Steve Pepper

Lexical transfer in Norwegian interlanguage

A detection-based approach

Master's Thesis in Linguistics
UNIVERSITY OF OSLO
Department of Linguistics and Scandinavian Studies
December 2012

Abstract

This study investigates cross-linguistic influence ('transfer') in Norwegian interlanguage using predictive data mining technology with a focus on lexical transfer. The impetus for the present work came from the publication of a series of studies (Jarvis & Crossley 2012) that explore the 'detection-based approach' to language transfer.

The following research questions are addressed:

1. Can data mining techniques be used to identify the L1 background of Norwegian language learners on the basis of their use of lexical features of the target language?
2. If so, what are the best predictors of L1 background?
3. And can those predictors be traced to cross-linguistic influence?

The study utilizes data from *Norsk andrespråkskorpus* (ASK), the Norwegian Second Language Corpus housed at the University of Bergen, and draws on resources from the ASKeladden project. The source data consists of texts written by 1,736 second language learners of Norwegian from ten different L1 backgrounds, and a control corpus of 200 texts written by native speakers. Word frequencies computed from this data are analysed using multivariate statistical methods that include analysis of variance and linear discriminant analysis, and the results are subjected to contrastive analysis.

The combination of discriminant analysis and contrastive analysis produces all three types of evidence called for by Jarvis (2000) in his methodological requirements for language transfer research: intragroup homogeneity, intergroup heterogeneity and cross-language congruity. Well-known transfer effects, such as the tendency for Russian learners to omit indefinite articles, are confirmed, and other, more subtle patterns of learner language are revealed, such as the tendency amongst Dutch learners to overuse the modal verb *skal* to a far greater extent than other learners. In addition to confirming the reality of lexical transfer, these results provide abundant material for further research, while the methodology employed can be harnessed in many areas of linguistic research.

An Sylvia

meine Muse und Studienobjekt

Was ist Sylvia, saget an,
Daß sie die weite Flur preist?
Schön und zart seh ich sie nahn,
Auf Himmelsgunst und Spur weist,
Daß ihr alles untertan.

Ist sie schön und gut dazu?
Reiz labt wie milde Kindheit;
Ihrem Aug' eilt Amor zu,
Dort heilt er seine Blindheit
Und verweilt in süßer Ruh.

Darum Sylvia, tön, o Sang,
Der holden Sylvia Ehren;
Jeden Reiz besiegt sie lang,
Den Erde kann gewähren:
Kränze ihr und Saitenklang!

Preface

When I wrote my first MA thesis for the course in Language Documentation and Description at SOAS in 2010 my plan was to follow Bob Dixon's injunction to go forth and document an endangered language. When that plan was scuppered by my dreadful hearing, I decided to take up the place I had been offered at the University of Oslo, embark on another Masters degree course, and apply for a doctoral position. I had completed the taught component of that course and made two false starts on a second thesis (on topics related to my doctoral projects in historical linguistics and language typology) when I stumbled upon the field of language transfer research. It was as if I had finally come home and I resolved to change my thesis topic (and doctoral project) yet again. This is the result.

I have a lot of people to thank. First of all, my primary supervisor, *Rolf Theil*, for infecting me with his passion for linguistics in the first place, and for guiding me through four wildly different MA topics. His polyglot knowledge was especially valuable in the chapter on contrastive analysis and he also forced me to make the technical stuff understandable for a linguist. (I'm only sorry we never got to reconstruct Proto-Mambiloid together.)

My second supervisor, *Anne Golden*, smoothed my path into the field of SLA, directed me towards the Norwegian Second Language Corpus (ASK) and oversaw the current project, showing great enthusiasm for the actual topic. (I hope we get to pursue this work in the future.)

Without *Kari Tenfjord* of the University of Bergen's ASKeladden project the data I have used wouldn't have existed, and her support was both moral and financial. *Scott Jarvis* of Ohio University provided the initial stimulus and has also been very helpful in clarifying a number of details. *Bjørn-Helge Mevik* of the University Centre for Information Technology helped me get started with R, and *Bård Uri Jensen* of Hedmark University College offered useful guidance as I grappled with the completely new field of statistics. My friends *Rafał* and *Naito-san* very graciously lent me samples of their respective interlanguages.

And last, but not least, it was *Sylvia*, who once again made it all possible. It is with love, admiration and affection that I dedicate this work to her.

Table of contents

Table of figures

Table of tables

Abbreviations

See also *Language codes* (p. xi) and *Appendix A. Glossary of terms* (p.118).

1-gram	unigram (e.g. an individual word, see gram)
2-gram	bigram, sequence of two grams
3-gram	trigram, sequence of three grams
4-gram	sequence of four grams
ART	article
ANOVA	analysis of variance
ASK	*Norsk andrespråkskorpus* (Norwegian Second Language Corpus)
CA	contrastive analysis
CAH	contrastive analysis hypothesis
CV	cross-validation
def	definite
DFA	discriminant function analysis
EFL	English as a Foreign Language
f	feminine
FL	foreign language
gram	a contiguous sequence of *n* items from a given sequence of text or speech; the items in question can be phonemes, syllables, letters, words etc. according to the application
ICLE	International Corpus of Learner English
ind	indefinite
L1	first language
L2	second language
LDA	linear discriminant analysis
LDDA	linear diagonal discriminant analysis
LDF	linear discriminant function
LOOCV	leave-one-out cross-validation
m	masculine
M	statistical mean
MANOVA	multivariate analysis of variance
n	neuter
***n*-gram**	(specific) sequence of *n* words
NL	native language
NOA	*Norsk som andrespråk* (Norwegian as a Second Language)
NP	noun phrase
PCA	Principal Components Analysis
pl	plural
POS	part of speech, word class; a POS *n*-gram is a sequence of *n* words expressed in terms of word classes, e.g. *on the road* as an instance of the POS 3-gram 'PRP ART SB'
PP	prepositional phrase
PRP	preposition
s	standard deviation
SB	noun
SD	standard deviation
SDDA	stepwise diagonal discriminant analysis
sg	singular
SL	source language
SLA	second language acquisition
SVM	Support Vector Machine
TL	target language
vb	verb

Language codes

For reasons of space, L1s represented in the Norwegian Second Language Corpus (ASK) and the International Corpus of Learner English (ICLE) or used in related studies, are usually referred to by their (two-letter) ISO 639 codes. Most of these are well-known and fairly transparent, but the reader should take special note of SH, SQ and ZH.

AR Arabic	FR French	**PL Polish**	SW Swedish
BU Bulgarian	HU Hungarian	PO Portuguese	TR Turkish
CS Czech	IT Italian	**RU Russian**	TS Tswana
DA Danish	JA Japanese	**SH**[2] **Serbo-Croat**	**VI Vietnamese**
DE German[1]	LA Latin	**SO Somali**	WO Wolof
EN English	**NL Dutch**	**SP Spanish**	ZH Mandarin
FI Finnish	**NO Norwegian**	**SQ Albanian**	

Table 1: Language codes

Note 1: L1s represented in ASK are shown in boldface.

Note 2: The code SH is now deprecated by ISO. The language formerly known as Serbo-Croat (or Serbo-Croatian) has been assigned the three-letter code HBS in ISO 639-2 and is now regarded by Ethnologue (Lewis 2009) as a "macro language" whose member languages are Croatian (HR), Bosnian (BS) and Serbian (SR). However, ASK labels the relevant texts as '*serbokroatisk*' and makes no distinction between the different varieties. The same practice is followed here and the code SH is used accordingly.

Typographical conventions

Norwegian words are italicised and glosses are occasionally supplied in parentheses, e.g. *barn* ('child'). A complete, glossed list of Norwegian words used in the analysis of word frequencies is provided in Appendix B. The names of files, worksheets and scripts are shown in **boldface**. Function names and extracts of program code are shown in a `monospace` font.

1. Introduction

1.1 Background

As a foreigner who has lived abroad for over 35 years, language has never ceased to intrigue me. As a father whose first language (L1) is his children's second, and whose second language (L2) is his children's first, the acquisition of foreign languages concerns me at a personal level. As a husband whose wife speaks his first language as her third, and with whom he often communicates in his second (her fourth) language, the delights of multilingualism and the tricks it plays on our use of language are a source of continual fascination.

As a traveller who had the good fortune to be brought up a native speaker of British English, I was always intrigued by the way you could pinpoint someone's nationality by the accent they have when speaking English. As a linguist, I now know that this phenomenon has nothing to do with nationality per se, but is rather due to the cross-linguistic influence of the mother tongue (which may or may not bear some relationship to nationality).

As a teacher of English as a foreign language to adults, I discovered that the influence of the mother tongue extends not only to pronunciation, but also to vocabulary, grammar and every other aspect of language. And as a dilettante polyglot who has learned at least bits of a dozen and more languages, I derive great pleasure from being able to trace specific influences back to what I know about the source language.

While the extent and nature of cross-linguistic influence has been (and to a certain extent, still is) hotly disputed by linguists, the fact of it seems to me indisputable. Given enough experience, a foreign language teacher can say a lot about the language backgrounds of learners from the ways in which they use the target language. Pronunciation is invariably the

biggest giveaway, but also written language can be "mined" for clues about the language background of its user. As an example, take the following text, written by an advanced learner of English:

(1) We thought we would move our stay in Klekotki to an earler (*sic*) weekend so that you could come. What about period Friday 26 to Monday 29 October, with some flexibility?

Apart from the typo, the only deviation from native speaker English here is the omission of the definite article in the second sentence. This tiny irregularity gives the text an unmistakeably foreign flavour: it is an error no native speaker would be likely to make – and a good example of how the first language (L1) can influence the second (L2). In this case it is not unlikely that influence from a source language that lacks the definite article is leading the user to omit it in the target language. To an experienced language teacher this might suggest an Eastern European, South Asian or Japanese provenance.

A more extensive text could provide further clues:

(2) I am still looking for Ryokans in Kyoto. You can search Ryokans at the following site: […] I found two english web site of Ryokans. Please enjoy them! I can make Ryokan reservation for you if you want.

Among the particular characteristics in (2) we again note the missing article in the final sentence, and also the absence of plural marking on the noun 'web site', the idiosyncratic capitalization, and the quaintly polite exhortative *Please enjoy them!* Now, the number of languages in the world that lack both articles and plural marking is relatively small: the largest by far are Mandarin Chinese and Japanese, both of which have writing systems that make no distinction between upper and lower case (which could account for the distinctive capitalization). Of the two, Japanese is the one most associated with a culture in which politeness plays a key role.[1]

[1] A more detailed contrastive analysis of the constructions 'you can search Ryokans' and 'web site of Ryokans', and comparison with 以下のサイトで旅館を探すことができます (ika no saito de ryokan wo sagasu koto ga dekimasu) and 旅館について二つの英語のサイトを見つけました (ryokan ni tsuite futatsu no saito wo mitsukemashita), might reveal additional clues pointing to Japanese.

Of course, the origins of (1) and (2) in Polish and Japanese, respectively, are given away by the references to place names, but the point is that *even if that were not the case*, it is fully possible to pinpoint the source language on the basis of particular features of the learner's 'interlanguage'. Knowledge of the language(s) concerned can help explain the 'interference' but it is not necessary: all that is required is sufficiently frequent exposure to "errors" of a particular type from speakers with a particular background. This explains why not only polyglot linguists with an interest in language typology, but also monoglot language teachers can correctly identify a learner's L1 – provided they have sufficient experience and enough data to go on.

It is possible, then, for humans to predict the L1 of learners on the basis of their written language. As a knowledge engineer, this leads me to wonder whether computers – given enough data – might be able to do the same.

That question was addressed in a recently published book (Jarvis & Crossley 2012) in which Scott Jarvis and his colleagues use "text classification" techniques associated with data mining to explore what he terms the "detection-based approach" to investigating language transfer. Those exploratory studies suggest that the predictive methods of advanced statistics (in particular, linear discriminant analysis) can indeed be used to classify learner texts by source language and also to identify which linguistic features are the best L1 predictors. This type of analysis represents a valuable new impetus for future directions in transfer research:

> Traditional studies dealing with second language acquisition, including transfer, have often investigated "bits and pieces of learners' language chosen for analysis because they caught the researcher's eye…" (Lightbrown 1984: 245). The detection-based, corpus-driven approach … expands the scope of transfer inquiry to include items that might otherwise never have attracted the researcher's eye. This is the nature of discovery, and … the detection-based approach is well suited to a program of discovery" (Jarvis & Paquot 2012: 100-101).

The Jarvis & Crossley studies are all exploratory in nature, and they are all based on texts whose target language is English. As far as is known, no other studies have to date applied the same techniques to texts with a different target language. The purpose of the present study is to remedy this situation using Norwegian as the target language, with the primary goal of either confirming or disconfirming the generalizability of the results obtained by

Jarvis *et al.* to other languages. A second goal is to pave the way for a more in-depth and theoretically grounded investigation to be conducted later – and perhaps reveal some "bits and pieces" of Norwegian interlanguage that have hitherto escaped notice.

The present work is loosely modelled on the first of the Jarvis & Crossley studies (Jarvis *et al.* 2012), which uses frequency data for 53 highly frequent words in texts written by learners of English as a foreign language from five different L1 backgrounds. The same detection-based approach is applied here to six sets of lexical data, each from texts written by learners from five different L1 backgrounds. However, the present work also differs from the study by Jarvis *et al.* in a number of ways, the most important of which are:

1. the target language is Norwegian rather than English;
2. the data comes from second language learners rather than foreign language learners;[1]
3. ten source languages are involved in the study rather than five;
4. the source texts are thematically heterogeneous;
5. open source tools have been employed in preference to proprietary software;
6. all data and scripts are publicly available.

1.2 Research questions

The general research question addressed in this study is the same as that pursued in Jarvis & Crossley (2012) – "is it possible to identify the L1 background of a language learner on the basis of his or her use of certain specific features of the target language?" – and the focus, as for Jarvis and his colleagues, is on the machine-learning capabilities of computer classifiers rather than the "psycholinguistic ability of human judges" (Jarvis 2012: 1).

