West Chandic language of Nigeria). *Birgit Hellwig*

23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*