The linguistic phenomena investigated by Jarvis *et al.* are of various kinds – lexical, grammatical and stylistic – as described in §2.3.2. For reasons of scope, the present study is restricted to lexical features, more specifically the frequency of occurrence of individual

[1] For the present purpose the most important distinction between the two is "the amount of exposure that learners have to the target language outside of the classroom" (Loewen & Reinders 2011: 68). The learners in the study by Jarvis *et al.* had all learned English as a foreign language (EFL) in their home countries, whereas those in the present study have (almost) all learned Norwegian as a second language (NOA) while living in Norway.

words. The principal research question can therefore be formulated more precisely as follows:

Q1 *Can data mining techniques be used to identify the L1 background of Norwegian language learners based on their use of lexical features of the target language?*

Assuming an affirmative answer to this question, subsidiary research questions are:

Q2 *What are the best (lexical) source language predictors?*
Q3 *Can those predictors be traced to cross-linguistic influence?*

1.3 Theoretical framework

This study is empirical and exploratory in nature and therefore to some degree theory-neutral. A central assumption, however, is the reality (and importance) of cross-linguistic influence and, by implication, the importance of frequency effects in language learning. The latter is a major tenet of cognitive linguistics, which holds that language learning is based on general cognitive abilities rather than a specialized language organ. Thus, theories of construction grammar (Croft 2007), the usage-based perspective (Tomasello 2003), the role of frequency effects (Bybee 2010), and the constructionist view of language acquisition (Ellis 2003) will provide the theoretical foundation for the present investigation and work that will follow from it.

An important lodestone is Jarvis' (2000, 2010) framework for methodological rigour in transfer research (see §2.1.3). The first two types of evidence that Jarvis calls for turn out to have their parallel in the underlying goal of linear discriminant analysis, which forms an important part of the theoretical foundation for this project.

In order to address the question of whether lexical behaviour found to be distinctive of particular L1 groups can be traced to cross-linguistic influence (and thus provide the third type of evidence called for by Jarvis), contrastive analysis will be employed – in its 'weak' or 'diagnostic' form (Gast to appear) – using the approach first developed by Fries (1945) and Lado (1957), and exemplified for Norwegian by Lie (2005), Golden *et al.* (2008), Næss (2011a), and other works cited therein.

1.4 Hypothesis and contribution to knowledge

The basic hypotheses of this study are as follows:

H1 *That data mining techniques can be used (up to a point) to identify the L1 background of a Norwegian language learner on the basis of his or her use of lexical features of the target language.*

H2 *That such techniques can also reveal subtle patterns of learner language that might otherwise not be detected.*

H3 *That many (but not all) of the L1 predictors thus revealed will be traceable to cross-linguistic influence.*

The study contributes to knowledge in a number of ways. First of all, it serves to validate the techniques developed by Jarvis *et al.* to a target language other than English. Secondly, it reveals some very interesting and subtle aspects of Norwegian interlanguage that have not been recognized earlier, as well as confirming certain other observations that are rather well-known. Thirdly, it goes considerably beyond Jarvis *et al.* in demonstrating how contrastive analysis can be applied to provide more compelling cross-linguistic explanations for the patterns revealed by the statistical analysis. Fourthly, the carefully documented methods and the publication of both source data and source code will mean that the techniques used here will be more readily available to other researchers. Finally, it is to be hoped that some of the lexical features that turn out to be the best predictors of L1 background in Norwegian interlanguage will lead to new insights into language transfer and provide material for future research.

1.5 Structure of this dissertation

Following this introductory chapter, Chapter 2 presents the theoretical background for the project, including a review of previous work on the application of discriminant analysis in linguistics and, in particular, transfer research. Chapter 3 provides an overview of the data sources and Chapter 4 describes both the methodology employed in the study and various methodological issues that arose. The findings are presented in Chapter 5 and these are discussed in the light of contrastive analysis in Chapter 6. Finally, Chapter 7 lays out answers to the research questions, evaluates the hypotheses, and suggests ideas for further work. The appendices contain additional material, in particular intermediate results and

examples of the methodologies used. A companion web site is planned and will be initially hosted at http://folk.uio.no/stevepe.

1.6 Terminology

The terms ‘first language’ (L1), ‘mother tongue’ and ‘source language’ (SL) are used interchangeably in this dissertation. The term ‘second language’ (L2) is used in two senses. Usually it refers to any language learned after the first language(s), irrespective of where or how it is acquired; in this sense it is usually synonymous with ‘target language’ (TL). However, occasionally ‘second language’ is used in contrast to ‘foreign language’ (as in footnote 1 on page 4) to denote a language acquired in the community in which it is spoken, as opposed to one acquired outside such a community.

Following Larson-Hall (2010: 402) I have adopted Kline’s (2004) recommendation to return the word “significant” to its common meaning of “important” and instead of talking about “significant results” and “significant differences” to simply use the adjective “statistical”.

Since the primary audience for this thesis is linguists and not statisticians or mathematicians, a glossary of statistical and mathematical terms is included in Appendix A. Mathematically and/or statistically challenged readers should be aware that they do not need to understand all the technical details of §2.2 and §4.4 in order to grasp the basic ideas underlying the methodology used in this study, nor to appreciate the results produced by that methodology – nor even to apply the same methodology in their own work.

Finally, a note (for the benefit of prescriptivists) on the word ‘data’, which is currently undergoing a transition in Modern English from count noun to mass noun. Originally the plural form of the Latin loanword ‘datum’, ‘data’ is still often used as a plural word within Academia. Without, on the other hand, and especially in IT, ‘data’ is far more frequently used as a mass noun. With one foot in each camp, the present writer reflects this indeterminacy and vacillates in his used of the word. No attempt has been made in this thesis to achieve consistency and the reader will therefore encounter both the plural form (the data *are*) and the singular (the data *is*).

2. Theoretical background

This chapter discusses the theoretical background for the present study and reviews relevant literature. It starts with a brief discussion of the field of study in which the project is located: language transfer. This is followed by an introduction to the technology employed: that is, the method of linear discriminant analysis (LDA). The chapter concludes with a review of previous work using LDA in the field of linguistics and a fairly detailed account of applications of LDA in transfer research.

2.1 Language transfer

2.1.1 Background

The term 'language transfer' is used in Second Language Acquisition research to refer to "the influence resulting from similarities and differences between the target language and any other language that has been previously (and perhaps imperfectly) acquired" (Odlin 1989: 27). It is generally used synonymously with 'cross-linguistic influence' (CLI), which Jarvis & Pavlenko (2008: 1) define as "the influence of a person's knowledge of one language on that person's knowledge or use of another."

Historically, the study of transfer has been through several "swings of the pendulum" (Gass 1996, quoted in Mitchell & Myles 2004: 19). Its origins can be traced to the work of Fries, Lado, Haugen and Weinreich in the post-war period (when the phenomenon was referred to as 'interference') and to the practice of contrastive analysis. In the 1970s the latter concept became tarred with the brush of behaviourism – undeservedly, according to Swan (2007). However, today most theorists accept that cross-linguistic influence plays an important role in second language learning and that it "can occur in *all* linguistic subsystems, including morphology and syntax" (Odlin 1989: 23). The earlier focus on "interference" and "errors"

(negative transfer) has been balanced by the recognition of positive transfer, described by (Odlin 1989: 26) as "the facilitating influence of cognate vocabulary or any other similarities between the native and target languages", and of other manifestations of cross-linguistic influence such as avoidance and over-use. The view of learner language as a 'deficient' form of the TL has similarly been replaced by the concept of 'interlanguage' as a linguistic system in its own right (Selinker 1972).

More recent developments in transfer research include "new findings and refinements in already established areas, such as lexis and phonology; new areas and directions of transfer research, such as reverse transfer, sociolinguistic transfer, and the study of the multilingual lexicon; and new theoretical accounts of CLI, such as conceptual transfer" (Jarvis & Pavlenko 2008: 212).

2.1.2 Lexical transfer

The focus of the current study is cross-linguistic influence in the lexicon, or lexical transfer. Whereas most previous work on lexical transfer (Ringbom 1987, 2001, 2007; Arabski 2006; Jarvis & Pavlenko 2008; Jarvis 2009; Llach 2010; etc.) deals primarily with *content* words, the methods and data used in this study bring *function* words to the fore. Nevertheless, the framework established previously is relevant here.

Lexical transfer can involve both form and meaning. Citing examples from Ringbom (1987, 2001) and Poulisse (1999), Jarvis & Pavlenko (2008: 75) present five kinds of lexical transfer, grouped according to whether they involve "morphophonological" (form-related) errors or "semantic" (meaning-related) errors, as follows:

Form-related transfer

(a) **False friends**:[1] *many <u>offers</u> of violence have not enough courage to speak about it* (SW *offer* 'victim')

(b) **Unintended code-switching**: *and then <u>nog</u> one* (NL *nog* 'another')

(c) **Cross-linguistic blends**: *we have the same <u>clothers</u>* (EN clothes + SW *kläder* 'clothes')

[1] Jarvis (2009) rightly prefers this term to "false cognates", since it covers loanwords and accidental resemblances as well as cognates. It also meets the justified objection of historical linguists that "false" cognates like 'offer' are in fact true cognates in their understanding of the term.

Meaning-related transfer

(d) **Semantic extension**, i.e. authentic TL words used in senses that reflect SL semantic ranges: *he bit himself in the language* (FI *kieli* 'language, tongue')

(e) **Calques**: *he remained a youngman all his life* (SW *ungkarl* 'bachelor' < *ung* 'young' + *karl* 'man')

Jarvis (2009: 100) reformulates this taxonomy in terms of a proposed distinction between 'lemma' and 'lexeme' (which are said to specify a word's semantic-syntactic and morpho-phonological properties, respectively). Lexemic transfer is essentially the same as form-related transfer and is divided into the same three subtypes; lemmatic transfer, on the other hand, includes the two types of meaning-related transfer (d) and (e), and two further types:

(f) **Collocational transfer**: *there is also people who wants to get married, do children and build a nice house* (FI *tehdä lapsia* 'have children, lit. do children')

(g) **Subcategorization transfer**, i.e. selection of the wrong type of complement to a headword: *she kissed with him* (PP instead of NP); *he thinking his mother* (NP instead of PP); *late from an appointment* (incorrect PRP)

Jarvis makes the point that lemmatic transfer, in contrast to lexemic transfer, does not appear to be constrained by language distance Jarvis (2009: 118).

2.1.3 Methodological framework

A methodological framework for achieving empirical rigour in the identification of instances of transfer was proposed by Jarvis (2000). It consists of three components:

(a) a theory-neutral definition that defines CLI as the relationship between source-language group membership and target-language behaviour,

(b) a specification of the types of evidence for CLI that the definition implies, and

(c) a list of the various types of factors that need to be taken into account in an investigation of CLI because of their potential to produce effects resembling CLI and, conversely, because of their potential to obscure the effects that CLI itself produces (Jarvis 2010: 170).

The three types of evidence originally adduced and reaffirmed in Jarvis & Pavlenko (2008: 35, 41ff) are shown in Table 2. (Note that the term 'group' stands here for 'L1 group', i.e. a group of learners with the same L1 background.)

Original terms	**Heuristic terms**	**Definition**
intragroup homogeneity	within-group similarities	Evidence that the behaviour in question is not an isolated incident, but is instead a common tendency of individuals who know the same combination of languages.
intergroup heterogeneity	between-group differences	Evidence that the behaviour in question is not something that all language users do regardless of the combination of L1s and L2s that they know.
cross-language congruity	between-language similarities	Evidence that a language user's behaviour in one language really is motivated by her use (i.e. the way she demonstrates her knowledge) of another language.

Table 2: Three types of evidence for transfer

All three types of evidence involve an interplay of quantitative and qualitative considerations. However, the first two tend to be used with a quantitative emphasis, e.g. by showing how common a particular pattern of TL use is among learners from a particular L1 background (intragroup homogeneity), and establishing whether learners from different L1 backgrounds use that pattern with statistically different rates of occurrence (intergroup heterogeneity). The third type of evidence, on the other hand, is "often used with a qualitative emphasis by showing that learners' … patterns of performance in the source and target language are qualitatively similar" (Jarvis 2010: 173).

The present study invokes all three types of evidence: the first two through the application of discriminant analysis (see §2.2) and the third through the use of contrastive analysis (see below).[1]

[1] Recognizing that these three types of evidence are based on the dichotomy of similarity vs. difference within and between two types of entity: languages and L1 groups, Jarvis (2010: 175) introduces a fourth type of evidence, 'intralingual contrasts', in order to complete the taxonomic model. For reasons of scope, it has not been possible to incorporate this type of evidence into the current project.

2.1.4 Contrastive analysis

Contrastive analysis (CA) investigates the differences between pairs of languages against the background of similarities, in order to provide input to disciplines such as foreign language teaching and translation studies (James 1980; Gast to appear). The contrastive methodology formulated by Fries (1945) and Lado (1957) was based on the 'contrastive analysis hypothesis' (CAH) which in its strong form suggested that "L2 acquisition consists of a transfer of L1 habits to the L2" (Loewen & Reinders 2011: 42). The CA methodology was applied extensively in the 1960s but fell into disfavour in the US in the 1970s, partly because of its links with behaviourism and partly because of the failure of the CAH to explain all the facts of second language acquisition.

In Europe, however, many CA projects were initiated, "most of them comparing English to the native language of the investigators" (Fisiak 1981; Gast to appear). Works such as Swan & Smith (2001) and König & Gast (2009) are evidence that this tradition is still strong. In Norway the period 1980-2005 saw the publication of many contrastive grammars in which Norwegian is compared with the languages of immigrants (Golden *et al.* 2007: 20), e.g. Andenæs (1984), Bruland *et al.* (1979), Husby (1989, 1991, 1999, 2000, 2001), Hvenekilde (1980) and Rosén (1999), and similar work has been carried out more recently by Wiull (2006a, 2006b, 2007, 2008, 2009, 2010).

A 'weak' or 'diagnostic' form of the CAH was formulated by Wardhaugh (1970):

> The weak version [of the contrastive analysis hypothesis] requires of the linguist only that he use the best linguistic knowledge available to him in order to account for observed difficulties in second language learning… [T]he starting point…is provided by actual evidence from such phenomenon as faulty translation, learning difficulties, residual foreign accents, and so on, and reference is made to the two systems only in order to explain actually observed phenomena.

Wardhaugh talks only in terms of negative transfer, but contrastive analysis in its diagnostic form is equally applicable to positive transfer, and it plays an important role in the present study (see §6.4).

2.1.5 Learner corpus research

Over the last decade an increasingly important role in SLA research has been played by computer learner corpora. These are a special kind of language corpus, defined by Granger (2008: 338) as "electronic collections of (near-) natural foreign or second language learner texts assembled according to explicit design criteria." One of the most influential foreign language corpora is the International Corpus of Learner English (ICLE), based at the University of Louvain in Belgium (Granger *et al.* 2009). The current project is based on ICLE's Norwegian counterpart, *Norsk andrespråkskorpus* (ASK), which is described in §3.1.

2.1.6 The detection-based approach

One of the most recent developments in the study of transfer has been the use of statistical methods to detect cross-linguistic influence. Inspired by techniques used in stylistics and data mining, Scott Jarvis of Ohio University and his collaborators have shown that classificatory methods such as discriminant analysis can be used to detect "subtle, complex, and unpredicted instances of L1 influence that can easily be overlooked – and may not even be anticipated – in the comparison-based approach" (Jarvis 2012). The present study represents the first attempt to apply such techniques to Norwegian data. Before discussing these studies in more detail, a general introduction to the concepts and goals of discriminant analysis is provided, along with some earlier examples of its application in the field of linguistics.

2.2 Discriminant analysis

2.2.1 Fundamentals of discriminant analysis

Discriminant analysis in its various flavours, along with other classification and clustering methods, is an important technique in the field of data mining, in which the goal is to attempt to discover patterns in large data sets. In the subfield of data mining known as text mining, the source data (as the name suggests) is textual. When the source data is textual and the goal is predictive or classificatory, the term 'text classification' is used – as it is in the subtitle of Jarvis & Crossley (2012).

Klecka (1980: 7) defines discriminant analysis as "a statistical technique which allows the researcher to study the differences between two or more groups of objects with respect to

several variables simultaneously."[1] Multivariate methods that deal with groups of objects are subdivided into clustering ('unsupervised learning') and classification ('supervised learning'), depending on whether or not the method has access to prior knowledge about how the objects are grouped. In cluster analysis the goal is to discover structure in the data (in the form of groupings of observations); in classification, which includes LDA and with a number of other techniques, prior knowledge about groupings is utilized towards goals that are either predictive or descriptive (or both).

The predictive goal of classification is to ascertain the group membership of objects on the basis of certain properties that they exhibit. In the classic example (reported in Everitt 2005: 137ff), the data are based on two groups of skulls found in Tibet. On each of the 32 skulls, five measurements were recorded, for length, breadth, height, face height and face breadth, respectively. Thus there were 32 objects (or 'cases') whose groupings were known, each of which was described by five variables. The problem to be solved was how to classify further skulls whose measurements were known but whose groupings were not. The method of discriminant analysis, developed by Fisher (1936), provided a means of predicting *group membership* (i.e., skull type) based on *discriminating variables* (i.e., the five measurements).

The same method can also be used with the (explanatory) goal of describing how groups differ, and which variables (or 'features') constitute the best discriminators (or 'predictors') of group membership. In an example given by Hand (2005) the aim is to use variables such as gender, responsiveness, gestational age, birth weight, etc. to explain how infants at high risk of dying from Respiratory Distress Syndrome differ from those at low risk. Using LDA it is possible to determine which of the variables correlate most closely with the risk of death.

In both kinds of application – predictive and descriptive – there is a set of *cases* which can be divided into two or more *groups*, and a set of discriminating variables or *features*. In the present project, cases are learner texts from the ASK corpus; these are grouped according to

[1] In fact it is a family of such techniques, with names such as linear discriminant analysis (LDA), discriminant function analysis (DFA), quadratic discriminant analysis (QDA), diagonal discriminant analysis (DDA), etc. The current study will not be overly concerned with the distinction between these different variants and will talk for the most part in general terms about discriminant analysis, referring to it as LDA; in practice, however, it will mostly be using linear discriminant analysis.

the learner's L1, and the features used in the analysis are word frequencies.[1] On the basis of this data, statistical models are constructed using LDA in order to discriminate between the L1 groups. The predictive approach is then used to test the accuracy of the models, and the descriptive approach to identify which features (i.e. lexical choices) are the best predictors of group membership (i.e. L1 background).

LDA is similar to the statistical method of multiple regression, with the important difference that whereas the outcome in regression is quantitative (i.e., a numeric value), in LDA it is a categorical variable – such as L1.

2.2.1.1 How discriminant analysis works

LDA can handle an arbitrary number of groups (provided there is a sufficient number of cases and variables), but the principles are best explained using a simple example in which there are only two groups (or classes), each with just two cases. In the following, a subset of the Tibetan skull data mentioned above is used for this purpose (see Table 3).

Case	Length (mm)	Breadth (mm)	Type	LDF
A	162.5	139.0	I	-2.27265
B	174.5	143.5	II	2.324759
C	178.5	135.0	I	-0.91175
D	182.5	137.0	II	0.859172

Table 3: Simple LDA example (Tibetan skulls data)

Ignoring the column labelled `LDF` for the moment, we observe four cases (A, B, C and D) that each belong to one of two groups (Type I and Type II). For each case there are two numeric variables: the measurements for `Length` and `Breadth`. A scatter plot of the data is shown in Figure 1.

[1] Note that there is a one-to-one correspondence in the present project between 'feature' (as used in its technical, LDA-related sense) and 'word', since each feature is a data variable consisting of the relative frequency of a particular word.

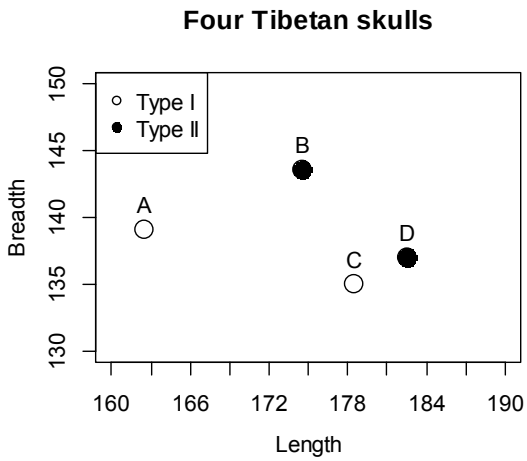


Figure 1: Simple LDA example – scatter plot

On their own, neither of the two variables `Length` or `Breadth` suffice to discriminate between the two groups. This is demonstrated for `Length` in Figure 2, in which the four data points are projected onto the `Length` axis; there is no point along this axis at which a line can be drawn to separate the two groups.

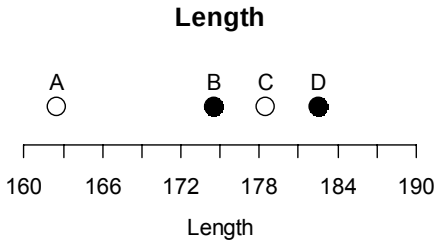


Figure 2: Cases projected onto the Length axis

Figure 3 shows the same four data points projected onto the `Breadth` axis. Again there is no point where a line can be drawn to separate the two groups.

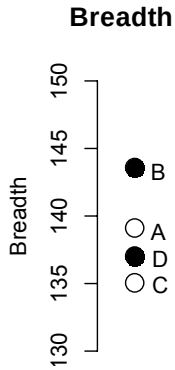


Figure 3: Cases projected onto the Breadth axis

However, by combining the two variables in a simple linear function a new axis can be produced (shown as a dotted line in Figure 4).

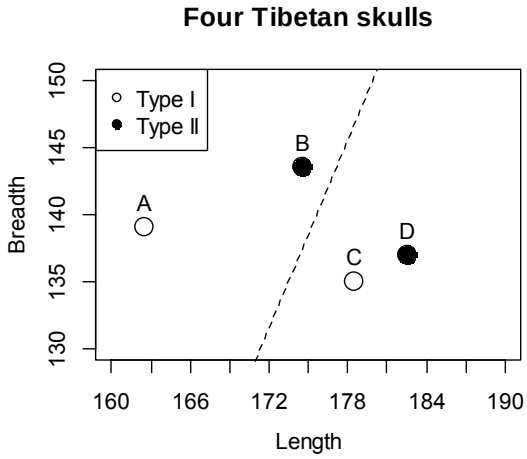


Figure 4: Scatter plot with LDF axis

This new axis *does* allow the two groups of cases to be separated. This is demonstrated in Figure 5, in which the four data points are projected onto the new axis. It may be observed that values less than zero are type I, while those greater than zero are type II.

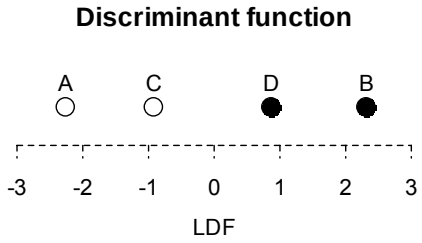


Figure 5: Cases projected against LDF axis

The values that have been plotted along this new LDF axis are those shown in the column headed `LDF` in Table 3. They were calculated using the simple formula shown in (1).

$$(1) \qquad z = 0.204 * L + 0.477 * B - 101.8$$

where L = `Length` and B = `Breadth`. The values 0.204 and 0.477 in this formula are called 'discriminant coefficients' (or 'discriminant loadings'). They were arrived at through a calculation in which the goal was to produce a linear function z for each of the four cases, such that the variance *within* each group was *minimized*, and the variance *between* the two groups was *maximized*. (The value −101.8 is simply a constant whose purpose is to position the zero point midway between the groups.)

When there are only two classes to be discriminated (as here), only one linear function is calculated. With two discriminating (or 'predictor') variables (as here), it has the general

form shown in (2). Comparing this to the actual function shown in (1), it may be observed that x_1 and x_2 are the variables (corresponding to `Length` and `Breadth`); a_1 and a_2 are the coefficients (corresponding to 0.204 and 0.477), which constitute weightings (or loadings) of the variables; and a_0 is a constant (corresponding to −101.8).

(2) $z = a_0 + a_1x_1 + a_2x_2$

Neither the number of groups nor the number of discriminating variables need be restricted to two, as they are in the simple example used here. Given n variables (instead of two) and two groups, the general form of the discriminant function z is as shown in (3), which now contains n+1 terms instead of the 2+1 (i.e. three) terms, a_0, a_1x_1 and a_2x_2, found in (2).

(3) $z = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \ldots + a_nx_n$

If the number of groups (G) is greater than two, additional discriminant functions are calculated up to a maximum of $G-1$. For example, with five groups (say, five L1s), there could be up to four functions, as in (4).

(4)
$$z_1 = a_{1,0} + a_{1,1}x_{1,1} + a_{1,2}x_{1,2} + a_{1,3}x_{1,3} + a_{1,4}x_{1,4} + \ldots + a_{1,n}x_{1,n}$$
$$z_2 = a_{2,0} + a_{2,1}x_{2,1} + a_{2,2}x_{2,2} + a_{2,3}x_{2,3} + a_{2,4}x_{2,4} + \ldots + a_{2,n}x_{2,n}$$
$$z_3 = a_{3,0} + a_{3,1}x_{3,1} + a_{3,2}x_{3,2} + a_{3,3}x_{3,3} + a_{3,4}x_{3,4} + \ldots + a_{3,n}x_{3,n}$$
$$z_4 = a_{4,0} + a_{4,1}x_{4,1} + a_{4,2}x_{4,2} + a_{4,3}x_{4,3} + a_{4,4}x_{4,4} + \ldots + a_{4,n}x_{4,n}$$

To sum up, the basic task of discriminant analysis is to calculate discriminant functions which together constitute a model (in n-dimensional space) that minimizes the variance within groups and maximizes the variance between groups.

2.2.1.2 LDA and the methodological framework

At this point it is worth noting the parallel between the method of discriminant analysis and the first two types of evidence called for by Jarvis (see §2.1.3):

- minimized within-group variance *equates to* Jarvis' intragroup homogeneity
- maximized between-group variance *equates to* Jarvis' intergroup heterogeneity

In other words, the method of discriminant analysis automatically provides two of the three kinds of evidence called for by Jarvis. This point will be discussed further in Chapter 6.

2.2.1.3 Cross-validation

In order to assess the accuracy of models constructed through LDA, they must be validated using data for which the correct classification is known. This should not be the same data that was used to construct the model, otherwise an overoptimistic estimate of the model's accuracy will result. The usual approach is therefore to partition the data into a 'training set', which is used to construct the model (or 'train' the classifier), and a 'test set', which is used to test it. Such 'cross-validation' (CV) is often performed ten times (so-called 10-fold CV) by dividing the data into ten test sets, each containing one-tenth of the cases. A model is constructed using the remaining nine-tenths and this is used to classify the cases in the test set. The process is repeated ten times, once for each test set, and the results are then collated. An alternative to 10-fold CV is leave-one-out cross-validation (LOOCV), in which the data is divided into the same number of partitions (n) as there are cases, with one case in each test set (and $n-1$ cases in each training set).

Either way, the grouping information for the training set is used to build the model which is then applied to the test set in order to 'predict' the grouping of the 'unknown' cases. The predictions are then compared to the known groupings in order to determine the accuracy of the model. The usual way to present the results is in the form of a 'confusion matrix'. In the study of Tibetan skulls the model was tested using LOOCV and the results were combined into the confusion matrix shown in Table 4.

	Predicted type	
Actual type	Type I	Type II
Type I	**14**	3
Type II	3	**12**

Table 4: Confusion matrix (Tibetan skulls data)

This shows that 14 of the 17 type I skulls were predicted correctly, and that 12 of the 15 type II skulls were predicted correctly. All in all, 26 of the 32 skulls were correctly identified, which gives a success rate of 26/32 = 0.81 (81%), and a corresponding error rate of 1 − 0.81 = 0.19 (19%). To determine whether this result is statistical, a null hypothesis is formulated stating that the type of skull bears no relation to its measurements. Since there are two possible choices, the probability of correctly predicting the type of any individual skull is 0.5

(50%). Using a simple binomial test in R (see §4.2.3) the null hypothesis can be tested as follows:

```
> binom.test( x=26, n=32, p=0.5 )

        Exact binomial test

data:  26 and 32
number of successes = 26, number of trials = 32, p-value = 0.0005351
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.6356077 0.9279238
sample estimates:
probability of success
                0.8125
```

The low *p*-value (< 0.01) allows the null hypothesis to be rejected and it may be concluded that the classification result is indeed statistical.

2.2.1.4 Feature selection

It is not always the case that every variable contributes to discriminating between groups, especially when the number of variables is very large. For this reason it is often desirable to identify an appropriate subset to use in the analysis. When the goal of the LDA is descriptive, irrelevant variables represent noise that confuses the interpretation of the results. When the goal is prediction, variables that play no useful role in the classification are a source of additional expense, in terms of cost, speed, etc. In both cases, such variables are best removed before constructing the model, through a process called 'feature selection'.

A number of approaches to feature selection are possible, the most popular being stepwise methods. These can operate in either a forward or a backward manner, or in a combination of the two. In a forward stepwise procedure, features are added to the model one step at a time. At each step the feature that contributes the most discriminatory power to the model (above a specified threshold for entry) is added, and the procedure continues until a cut-off point is reached, defined in terms of a certain number of features or the threshold for entry. A backward stepwise procedure starts off with all the features and removes them one by one depending on which one contributes the least discriminatory power to the model at each step. Again there is a cut-off point at which the procedure is halted.

Feature selection can be combined with cross-validation in one of two ways: either it can be 'embedded' within each fold of the CV – which typically results in different (but overlapping)

sets of features in each fold; or it can be performed 'up-front' and the same set of features used in each fold of the CV. Both methods are used in the present study.

2.2.1.5 Underlying assumptions of LDA

Figure 6 depicts the process flow of a typical discriminant analysis, showing the partitioning of the source data into a training set and a test set, the construction of a model with optional feature selection, the application of the model to predict the classes of cases in the test set, and the summary of the output in the form of a confusion matrix. When feature selection is performed the output includes a feature set which can be analysed in order to gain insight into which features contribute the highest degree of discriminating power to the model.

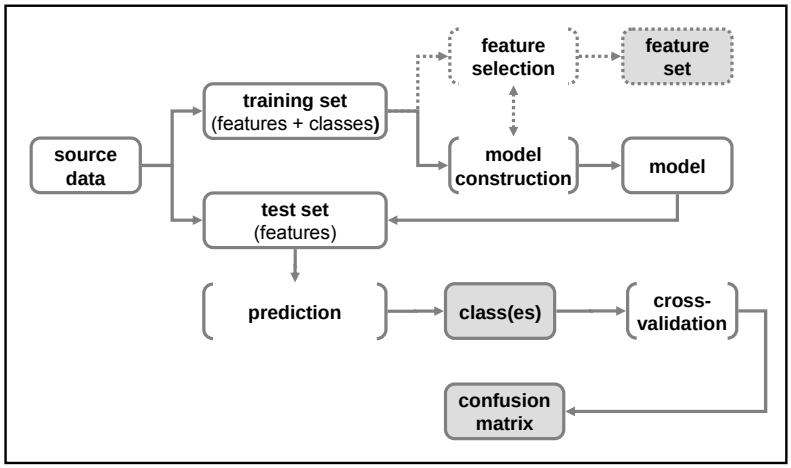


Figure 6: Discriminant analysis flowchart

Like all statistical methods, discriminant analysis is based on certain underlying assumptions about the data which must be satisfied in order for the method to work correctly. Klecka (1980) summarizes those assumptions as follows:

1. two or more groups: $G \geq 2$
2. at least two cases per group: $n_i \geq 2$;
3. any number of discriminating variables, provided that it is less than the total number of cases minus two: $0 < p < (N - 2)$;
4. discriminating variables are measured at the interval level;
5. no discriminating variable may be a linear combination of other discriminating variables;

6. the covariance matrices for each group must be (approximately) equal, unless special formulas are used;
7. each group has been drawn from a population with a multivariate normal distribution on the discriminating variables.

There is a lack of consensus in the literature regarding the ratio of cases to features. Jarvis (2012: 16) refers to a convention of having "at least 10 texts [i.e. cases] for every feature" but provides no references. Burns & Burns (2008: 591) state that group sizes should be "at least five times the number of independent variables." On the other hand, Venables & Ripley (2002: 331ff),[1] Everitt (2005: 142ff), Baayen (2008: 154ff) and Field *et al.* (2012: 738ff) all introduce LDA without stating requirements that go beyond those articulated by Klecka. According to Lachenbruch (1975: 17), "in general, the discriminant function performs fairly well with samples of moderate size" and no more than 3.5 cases are required per feature in order to be "within 0.05 of the optimum error rate." The extent to which the data used in the present study follows these assumptions is discussed in §4.5.1.

2.2.2 Applications of discriminant analysis in linguistics

Until now discriminant analysis does not appear to have been very widely used in linguistics, except in stylistics, and hardly at all in SLA. Larson-Hall (2010) makes no mention of LDA (nor of any other classification technique), despite covering equally sophisticated methods, such as multiple regression, and despite the fact that SPSS, the tool she uses, has excellent support for LDA (see §4.2.2). However there are some examples. The earliest (Mustonen 1965) makes an excellent introductory case study and is therefore described here in some detail, after which a number of later applications of LDA are listed, along with brief descriptions and bibliographic references.

2.2.2.1 Mustonen (1965)

One of the earliest applications of discriminant analysis in the field of linguistics was Seppo Mustonen's 1965 paper *Multiple discriminant analysis in linguistic problems*. The paper's purpose was to "indicate the possibilities of applying statistical multivariate analysis to some linguistic problems" and it did so by "teaching the computer to decide to which language a

[1] The R package MASS that is used in this project (see §4.2.3 and §4.5.4) is based on this book.

given word most probably belongs" (Mustonen 1965: 37). The author cautions that the example was planned "only for fun" and was therefore not to be taken too seriously, except as an example that can be extended "to have practical linguistic applications."

The source data in this experiment was a sample of 900 words – 300 each from English, Swedish and Finnish – chosen at random from dictionaries. For each word, 43 quantitative variables were computed, one for each of the properties shown in Table 5. The numbers in the right-hand column headed 'Ex.' give the value of each variable for the example word 'always'. This word consists of five different letters, including two different vowels, and has two syllables (according to the rules of Finnish), one of which has a length of two letters; there are two letters in the first syllable and four in the last, one syllable of type VC, etc.

#	Variable	Ex.	#	Variable	Ex.
1	different letters	5	23	letter F	0
2	different vowels	2	24	letter G	0
3	syllables	2	25	letter H	0
4	1-letter syllables	0	26	letter I	0
5	2-letter syllables	1	27	letter J	0
6	3-letter syllables	0	28	letter K	0
7	letters in first syllable	2	29	letter L	1
8	letters in last syllable	4	30	letter M	0
9	syllables of type VC	1	31	letter N	0
10	syllables of type CV	0	32	letter O	0
11	syllables of type VCC	0	33	letter P	0
12	syllables of type CVV	0	34	letter R	0
13	syllables of type CVC	0	35	letter S	1
14	twin (double) letters	0	36	letter T	0
15	diphthongs	1	37	letter U	0
16	first letter (V=0, C=1)	0	38	letter V	0
17	last letter (V=0, C=1)	1	39	letter W	1
18	letter A	2	40	letter Y	1
19	letter B	0	41	letter Å	0
20	letter C	0	42	letter Ä	0
21	letter D	0	43	letter Ö	0
22	letter E	0			

Table 5: Set of features in Mustonen (1965)

These variables constitute the 43 features (or predictor variables) used in the discriminant analysis. There were 900 cases (i.e., the words), and three groups (EN, SW and FI). Other than that, as Mustonen points out, "the computer was not given any other special information to improve accuracy of discrimination (*sic*)":

> For instance, no deterministic rules for identifying the language were given, i.e., nobody told the machine that a Finnish word never ends with two consonants, that the English alphabet does not contain Å, Ä, Ö, etc. Thus all the information the computer could use was restricted to 900 words described by 43 quantitative variables (Mustonen 1965: 39).

The input data may be envisaged as a matrix of 900 rows (one per word) and 43 + 1 = 44 columns (one per quantitative variable, plus the grouping variable, Language). The data was fed into a computer program which then computed two discriminant functions of the form shown in (1) (cf. equation (3) on page 18):

(1) $z = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \ldots + a_{43}x_{43}$

The values of the coefficients are shown in Table 6: column I contains the coefficients for the first function and column II contains the coefficients for the second. A portion of the first function is shown in (2) and the value of this function for the example word 'always' is shown in (3).

(2) $z_1 = 0.021x_1 + 0.028x_2 + 0.414x_3 - 0.467x_4 + \ldots - 0.043x_{43}$

(3) $z_{\text{always}} = (0.021 \cdot \mathbf{5}) + (0.028 \cdot \mathbf{2}) + (0.414 \cdot \mathbf{2}) - (0.467 \cdot \mathbf{0}) + \ldots - (0.043 \cdot \mathbf{0})$

Mustonen points out how the first (and stronger) discriminant function serves to separate Finnish from English and Swedish: high negative loadings indicate a tendency towards Finnish, as for syllables of one or two letters (–.467 and –.232, respectively), and the letters H, J, K, L, M, N, P, S, T, U, V and Ä. The second function serves to discriminate between English and Swedish: here, high positive loadings indicate a tendency towards Swedish, and thus, on the basis of this data, we may observe that letters such as F, J, K and Å are more frequent in Swedish, while C and W are more typical of English.

#	Variable	I	II	#	Variable	I	II
1	different letters	.021	–.024	23	letter F	–.078	.283
2	different vowels	.028	.067	24	letter G	–.011	.184
3	syllables	.414	–.209	25	letter H	–.215	.032
4	1-letter syllables	–.467	.176	26	letter I	–.089	–.090
5	2-letter syllables	–.232	.073	27	letter J	–.224	.208
6	3-letter syllables	–.011	.058	28	letter K	–.261	.230
7	letters in first syllable	.056	–.020	29	letter L	–.142	.126
8	letters in last syllable	.078	–.049	30	letter M	–.127	.191
9	syllables of type VC	.076	–.073	31	letter N	–.167	.143
10	syllables of type CV	–.038	.017	32	letter O	–.071	–.093
11	syllables of type VCC	–.084	.167	33	letter P	–.146	.132
12	syllables of type CVV	–.160	.137	34	letter R	–.099	.127
13	syllables of type CVC	–.161	–.009	35	letter S	–.164	.149
14	twin (double) letters	–.033	–.004	36	letter T	–.132	.171
15	diphthongs	–.247	–.134	37	letter U	–.107	–.121
16	first letter (V=0, C=1)	–.043	–.120	38	letter V	–.175	.155
17	last letter (V=0, C=1)	.071	.054	39	letter W	.060	–.299
18	letter A	–.055	–.033	40	letter Y	–.095	–.074
19	letter B	.009	.187	41	letter Å	–.025	.408
20	letter C	.076	–.163	42	letter Ä	–.126	–.002
21	letter D	–.073	.173	43	letter Ö	–.043	.166
22	letter E	–.001	–.122				

Table 6: Discriminant loadings in Mustonen (1965)

In order to test the accuracy of the statistical model described by the discriminant functions, Mustonen took a new random sample of 300 words, 100 from each language, and had the computer "predict" which language each one belonged to. Discriminant function values were calculated for each new word, which was then classified according to its "distance" from the means of each of the three languages. In Figure 7 the *x* and *y* axes correspond to the two discriminant functions, z_1 and z_2, and the small circles represent the positions on the "map" of (some of) the cases in the test set. Finnish words are shown in green, Swedish in red and English in blue. Words that belong to two or more languages (e.g. *anemone*) are in black. The large circles represent the means of the three language groups and they divide the coordinate space into three sectors. It may be observed that words located to the West are typically Finnish, while those in the North-East are typically Swedish and those in the South-East are typically English.

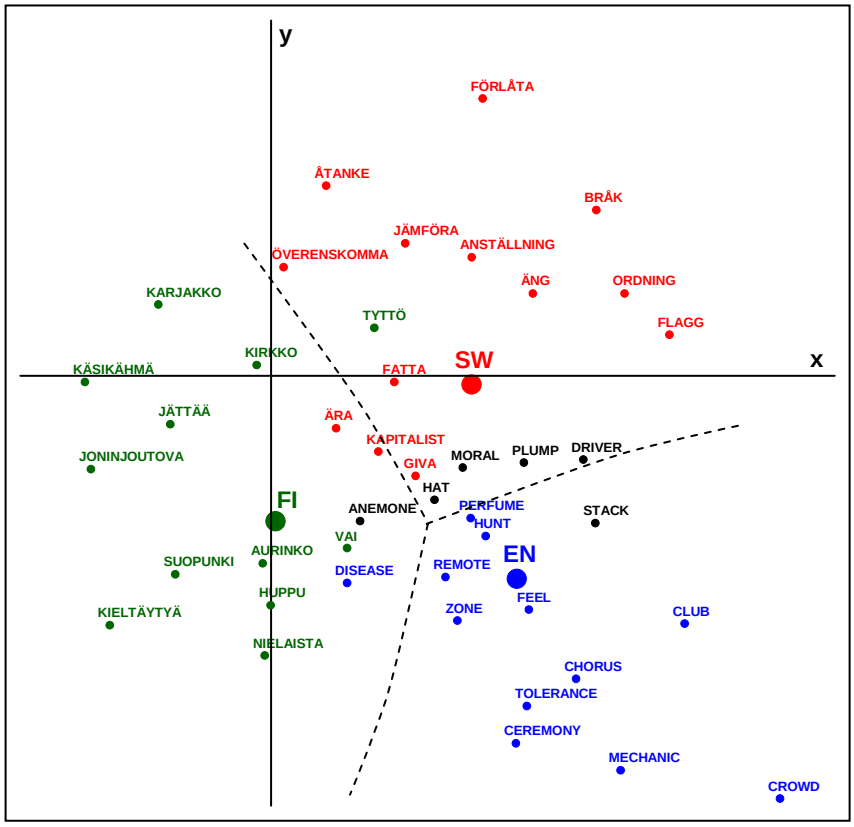


Figure 7: Scatter plot of test cases in Mustonen (1965)

The classification is not perfect (e.g. FI *tyttö* is classified as SW, and SW *ära* as FI), but neither is this to be expected. As Mustonen writes:

> One must remember that complete accuracy of classification is hardly achieved in applications like this where the different groups (as languages here) are not disjoint but have a great many members (words) which are, or at least could be, common to several of those groups (Mustonen 1965: 42).

The overall success rate for the classification is shown in Table 7. This indicates a high success rate for Finnish (91%) and somewhat lower rates for English (59%) and Swedish (79%). The overall success rate is 76.3% (alternatively, the error rate is 33.7%).

		Predicted language			
		EN	**SW**	**FI**	**Total**
Actual language	**EN**	**59**	28	13	100
	SW	11	**79**	10	100
	FI	1	8	**91**	100

Table 7: Confusion matrix (Mustonen 1965)

Using R to perform a binomial test (with a chance probability of 0.333, since there are three groups) leads to rejection of the null hypothesis that there is no difference between EN, SW and FI. The results are therefore statistical ($p < 2.2e\text{-}16$), which means that the discriminant analysis has found significant structure in the data.

```
> binom.test( 59+79+91, 300, 0.333 )

        Exact binomial test

data:  59 + 79 + 91 and 300
number of successes = 229, number of trials = 300, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.333
95 percent confidence interval:
 0.7110981 0.8102844
sample estimates:
probability of success
             0.7633333
```

Of course, this kind of classification cannot compete with deterministic methods based on a strict set of rules (assuming it is possible to formulate such a strict set of rules, which in the present case it is not) but, as Mustonen points out:

> deterministic classification is not very interesting, since it hardly can reveal anything new about the subject. All the information about the rules of classification must be given by the investigator. In the method based on Discriminant Analysis no rules for classification are given in advance (Mustonen 1965: 44).

This comment underlines the importance of LDA as tool for discovering new facts.

2.2.2.2 Other case studies

While Mustonen's pioneering paper did not lead to a major breakthrough in the use of LDA in linguistics, there have been some applications, including:

- Bellinger (1979) – an analysis of the pattern of change in mothers' speech.
- Marckworth & Baker (1980) – a discriminant function analysis of co-variation of 36 syntactic devices (sentence type, focus phenomena, verb structure, conjoining and embedding, etc.) in five prose genres.
- Fletcher & Peters (1984) – an exploratory study of language impairment in children.
- Fox (1991) – a discriminant analysis of yes-no questions in Quebec French.
- Nicoladis (1994) – a PhD thesis on code-mixing in French-English bilingual children.

- Ganschow & Sparks (1996) – a psychological study of anxiety about language learning among high school women, which examines the relationship between three levels of anxiety (high, average, low) and nine measures of NL skill and FL aptitude.
- Peng & Hengartner (2002) – a quantitative analysis of the literary styles of nine authors, including Shakespeare, Dickens, Kipling and London.

2.3 The detection-based approach

The impetus for the current project came with the publication in March 2012 of the book *Approaching Language Transfer through Text Classification* (Jarvis & Crossley 2012). Subtitled "explorations in the detection-based approach", this collection of five closely related studies reports on the first systematic attempt to apply discriminant analysis to transfer research. The cases in each of these studies are learner texts, the features are various frequency data and other metrics, the groups are the learners' L1s, and the goal is to 'predict' the L1 background of the learners on the basis of the lexical features exhibited in their texts.

The Jarvis & Crossley studies were informed by earlier work on text mining and in particular by seven studies that focus on L1 detection. This section gives a brief overview of those "pioneer" studies and then presents the Jarvis & Crossley studies in more detail, in particular the one on which the present study is modelled.

2.3.1 Pioneer studies

A fairly detailed overview of previous studies on L1 detection is given by Jarvis in his introductory chapter (Jarvis 2012: 20–27). This section therefore contains only a brief summary of each study in chronological order.

2.3.1.1 Mayfield Tomokiyo & Jones (2001)

The primary goal of the first study to use text classification techniques in SLA was to test the ability of a Naïve Bayes classifier (an alternative to LDA) to distinguish between native and non-native speakers of English using transcripts of spontaneous speech. A related goal was to test whether such a classifier could distinguish between two groups of six Chinese- and 31 Japanese-speaking learners of English. The features used in the study were frequencies of

(1) individual words (1-grams),[1] (2) word sequences (2- and 3-grams), (3) word classes (POS 1-grams), (4) sequences of word classes (POS 2- and 3-grams), and (5) word sequences in which nouns were replaced by their word class (as in *on the SB*). Cross-validation was performed by selecting 70% of the texts at random as the training set and using them to classify the remaining 30%. This procedure was performed 20 times and the results averaged. The accuracy obtained varied from 74% to 100% depending on the combination of L1s (ZH, JA, EN) and the kinds of features used. Greater accuracy was achieved using word tokens than with POS tokens, and when distinguishing between ZH and native speakers. As Jarvis points out, the level of accuracy is "quite phenomenal, but the fact that there were only two L1 groups to distinguish between, and the fact that the two L1 groups were so small and unevenly balanced, casts some doubt on the generalizability of the results" (p. 21).[2]

2.3.1.2 Jarvis et al. *(2004)*

This study is the direct precursor of the first Jarvis & Crossley study (see §2.3.2.1). Its purpose was to determine how well an LDA classifier could predict the L1 backgrounds of 446 adolescent EFL learners from five different L1 backgrounds: DA, FI, PO, SP and SW. The source data consisted of written narrative descriptions of a segment of the Charlie Chaplin film *Modern Times*, and the features fed into the classifier were the 30 most frequent words used by each L1 group, for a total of 53 words. No automatic feature selection was performed and validation was by 10-fold CV. The overall accuracy rate achieved in the classification was 81%. However, Jarvis, Castañeda-Jiménez & Nielsen subsequently concluded that their CV procedures were "overly simplistic and somewhat positively biased" (p. 22), which is why they chose to re-implement the study, as discussed in §2.3.2.1, below.

---

[1] The term 'gram' is widely used in the fields of computational and corpus linguistics to denote a contiguous sequence of *n* items from a portion of text or speech; the items in question can be phonemes, syllables, letters, words, etc. depending on the application. A word unigram (1-gram) is a single word; a word bigram (2-gram) is a sequence of two words, etc. A POS *n*-gram is a sequence of *n* words expressed in terms of their word class (or part of speech), e.g. *on the road* is an instance of the POS 3-gram [PRP ART SB].

[2] Bare page references such as (p. 21) are references to Jarvis & Crossley (2012).

2.3.1.3 Koppel et al. *(2005)*

Koppel *et al.*'s study used 1,290 texts from the International Corpus of Learner English (ICLE) written by learners from five L1 groups: BU, CS, FR, RU and SP. They used a Support Vector Machines classifier (another alternative to LDA) and a total of 1,035 different features: 400 function words, 200 frequent letter sequences (or letter *n*-grams), 185 error types and 250 rare POS 2-grams. A 10-fold CV was performed, but without feature selection and classification rates of up to 80% were achieved.

2.3.1.4 Estival et al. *(2007)*

The goal of this study was to test many different classifiers on a number of tasks, including the prediction of L1, age, gender, level of education, country of origin, etc. Using the WEKA toolkit (Witten & Frank 2005), several machine learning algorithms were tested. The data consisted of 9,836 emails written by 1,033 people from three L1 backgrounds (EN, AR and SP), and the features fed into each classifier included word-length indices, and the relative frequencies of punctuation, function words, POS categories, paragraph breaks and HTML tags. A number of feature selection methods were tested and the best results (84% accuracy) were obtained using the Random Forest classifier using an unspecified number of features that passed an information-gain criterion "after features that were used by speakers of only one L1 background were removed" (p. 23).

2.3.1.5 Tsur & Rappoport (2007)

This study replicates Koppel *et al.* (2005) in many respects, not least its selection of L1s and use of SVM with 10-fold CV. However it was based on a different sample of 1,190 texts from ICLE and a slightly different set of features. For purposes of comparison a baseline classifier was implemented in which each document was represented by the normalized frequencies of the (de-capitalized) letters it contains, and this resulted in an accuracy rate of 46.8%. Classification based on the 200 most frequent 2-grams in each sub-corpus had a 65.6% success rate, while a corresponding experiment using 3-grams achieved 59.7%, and one based on 460 function words achieved 66.7%. In a further test, the 84 2-grams with the greatest "separating power" produced a result of 61.4%, a drop of just 4% compared to the full 200 2-gram test. As Jarvis points out, Tsur & Rappoport's results are statistically lower than those achieved by Koppel *et al.*, although still far higher than chance. As possible

explanations, Jarvis suggests the smaller sample in the later study, differently composed sets of features, and the possibility that different classifier settings were used.

2.3.1.6 Wong & Dras (2009)

This is another follow-up study to Koppel *et al.* (2005), using the same classifier and the same five L1s, but with the addition of ZH and JA and a smaller sample of texts (95 per L1, 665 in total). A simple split CV was used, with 490 texts in the training set and 175 texts for testing. The study was in two parts. The first part investigated whether syntactic errors correlate with L1. Three types of error were explored: subject-verb disagreement, noun-number disagreement and the misuse of determiners. The last of these turned out to be highly statistical, whereas subject-verb disagreement was only marginally so, and noun-number disagreement not at all. After tuning, these errors yield a small but statistical improvement in accuracy (24.6%) compared to the baseline of 14.3% (i.e. 1/7). The second part of the study used various lexical features (function words, letter *n*-grams and POS *n*-grams). Four types of classification were performed for each set of features, both individually and also in combination with each other and the previously mentioned syntactic error values. The best results (of 73.7%) were obtained by combining function words with POS *n*-grams and were not affected either by tuning or the inclusion of syntactic errors.

2.3.1.7 Jarvis (2011)

The primary purpose of this study was to test a number of different classifiers in order to find out which one offered the best performance when attempting to detect a learner's L1 background. The data used in the study were identical to those used in the study by Jarvis and Paquot, discussed in §2.3.2.2 below, i.e. frequency data for 722 *n*-grams extracted from 2,033 ICLE texts representing 12 L1s. A total of 20 classifiers were tested using 10-fold CV. The best results (53.6% accuracy) were obtained using LDA with stepwise feature selection. Three other methodologies performed almost as well using the full set of features.

2.3.2 Jarvis & Crossley (2012)

Notwithstanding the ground-breaking work presented in the preceding section, the five case studies published in Jarvis & Crossley (2012) represent the first *systematic* application of the detection-based approach to transfer research. All of these studies use discriminant analysis

to analyze texts written by intermediate to advanced level learners of English, and most of them use data from ICLE. They each attempt to detect the authors' L1s on the basis of characteristic linguistic features and, in the process, identify those features that are the best L1 predictors. Three types of feature are employed in these studies – lexical choices, stylistic metrics and error types, and the success rates achieved – despite varying according to the number of languages and the combination of features – consistently exceed chance probability ($p < 0.01$).

The first two studies investigate lexical style. In Chapter 2, Jarvis, Castañeda-Jiménez & Nielsen (hereafter Jarvis *et al.*) replicate in a more rigorous fashion their 2004 study (see §2.3.1.2). This is the study that provides the model for the present work. Then, in Chapter 3, Jarvis & Paquot broaden the investigation from five L1s to 12, and to larger numbers of both 1-, 2-, 3- and 4-grams, using data from ICLE. Since the present study also focuses on lexical style, both of these studies are presented in some detail below, followed by brief descriptions of the three remaining studies.

2.3.2.1 Jarvis et al. (53 1-grams)

Jarvis *et al.* (2012) set out to determine "whether feature vectors (i.e. lexical styles) made up of roughly 50 of the most frequent words in a learner corpus serve as successful indicators of learners' L1s" (p. 35). The same source data was used as in the 2004 study: i.e. descriptions of a Chaplin movie written by 446 learners whose L1 backgrounds (and home countries) were DA (Denmark), FI (Finland), PO (Brazil), SP (Mexico) and SW (Finland). Two pairs of closely related languages (North Germanic DA and SW, and West Iberian PO and SP) were chosen deliberately in order to make the classification task more challenging and test "the sensitivity of the classifier and the uniqueness of L1-related lexical styles" (p. 45).

The texts varied in length from 15 to 608 words ($\bar{x} = 218$; $s = 106$). The learners' ages ranged from 11 to 18 and they had been exposed to from 2-13 years of English tuition. Their proficiency levels were assumed to range from A2 to C1 on the CEFR scale "with very few cases that qualify as A1 or C2" (p. 46). Jarvis *et al.* state that they "deliberately recruited learners representing a wide range of L2 proficiency in order to test the sensitivity of our

classifier, and more importantly to test whether clearly measurable L1-related characteristics of learners' lexical styles remain consistent across proficiency levels" (p. 47).

The investigation was thus based on 446 cases, distributed (somewhat unevenly) across 5 groups, and the features used were frequencies for the 53 1-grams shown in Figure 8. This list was arrived at by pooling the 30 most frequent words from each L1 group. A series of one-way ANOVAs showed that 45 of the 53 variables differed statistically across groups (those that did not are shown in parentheses).

> *a, and, away, be, bread, bus, (but), car, Chaplin, Charlie, come, do, down, (eat), (for), get, girl, go, have, he, (house), imagine, in, into, it, lady, (live), man, of, one, out, police, policeman, run, say, see, she, sit, so, steal, take, tell, that, the, then, there, (they), to, (up), when, (who), with, woman*

Figure 8: The 53 features used by Jarvis *et al.*

LDA was performed using SPSS (see §4.2.2) version 17.0 with L1 as the grouping variable, a setting of equal prior probabilities, 10-fold CV with embedded stepwise feature selection, and default values for feature entry ($p < 0.05$) and removal ($p > 0.10$).[1] Following this, SNK post-hoc tests were run to reveal homogeneity subsets, as described below.

Jarvis *et al.* exploited both the predictive and the descriptive aspects of discriminant analysis and their results can be discussed accordingly. As far as prediction is concerned, accuracy rates ranged from 63.6% to 89.7% with an overall rate of 76.9%. This is "considerably and significantly higher than the chance level of 20% ... It is also significantly and substantially higher than the baseline accuracy of 31.4% … which is the accuracy that would be attained if all texts were classified as belonging to the largest L1 group" (p. 53).

The confusion matrix shown in Table 8 reveals rates of around 65% for DA and FI and well over 80% for the other L1s. Misclassification occurs mostly between related languages, i.e. SW for DA (21.7%) and *vice versa* (8.6%), and SP for PO (15%) and *vice versa* (5.2%). The surprising number of DA texts that are misclassified as FI (11.7%) is explained by the fact

[1] Since SPSS does not support the embedding of feature selection in the folds of the cross-validation procedure, Jarvis *et al.* implemented the functionality using Perl scripts (p. 51). See Appendix E for more information about SPSS settings.

that several DA participants "either did not understand the task correctly or did not have the required L2 proficiency to complete it successfully" (p. 64). The even higher proportion of FI misclassified as SW (20%) is shown to be due to the influence of L2 Swedish learned at school by L1 Finnish speakers. These results are commensurable with those obtained in previous studies that used five groups (Koppel *et al.* 2005; Tsur & Rappoport 2007), despite (or perhaps because of) using far fewer features.

		Predicted L1					
		DA	**FI**	**PO**	**SP**	**SW**	total
Actual L1	**DA**	**63.3**	11.7	1.7	1.7	21.7	100.0
	FI	5.7	**66.4**	0.7	7.1	20.0	100.0
	PO	0.0	0.0	**83.3**	15.0	1.7	100.0
	SP	0.9	0.9	5.2	**89.7**	3.4	100.0
	SW	8.6	4.3	0.0	4.3	**82.9**	100.0

Table 8: Confusion matrix (Jarvis *et al.*)

As far as the descriptive function of LDA is concerned, the use of feature selection enabled Jarvis *et al.* to identify 36 features that contributed statistically to the discriminating model. These were the 34 that were selected in all 10 folds of the cross-validation, plus *go* and *she*, which were selected in a majority of the five models whose accuracies were higher than 80%. Student-Newman-Keuls (SNK) post-hoc tests were run on these variables in order to reveal what are known as "homogeneity subsets". These are subsets of the set of five L1 groups in which the values of a given variable show no statistical difference. 18 of the 36 features resulted in subsets that overlapped one another (and are therefore somewhat harder to interpret); the remaining 18 are shown in Table 9.

Each entry in this table shows how the five L1s group themselves with respect to each feature. For example, FI is clearly separated from the other groups by a relatively infrequent use of the words *a, be* and *the*, and a relatively frequent use of the words *Chaplin* and *go*. The word *a* also helps separate PO, but the other words that separate PO (*bread, girl, then* and *woman*) are different from the ones that isolate FI (p. 59-60).

These 18 words are thus the best predictors of L1 group membership. The crucial question from the perspective of transfer research is "whether these L1-specific word choice patterns are due to direct L1 influence, or do they simply reflect differences in the L1 groups' cultural

Feature	Homogeneity subsets	Feature	Homogeneity subsets
a	FI < PO < (SW, SP, DA)	*go*	(PO, SW, DA, SP) < FI
away	(SP, PO) < (DA, SW, FI)	*imagine*	(DA, FI, SW) < (PO, SP)
be	FI < (SP, SW, DA, PO)	*into*	(PO, FI, SP) < SW < DA
bread	PO < (DA, SW) < (SP, FI)	*police*	(SW, DA, FI) < (PO, SP)
bus	(DA, FI, SW) < (PO, SP)	*policeman*	SP < (SW, PO, DA, FI)
Chaplin	(SP, DA, PO, SW) < FI	*the*	FI < (SW, DA) < (PO, SP)
come	(PO, SP) < (FI, SW, DA)	*then*	PO < (DA, SW) < (SP, FI)
do	(DA, SW) < (PO, SP, FI)	*there*	(PO, SP) < (FI, SW) < DA
girl	PO < SP < (SW, DA, FI)	*woman*	(DA, SW, FI, SP) < PO

Table 9: Homogeneity subsets (Jarvis *et al.*)

and education backgrounds" (p. 60). Answering that question requires evidence of the third kind called for by Jarvis (2000), i.e. cross-language congruity, and this entails contrastive analysis. Jarvis *et al.* reserve such analysis for future research, but offer a few interesting observations (p. 60-62), such as:

- FI low use of *a* and *the* probably reflects the lack of articles in Finnish (Karlsson 2008: 7). Jarvis *et al.* suggest furthermore (p. 61) that the "much wider distribution" of definite articles in PO and SP than in DA and SW can account for the pattern exhibited by *the*.
- FI high use of *Chaplin* and *girl* has its complement in low use of *he* and *she* by the same group (not shown in Table 9). A likely contrastive explanation is that the absence of pronominal gender in FI (Karlsson 2008: 203) leads to underuse of pronouns and greater use of (referential) nouns in order to distinguish between the two main characters – Chaplin and the woman.
- PO and SP low use of *away* and *come* can be explained by the verb-framed nature of Romance languages compared to the satellite-framed nature of the others (Talmy 1985, 2000; Slobin 2004).[1]

[1] Jarvis *et al.* note the existence of direct counterparts to the phrasal verb 'run away' in DA and SW but do not explain their observation in terms of verb-framing.

2.3.2.2 Jarvis & Paquot (722 n-grams)

Jarvis & Paquot (2012) represents a broadening of the investigation into lexical transfer carried out by Jarvis *et al.* Their study has two parts, referred to here as Parts 1 and 2, both of which deal with *n*-grams of varying sizes, from individual words (1-grams) to multiword sequences (2-, 3- and 4-grams). Those portions of Parts 1 and 2 that deal with 1-grams are most pertinent to the present work and are therefore discussed in some detail. This is followed by a summary of the 'polygram' portions of the study.

Jarvis & Paquot aimed to find out whether 1-grams remain effective discriminators of learners' L1 backgrounds "even when (a) the number of L1 groups is increased substantially beyond five, (b) the texts are longer and reflect somewhat higher levels of proficiency and (c) the texts represent a range of open-ended argumentative topics rather than involving controlled narratives" (p. 72). Their investigation used texts from ICLE written by 2,033 learners from 12 L1 backgrounds (BU, CS, NL, FI, FR, DE, IT, NO, PO, RU, SP and SW) and representing a range of topics and task conditions.[1] The learners were mostly university undergraduates in their twenties (Granger *et al.* 2009: 7) and their proficiency levels were expected to range B1 to C2 (p. 78).[2] Their texts ranged from 500 to 1,000 words in length.

The features used in the discriminant analysis were the 200 most frequent 1-grams in the data that were not "prompt-induced" (see §4.4.3). Once again SPSS 17.0 was used with L1 as the grouping variable and equal prior probabilities. Part 1 of the analysis used 10-fold CV with all 200 features submitted using the 'Enter' method in SPSS – i.e., with no feature selection; Part 2 used 10-fold CV with embedded feature selection and strict values for feature entry ($p < 0.01$) and removal ($p > 0.05$).

Prediction accuracy rates of 53.0% and 49.9% were achieved for 1-grams in Parts 1 and 2 respectively, thereby confirming Jarvis *et al.*'s contention that embedded feature selection is to be preferred "in order to avoid positive bias (i.e. overly optimistic results)" (p. 51). This is

[1] Four other L1 groups found in ICLE (Japanese, Turkish, Tswana and Mandarin) were not included as these consist of "a high proportion of lower proficiency texts" (p. 78).

[2] Note, however, the unevenness reported by Granger *et al.* (2009: 11) based on a random sample of 20 essays from each of the 16 sub-corpora in ICLE. For example, all 20 SW were rated 'advanced' (C1 or C2), whereas only 40% of the SP were in this category, the rest were rated 'intermediate' (B1 or B2).

"significantly and substantially higher than the chance level of 8.3% … and also significantly and substantially higher than the baseline accuracy of 14.2% … which is the accuracy that would be attained if all texts were classified as belonging to the largest L1 group" (p. 84). As in the study by Jarvis *et al.*, some L1s were identified more accurately than others and misclassification patterns again pointed to genetic affiliation and L2 influence as the main factors (e.g. NO for SW and SW for FI, respectively). No feature selection was performed in Part 1 and Jarvis & Paquot only report the number of features selected in Part 2 without stating what they were or performing any form of contrastive analysis.

The remaining portions of the Jarvis & Paquot study show that 'polygrams' (i.e. *n*-grams where $n > 1$) function less well as L1 predictors than 1-grams, with statistically lower prediction rates being achieved in Part 1 for 2-grams (39.5%), 3-grams (31.2%) and 4-grams (22.0%). However, these results were still statistically higher than chance or baseline rates. Best results (53.6%) were obtained in Part 2 using a combination of 722 1-, 2-, 3- and 4-grams.

2.3.2.3 Crossley & McNamara (Coh-Metrix indices)

In Chapter 4 (Crossley & McNamara 2012) the focus is on stylistic rather than lexical characteristics of the learner texts. They restrict themselves to four L1s (CS, SP, DE and FI) and use 10-fold CV with embedded feature selection. Their feature set consists of 19 indices of textual cohesion, lexical sophistication, syntactic complexity and conceptual knowledge generated by the computational tool Coh-Metrix (McNamara & Graesser 2011).

The study shows that such indices "can significantly predict group membership" (p. 120). Their overall classification rate is 67.6% and the best results are achieved for DE, with 88% precision and 75% recall.[1] The corresponding figures for the other L1s are CS 62%–74%; FI 52%–59% and SP 58%–52%. The most significant differences concerned (1) word concreteness, with DE writers tending to use the most concrete words and SP writers the least concrete

[1] Recall is the number of correctly predicted L1s as a proportion of actual L1s and thus the value that corresponds to the accuracy rates reported elsewhere; precision is the number of correctly predicted L1s as a proportion of all predictions for this L1, including false positives.

words, i.e. (SP, FI) < CS < DE; (2) "word imagability",[1] which exhibited the same pattern; and (3) motion verbs, where the pattern was (SP, CS) < FI < DE. A brief discussion of linguistic features in relation to the manner in which they help characterize writers from different language backgrounds makes interesting reading, but has the limitation, acknowledged by the authors, that it does not include any form of contrastive analysis and therefore does not provide the third kind of evidence called for by Jarvis in his methodological framework.

2.3.2.4 Bestgen et al. (Error types)

The focus in Chapter 5, by Bestgen, Granger and Thewissen (Bestgen *et al.* 2012), shifts to error patterns. The data was a sub-corpus of ICLE that had been exhaustively error-tagged and comprised 223 texts from three L1 groups: FR, DE and SP. The features submitted to the classifier were the relative frequencies of 46 error types spread across seven broad categories: formal errors, grammatical errors, lexical errors, lexico-grammatical errors, punctuation errors, style errors and errors involving redundant, missing or incorrectly ordered words.[2] The analysis was performed using LOOCV with embedded stepwise feature selection and the model was able to correctly predict the L1 of 65% of the texts.

Of the 46 features submitted to the classifier, only 12 exhibited statistical differences across L1 groups. Seven of these served to discriminate SP from the others, three were closely associated with DE and two with FR. The authors conclude that it is possible to successfully discriminate the learners' L1 backgrounds on the basis of error types. However, they also acknowledge that different levels of proficiency may have played a part in the discriminant analysis. A closer examination of the facts, based on CEFR level ratings (see §3.4 below), showed statistical differences. After conversion to numeric values, SP exhibited a mean proficiency score of 1.44, statistically lower than FR 2.64, which in turn was statistically lower than DE 3.03.

This prompted a repeat investigation using a sample that was controlled for proficiency level. 30 FR and 30 DE texts that were all assessed as C1 were subjected to the same analysis using the same set of features. 75% were correctly classified, statistically more than chance, and

[1] This concept is not defined by the authors, but Graesser *et al.* (2004) define imageability as "how easy it is to construct a mental image of the word in one's mind, according to human ratings."

[2] For a complete description of the error tags used in this study see the authors' appendix (p. 150–153).

three types of error[1] were found to be particularly good discriminators. The authors caution against assuming that these necessarily result from language transfer; other factors (and not just proficiency) could be at work, such as different English teaching methods (p. 139–140).

2.3.2.5 Jarvis, Bestgen et al. (Combined features)

The final study (Jarvis, Bestgen *et al.* 2012), co-authored with Crossley, Granger, Paquot, Thewissen and McNamara, investigates the comparative and combined contributions of *n*-grams, Coh-Metrix indices, and error types. The same texts were used as in the study by Bestgen *et al.*, along with features from all three of the ICLE-based studies (i.e. Jarvis & Paquot's 722 *n*-grams, Crossley & McNamara's 19 Coh-Metrix indices, and Bestgen *et al.*'s 46 error categories). The analysis was run with LOOCV and embedded stepwise feature selection.

When the three groups of features were used separately, the best results (65.5%) were achieved with error categories, followed by Coh-Metrix indices (64.1%) and *n*-grams (63.2%). However, combining the three kinds of feature produced an even better result of 79.4%. Particularly remarkable, as Jarvis notes (p. 29), is the fact that this result was achieved after feature selection had chosen just 22 features; of these, seven were error types, six were Coh-Metrix indices, and nine were *n*-grams (p. 167).

To find out how important error types were to the analysis, a final test was performed using just *n*-grams and Coh-Metrix indices. This resulted in an L1 classification accuracy of 67.7%, which is statistically lower than the 79.4% achieved when error types were included. This leads the authors to conclude that error categories have an important role to play in this kind of research and that more work should therefore be done on the automatic identification and tagging of errors in learner texts.

[1] QL (Punctuation, Lexical), GNN (Grammar, Noun Number) and FS (Form, Spelling) (Bestgen *et al.*:139).

3. Data sources

Every application of the detection-based approach to transfer research to date has been based on English interlanguage texts. A central purpose of the present study was to try out the same approach using texts written in Norwegian interlanguage. This is the major point of difference between the present work and earlier work, and it could potentially lead to quite different results. This chapter therefore provides a detailed description of the data on which the present study was based.

3.1 The ASK corpus

The data comes from Norsk andrespråkskorpus (ASK), a learner corpus compiled under the direction of Kari Tenfjord at Bergen University (Tenfjord 2007; Tenfjord *et al.* 2009). ASK consists of Norwegian interlanguage texts written by adult learners of Norwegian L2 as responses to two officially recognized language tests. The tests were designed to measure language proficiency at two different levels: *intermediate*, corresponding to what is needed "to cope in most situations encountered at the workplace and in everyday life", and *advanced*, a pre-academic university entrance test (Carlsen 2012: 169). Following Carlsen's usage, the two tests will be hereafter referred to as follows:

- **IL test** (Intermediate) *Språkprøven i norsk for voksne innvandrere*
- **AL test** (Advanced) *Test i norsk – høyere nivå*[1]

The main corpus consists of 1,736 texts written by learners from ten different language backgrounds and a rich set of metadata, covering factors such as proficiency level, age,

[1] 'Language test for adult immigrants' and 'Test of Norwegian – advanced level', respectively.

length of residence in Norway, hours of tuition, English skills, etc. All 1,736 texts are error coded using XML markup and annotated with corrections in such a way that a parallel 'correct' corpus can be automatically generated. In addition to the learner texts there is a control corpus of 200 texts written by native speakers. The 1,936 texts vary in length from 66 to 1,068 words [66, 242, 327, 442, 1068; M = 355; SD = 138.6].[1] ASK provided the data for ASKeladden, a research project whose focus has been transfer effects in Norwegian interlanguage (see the project web site http://www.uib.no/fg/askeladden and Johansen *et al.* (2010) for an overview of outcomes from this project). Aspects of the corpus that are particularly salient to the present study are covered in the following sections. The include the L1 groups that are represented, the number of texts (and their distribution across L1 groups), proficiency levels, and thematic variation.

3.2 L1 groups

The data covers ten different L1 groups: German (DE), English (EN), Dutch (NL), Polish (PL), Russian (RU), Serbo-Croat (SH), Spanish (SP), Albanian (SQ), Somali (SO) and Vietnamese (VI). In addition there is a control corpus of texts in Norwegian (NO). Eight of the ten source languages belong to the Indo-European phylum, but they are distributed across four families. DE, EN and NL are West Germanic and thus closely related to NO, which is North Germanic; SP, as a Romance language, and PL, RU and SH, as Slavic languages, are more distantly related; while SQ is (probably) the most distantly related of the Indo-European languages.[2] The remaining two languages, SO and VI, are completely unrelated to NO (Table 10).

Their close genetic relationship mean that a lot of vocabulary items in DE, EN and NL have clearly recognizable cognates in NO. For example, EN 'house', DE 'haus' and NL 'huis' are cognate with – and have more or less the same meaning as – NO *hus*. In contrast, PL 'dom', RU 'дом' and SH 'кућа / kuća', SP 'casa', SQ 'shtëpi', SO 'daar, guri'; and VI 'nhà', are all quite different in form.

[1] The figures in square brackets give the five-number summary (minimum, lower quartile, median, upper quartile and maximum), plus the mean (M) and the standard deviation (SD).

[2] The Indo-European family tree supplied by Campbell & Poser (2007: 84–85) has the SQ subfamily branching off first from Proto-Indo-European, followed by Italo-Celtic (which includes the Romance group), after which come Germanic and Balto-Slavic. As Campbell & Poser point out, this tree is representative but far from universally agreed upon, and the position of SQ in particular is still unclear.

Phylum	Family	Language
Indo-European	Germanic	**NO, DE, EN, NL**
	Slavic	PL, RU, SH
	Romance	SP
	Albanian	SQ
Afro-Asiatic	Cushitic	SO
Austro-Asiatic	Mon-Khmer	VI

Table 10: Genetic affiliations of the ASK source languages

Cross-linguistic similarity in the lexicon can also take the form of loanwords. Many words were borrowed from Old Norse (the ancestor of NO) into Old English (the ancestor of EN) in the Viking period. Later, extensive trade contact with the Hanseatic League led to massive borrowing from Middle Low German (MLG) into the (mainland) Scandinavian languages. According to Haugen (1976: 316), estimates of the proportion of MLG words in those languages run from one half to three-fourths of the vocabulary, and today, as Seip (1934b: 25) pointed out, "two Norwegians cannot … carry on a conversation of 2-3 minutes without using [MLG] loanwords." In more recent times NO, like many other languages, has absorbed a great number of words from EN, but relatively few from any of the other ASK languages.

The sum result of these two factors is that DE, EN and NL learners have a great advantage over other L1 groups when it comes to acquiring NO vocabulary. This can be expected to have a facilitative effect (positive transfer), but may also lead to negative transfer, especially in the form of false friends and semantic extension.

3.3 Number of texts

The original goal in compiling the ASK corpus was to include 100 texts for each L1 group at each of the two levels of proficiency represented by the IL test and the AL test. However, this turned out not to be possible in the case of SO, SQ and VI because of the small numbers of learners from these language backgrounds who had taken (and passed) the AL test. The corpus is thus somewhat unbalanced, as emphasized in Table 11, which shows the breakdown of texts by L1 and test type. (Note that it is precisely the most distantly related of the 10 languages that are most underrepresented.) The 200 texts in the control corpus written by native speakers were composed under conditions approximating those of the two Norwegian language tests and are based on the same kinds of essay prompts.

	DE	EN	NL	PL	RU	SH	SP	SQ	SO	VI	Total
AL test	100	100	100	100	100	100	100	**24**	**7**	**5**	736
IL test	100	100	100	100	100	100	100	100	100	100	1000
Both tests	200	200	200	200	200	200	200	**124**	**107**	**105**	1736

Table 11: Number of texts in ASK at each test level (by L1 group)

3.4 Proficiency levels

The most widely accepted scale for measuring language proficiency in Europe is the Common European Framework of Reference, or CEFR (CoE 2001), which describes proficiency in terms of language functions (what learners can do) using six different levels, A1, A2, B1, B2, C1 and C2, that represent increasing levels of proficiency (Figure 9).

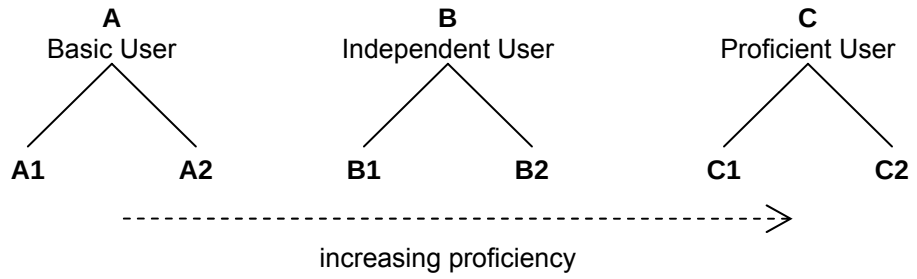


Figure 9: CEFR proficiency levels

Of the two Norwegian tests, the IL test is intended to correspond roughly to B1 on the CEFR scale (Carlsen 2012: 180), while the AL test corresponds to B2/C1 (Carlsen 2012: 180), that is, a level mid-way between B2 and C1.[1] However, it cannot be assumed that all texts that have passed the IL test represent the same level of proficiency (B1), nor that all those that have passed the AL test exhibit a uniform level of B2/C1. This is because B1 and B2/C1 only represent the *minimum* levels necessary in order to pass the tests. As Carlsen (2012: 170) says, "some texts would have passed with a small margin and others with excellence," so some variability of proficiency is inevitable. In order to obtain a more finely tuned and reliable categorization of proficiency, texts from seven of the ten L1 groups were independently re-assessed under strictly controlled conditions according to the CEFR scale. (For the details of this re-assessment, see Carlsen 2012.) The resulting categorization (across both tests) for those seven L1 groups is shown in Table 12 and illustrated graphically in Figure 10 (note that CEFR-based scores are not available for NL, SH and SQ).

[1] Descriptors for these three levels are given in Appendix C.

	DE	EN	PL	RU	SP	SO	VI	sum
C1	5	7	5	3	3	0	0	23
B2/C1	23	13	16	14	14	1	0	81
B2	70	69	58	46	47	2	2	294
B1/B2	46	30	52	60	37	4	3	232
B1	45	61	42	51	60	31	45	335
A2/B1	8	18	25	23	36	55	52	217
A2	3	2	2	3	3	14	3	30
Total	200	200	200	200	200	**107**	**105**	1012

Table 12: CEFR ratings across both tests

In addition to reflecting the imbalance in terms of the number of texts per L1 group (see §3.3), these figures also bring out a marked difference in proficiency levels. Again, it is the most distantly related L1 groups (SO and VI) that stand out the most, and there appears to be a clear correlation between proficiency levels and linguistic and cultural distance.

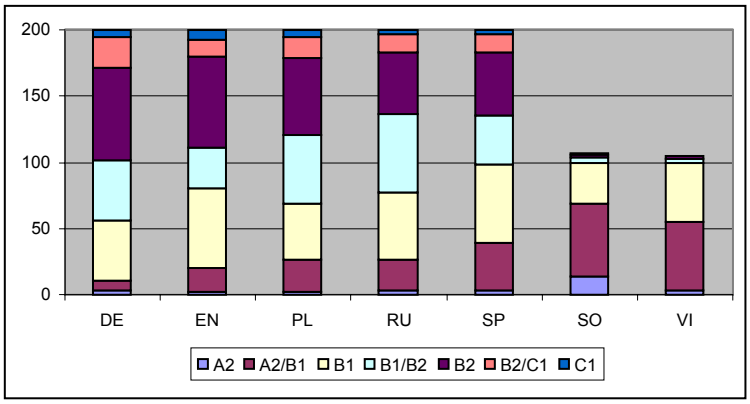


Figure 10: CEFR ratings across both tests

3.5 Thematic variation

Another factor that can influence lexical choice is thematic bias: the vocabulary used in a text is clearly related to the topic of the text. While this is especially obvious in the case of content words, the topic can also influence the choice of function words. For example, a text about a topic which is largely concerned with future events is likely to contain more words that convey future meaning, such as the auxiliary *skal* ('shall'), which is (often) used to form the future tense in Norwegian (especially by learners, see §6.4.2).

	DE	EN	NL	NO	PL	RU	SH	SO	SP	SQ	VI	sum
Alkohol og alkoholvaner 'Alcohol and drinking habits'	2		1		1	3	4	1		1		**13**
Barneoppdragelse 'Bringing up children'			3		**23**		9	1	4	9	4	**53**
Bilbruk 'Motoring'								3			2	**5**
Boformer 'Living arrangements'	6				1			2				**9**
Bolig og bosted 'Home and dwelling place'			9					4			6	**19**
Bomiljø 'Residential environment'		**38**	**16**				9		**23**		3	**89**
Den norske naturen 'Norwegian nature'	1					**27**		4	**12**	8	4	**56**
En bok du har lest 'A book you have read'								5		6	5	**16**
En forfatter og om en bok 'An author and a book'						**16**		1	5	4	3	**29**
En hyggelig opplevelse 'A pleasant experience'								3		**12**	6	**21**
En interesse du har 'An interest you have'								2		3	1	**6**
En kjent person 'A famous person'								2			1	**3**
En person som har betydd mye for deg 'Someone who meant a lot to you'									1			**1**
En religion du kjenner 'A religion you are familiar with'						6		5	1	1	1	**14**
En sport du liker 'A sport you like'	1		2		1			2			9	**15**
Et yrke 'A profession'							7			6	4	**17**
Fjernsynsprogram for barn 'TV programmes for children'	10		2		5	5	8					**30**
Flytting 'Moving'							**12**	1		5	1	**19**
Folk reiser så mye i våre dager 'People travel so much in our times'			2					**11**			1	**14**
Før og nå 'Then and now'			1		4		2	2		4		**13**
Framtida 'The future'			8					**22**	10	10	**13**	**63**
Fremmedspråkslæring 'Learning a foreign language'					**11**			2	9	7	2	**31**
Frihet og ansvar 'Freedom and responsibility'				**22**							1	**23**
Frivillig hjelp i organisasjoner 'Voluntary organisations'	5		2		2		5					**14**
Gjenbruk 'Recycling'			4		2					1		**7**
Glede 'Joy'	2				1						2	**5**
Konkurranse 'Competition'		1	3				1		2		2	**9**

Kultur og idrett 'Culture and sport'					2		1		2	1		**6**
Likestilling 'Equal opportunity'	1										5	**6**
Min første jobb 'My first job'								10			2	**12**
Mobiltelefoner 'Mobile phone'	**23**	**24**	3		**13**		5	2	1		1	**72**
Møte med norsk kultur 'Encountering Norwegian culture'	8	6	1		3	**16**	2	1		3		**40**
Nordmenn 'Norwegians'			1									**1**
Nyheter 'The News'	10	9	7		8		4		4	3		**45**
Penger og andre verdier 'Money and other values'				**35**							2	**37**
Reise 'Travel'			**14**									**14**
Røyking 'Smoking'			4					2			1	**7**
Spillemaskiner 'Gambling machines'				1								**1**
Sunnhet og helse 'Good health'				**42**							6	**48**
Tanker om det å bli gammel 'Thoughts on old age'			2					1				**3**
Trafikk og boligområder 'Traffic and residential areas'	10		2		5	10	4					**31**
Vaner og tradisjoner å ta vare på 'Important traditions to take care of'		**15**	6				5		**18**		5	**48**
Vennskap 'Friendship'	**15**	1	2		5	**17**	**13**	3		8	4	**1**
Viktige verdier i livet 'Important values in life'	6	6	2		**13**		9	3		4	2	**68**
Å treffe andre mennesker 'Meeting other people'			3					4	9	4	1	**45**

Table 13: Essay topics in ASK (Språkprøven)

Table 13 lists the essay topics for the texts used in the present study together with their distribution across L1 groups.[1] Topics that constitute more than 10% of the texts for any one L1 group are highlighted (for example, 23% of the texts written by PL learners are on the topic of *barneoppdragelse* 'bringing up children'). Two things are apparent: (1) the number of topics is quite large, and (2) their overall distribution across L1 groups is very uneven. It will be necessary to bear this issue in mind when reviewing the findings of the analysis.

[1] Note that 'topics' here equates to *essay titles* and should not be confused with the metadata field *tema* 'theme'.

4. Methodology

4.1 Study design

The aim of this study is to replicate as closely as possible the study described in Jarvis *et al.* (2012) using Norwegian data from the ASK corpus. The investigation should therefore involve five L1s and approximately 53 features that represent the word frequencies. The focus was restricted to 1-grams in order to limit the scope of the project, leaving *n*-grams, stylistic metrics and error types for future work.

The main difference between the present study and the one on which it is based is the use of Norwegian interlanguage data instead of English.[1] A second difference is due to the use of different software, which was motivated by the desire to use familiar and/or open source tools where possible, and which affects the methodology to a certain degree. This chapter starts by describing the tools that were employed. It then goes on to discuss a number of design issues and describe how the data was processed, before presenting details of the statistical analysis.

4.2 Software tools

This section describes the software tools used in the project: Omnimark, R and SPSS. Since an important goal was to make it as easy as possible for other investigators to replicate the results and use the same kinds of techniques in their own work, the source data, scripts and intermediate forms of the data are being made available at http://folk.uio.no/stevepe.

[1] Some of the consequences of these differences are discussed in §0.

4.2.1 Omnimark

The software platform for the ASK corpus, Corpuscle (Meurer 2012), offers a range of features for querying, collating and calculating distributions, and for exporting the results, but sometimes it is necessary to perform additional processing on the data. This is especially the case when there is a need to run a lot of different analyses with varying parameters on different subsets of the data – a task that is best performed in batch mode rather than manually, for reasons of efficiency and repeatability.

Jarvis *et al.* used Perl for this purpose, but for the present project Omnimark was preferred. The researcher's familiarity with the tool played a part in this decision, but Omnimark is also ideally suited to the task, being custom-designed for processing structured text in highly sophisticated ways, using its own, rather idiosyncratic but very powerful language. Originally developed for use with SGML, it can handle both SGML-encoded and plain text. The sample script in Appendix C gives a flavour of the language, and documentation is available online at http://florin.bjdean.id.au/docs/omnimark/omni55/docs/html/index.htm. In this project, Omnimark's cross-translation mode was used both to extract information directly from the source files, to process tab-delimited data files downloaded from ASK, and to post-process the output from the statistical analyses.

4.2.2 SPSS

The statistical software package used by Jarvis *et al.* was SPSS, described by Wikipedia as "a computer program used for survey authoring and deployment, data mining, text analytics, statistical analysis, and collaboration and deployment." It is among the most widely used programs for statistical analysis in social science, and it is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations and others. Originally released in 1968, SPSS is now owned by IBM and as of the time of writing the product has reached version 21.0. Details of its many components and their enormous range of functionality can be found at the product web site.[1] Some trial downloads are available, and student editions may be purchased in the US, but otherwise the software is beyond the reach of many researchers. For that reason SPSS was only used in this project in order to enable a proper comparison with the results obtained by Jarvis *et al.* and preference was otherwise given to the open source package R wherever possible.

[1] http://www-01.ibm.com/software/analytics/spss/

4.2.3 R

Wikipedia describes R as "an open source programming language and software environment for statistical computing and graphics[, which] is widely used among statisticians for developing statistical software and data analysis." R provides a wide variety of statistical techniques (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible (RDCT 2012). As an open source tool, R has major advantages over SPSS: it can be used for free, the way it operates is completely transparent to anyone who is able to read the source code, it can be easily extended, and it is supported by an active community of developers. However, it has a steep learning curve and usually requires a certain amount of programming. For the present project a number of add-on packages were used, including MASS, SDDA and klaR. An example R script can be found in Appendix G; in addition, all scripts used in the project are available at the project web site.

4.3 Design issues

The basic methodology used in this project was similar to that used by Jarvis *et al.* and consisted of the following steps:

1. Choose which texts to use in the study.
2. Determine the set of linguistic features to be used in the study.
3. Extract frequency data from the corpus for this set of features.
4. Massage that data into a shape suitable for statistical analysis.
5. Perform the discriminant analysis.
6. Interpret the results using contrastive analysis.

This section describes the main design issues that were faced in choosing the texts on which to base the study. Following that, the practical details involved in extracting and preparing the data are discussed in §4.4, after which the statistical analyses are described in §4.5. The results of the latter are presented in Chapter 5 and these are subjected to contrastive analysis in Chapter 6.

4.3.1 Choice of L1s

Jarvis *et al.* (2012: 26-27) report that the five L1s in their study were deliberately chosen in order to test the effects of genetic proximity. They were DA and SW (two closely related

Germanic languages), PO and SP (two closely related Romance languages) and FI, a non-Indo-European language that is unrelated to the other languages but spoken in the same general area as DA and especially SW. In order to approximate these conditions as closely as possible, it was decided for the present project to choose two closely related Germanic languages and two closely related Slavic languages and to supplement these with a fifth, unrelated language. In case it should prove necessary to investigate the role of proficiency, the initial intention was to select all five languages from the seven L1s for which ASK includes information on CEFR levels (see §3.4). This meant choosing DE and EN as the two Germanic languages, and PL and RU as the two Slavic languages.

Choosing a fifth language to supplement DE, EN, PL and RU was more problematic. The options were limited to SP, SO and VI, since CEFR data does not exist for NL, SH and SQ. Of these, SO and VI are completely unrelated to (and typologically very different from) both the target language (NO) and the four L1s already selected. Either of them could therefore play the role of unrelated L1 that FI played in the model study. However, Jarvis *et al.* (2012: 47) make an explicit point of the fact that speakers of their unrelated L1 "share a culture and educational system" with another of the L1 groups being tested, the implication being that cultural and educational differences might interfere with the ability to detect L1 influence and should therefore be controlled. The problem here is that the cultural distance between (speakers of) SO and VI on the one hand, and DE, EN, PL and RU on the other, is much greater than that between FI and DA, SW, PO and SP. For one thing, all of the latter share a common European culture, which is very different from the SO and VI cultures. Moreover, the SO and VI in Norway are refugee communities: immigration from Vietnam dates back to the time of the boat refugees who fled their country after the fall of Saigon in 1975, while Somalis have been arriving in Norway as refugees since the breakdown of law and order following the fall of Siad Barre in 1991. This is not the case for speakers of DE, EN, PL and RU, who tend to come to Norway seeking work – or in order to study – as Figure 11 shows.

This raises the question of which factor to accord greater weight in choosing the fifth L1 group for this study: genetic (and typological) distance, which would favour SO or VI, or cultural homogeneity, which would favour SP? Rather than having to make a choice here, it was decided to conduct a series of parallel investigations, combining the four base L1s with each of the other L1s in turn. This approach was facilitated using scripts to automate the processing and in the event it proved possible to run the analyses on a total of six different

groups of five L1s. Each of these "pentagroups" consisted of the four base L1s (DE, EN, PL and RU) plus one of the other six L1s (NL, SH, SO, SP, SQ and VI). The pentagroups will hereafter be denoted the "NL-group", the "SH-group", etc.[1]

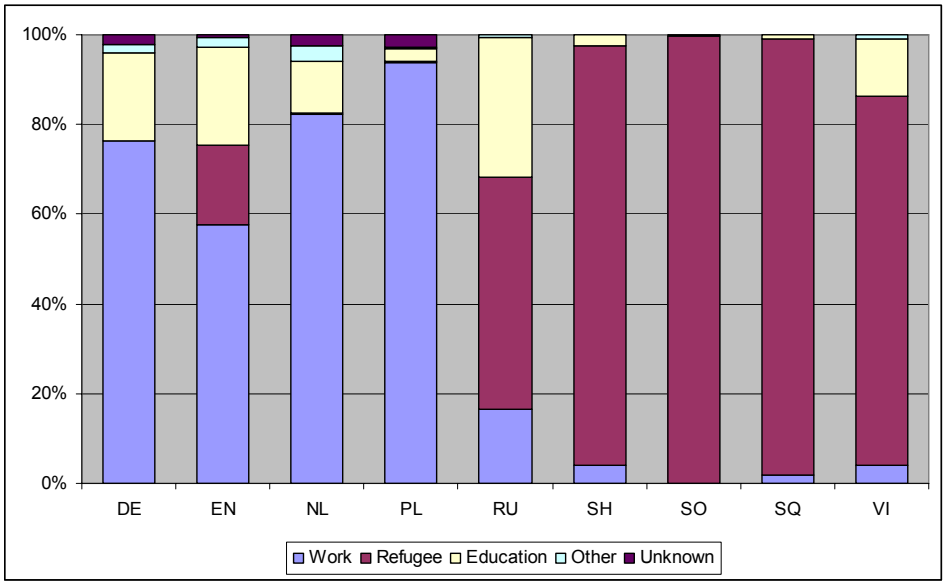


Figure 11: Reasons for immigrating to Norway, by nationality[2]

4.3.2 Sample size

The next question to be addressed was how many texts to use. Jarvis *et al.*'s set of 446 was somewhat unevenly distributed across their five L1s: DA (n = 60), SW (n = 70), PO (n = 60), SP (n = 116) and FI (n = 140) (p. 46). As mentioned in §3.3, ASK contains 200 texts for each of the L1 groups DE, EN, NL, PL, RU, SH and SP, but only slightly over 100 for each of SO, SQ and VI. This is quite a large difference which it was feared might compromise the results of the analysis, especially as it is compounded by uneven proficiency levels (see §3.4). It was therefore decided to use just 100 texts from each L1 group rather than all the available texts (which also had the nice side-effect that figures in confusion matrices can be read as either absolute numbers or percentages).

---

[1] The term pentagroup, meaning a group of five L1 groups, was coined for the purpose of this thesis in order to avoid confusion with 'group' used in the sense of 'L1 group', i.e. all learners from a particular L1 background.

[2] The EN figures represent immigration from USA and the UK; those for SH, Serbia and Montenegro; and those for SQ, Kosovo and Bosnia-Hercegovinia; the others represent Germany, Netherlands, Poland, Russia, Somalia and Vietnam. Figures for immigration from Spanish-speaking countries were not available. *Source:* Statistics Norway: *Innvandringer etter innvandringsgrunn og statsborgerskap. 1990-2010*.

4.3.3 Proficiency level

The purpose of this study is to investigate the extent to which Norwegian L2 learners' lexical choices are influenced by their L1 background. However, lexical choices can be influenced by a number of other factors, one of which is proficiency level. The imbalance that we see in the sample of interlanguage texts available in ASK is therefore a major challenge. In order to address this it was decided to restrict the study to texts written for the IL test.[1]

	DE	EN	PL	RU	SP	SO	VI	sum
C1	0	0	0	0	0	0	0	0
B2/C1	1	1	0	0	0	0	0	2
B2	19	5	8	4	0	1	0	37
B1/B2	27	15	24	24	7	3	1	101
B1	42	59	41	46	54	27	44	313
A2/B1	8	18	25	23	36	55	52	217
A2	3	2	2	3	3	14	3	30
Total	100	100	100	100	100	100	100	700

Table 14: CEFR ratings for the IL test data

The distribution of proficiency levels in this reduced sample is much more even, as shown in Table 14 and Figure 12 (cf. Table 12 and Figure 10). Furthermore, the variation in text length is also reduced [66, 216, 253, 305, 619; M = 265; SD = 71.6]. And since there are 100 IL texts for each L1, this accords nicely with the solution to the problem of sample sizes.

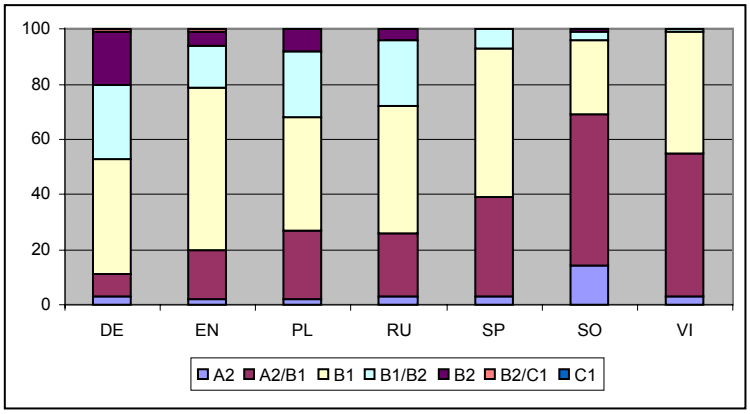


Figure 12: CEFR ratings for the IL test data

[1] Jarvis *et al.* (2012: 47) explicitly recruited learners representing a wide range of L2 proficiency "in order to test the sensitivity of [their] classifier." However, that was not a goal of the present project.

4.4 Data processing

The practical procedure that was followed is described here in some detail in order that other researchers might follow the same steps.

4.4.1 Querying ASK for word frequency lists

Once it had been decided which texts to base the analysis on, it was necessary to arrive at a set of features comparable to the 53 1-grams used by Jarvis *et al*. This involved extracting word lists from ASK using the Corpuscle query interface. In order to replicate the study by Jarvis *et al.*, separate word lists were required for each L1 group. In addition, word lists were generated for the corpus as a whole, both with and without the texts of the control corpus. The resulting word lists were downloaded as 13 text files, each of which contained a header followed by several thousand lines of frequency data

4.4.2 Fixing case sensitivity in the word lists

The word lists thus obtained were case-sensitive. Words written in different combinations of upper and lower case were treated as different words: for example, *Jeg* and *jeg* were treated as two different words, which of course they are not – they are merely orthographic variants of the same word. For the purpose of the current investigation, such differences were not relevant and case-*in*sensitive data was required. However, while it was possible to produce and display case-insensitive word lists on-screen in the version of ASK available at the time of the study, it was not possible to download them in this form. To remedy this, the word list files were processed using Omnimark and at the same time tidied (by removing headers and relative frequencies) and trimmed to 500 lines. The resulting data was collected in an Excel spreadsheet.

4.4.3 Prompt-induced words

The next step was to remove words whose variations in frequency might be due to the topic of the essay, which varies considerably in ASK. Such thematic bias could skew the analysis if (for whatever reason) the distribution of topics were to vary across different L1 groups. This issue did not arise for Jarvis *et al.* because their texts were thematically homogeneous, but it was faced by Jarvis & Paquot (2012) in their study based on (the equally heterogeneous) texts from ICLE. They operationalized the concept of "prompt-induced words" as "all content words (and their families) that appeared in any of the essay prompts and were used by more

than 35 learners…" (p. 83) and on this basis they disqualified from their list of 200 1-grams words such as *society*, *prison*, *science*, *technology*, *television*, *religion*, *imagination* and *dream*. However, Jarvis & Paquot were using 200 features whereas the present study was to use approximately 53, as Jarvis *et al.* had done. As it turned out, these included relatively few prompt-induced words, and it was therefore decided not to attempt to identify and remove such words at this stage, but rather to bear in mind the possibility of thematic bias when analysing the results (see §3.5 and §6.4.9).

4.4.4 Preparing word frequency data

Frequency data was extracted for the 200 most frequent words across the learner corpus as a whole for each of the 1,936 texts using the ASK query interface. Once again it was necessary to post-process the data using Omnimark in order to merge frequency counts for different orthographic variants of the same word and to remove percentage values. Two further operations remained to be performed before this data could be used in the statistical analyses. First, to compensate for the fact that the texts varied quite substantially in length, the absolute values had to be converted to relative frequencies per 1,000 words. Second, columns needed to be added for the grouping variable, L1, and for CEFR level (where available), the latter in order to enable filtering based on proficiency should this prove necessary. Both of these operations were performed using Excel functions.

4.4.5 Choosing the feature set

The feature set used by Jarvis *et al.* was arrived at by pooling the 30 most frequent words in each of their five L1 groups. Because of overlap between groups, this resulted in a set of 53 words. The initial intention in the present study was to do exactly the same. However, the ASK data turned out to be more homogeneous in terms of word choice across the various L1 groups than the data used by Jarvis *et al.* Pooling the 30 most frequent words from the four base L1s (EN, DE, PL and RU) resulted in a set of just 40 words. Adding the 30 most frequent words from SP and VI made no difference: the result was the same 40 words. Including the 30 most frequent SO words merely increased the total to 43, which was still considerably fewer than Jarvis *et al.*'s 53. Table 15 shows how increasing the number of words from each L1 group affected the total number of features in each pentagroup. In order to arrive at numbers close to those used in the model study, it was decided to use the 40 most frequent words in each L1 group (highlighted).

	EN, DE, PL, RU +					
Word count	**NL**	**SH**	**SO**	**SP**	**SQ**	**VI**
30	40	41	43	40	41	40
35	50	49	52	48	51	49
40	57	56	58	55	58	58
42	62	60	60	58	61	60
45	64	63	63	61	64	63

Table 15: Number of features in each pentagroup

The final set of data thus consisted of six sets of word frequencies, one for each pentagroup of L1s. Each pentagroup consisted of data from the four base L1s (EN, DE, PL and RU) plus one of the six additional L1s (NL, SH, SO, SP, SQ and VI). In LDA terms, there were six data sets, each consisting of 500 *cases* (i.e. texts), with 100 from each *group* (i.e. L1), and the number of *features* in each data set varied (as shown in the highlighted row of Table 15) from 55 (for the SP-group) to 58 (for the SO-, SQ- and VI-groups). By way of illustration, the set of features actually used in the analysis of the SP-group is shown in Figure 13.

> *alle, andre, at, av, bare, barn, barna, bo, da, de, den, det, du, eller, en, er, et, for, fordi, fra, ha, han, har, hvis, i, ikke, jeg, kan, liker, må, man, mange, med, meg, men, mye, når, norge, norsk, og, også, om, på, så, seg, som, sted, til, være, var, veldig, venner, vi, viktig, å*

Figure 13: The 55 features used for the SP-group

4.5 Statistical analysis

Jarvis *et al.* performed three kinds of statistical analysis on their data: analysis of variance (ANOVA), discriminant analysis (LDA), and a post-hoc test called Student-Newman-Keuls (SNK). In addition, Baayen (2008: 155), in discussing the application of LDA to linguistic data, includes an "unsupervised exploration" of his data using principal component analysis (PCA). This section describes all those methods as they were performed on the data from ASK. It starts by discussing the assumptions that these methods make about the data and the extent to which the data conforms to those assumptions. This is followed by sections on ANOVA and PCA, four sections on LDA, and a brief concluding mention of SNK.

A secondary aim of this project, as previously mentioned, was to use more freely available tools, which meant opting for R in place of SPSS. As it turned out, the most widely used R package for performing discriminant analysis (MASS) does not support feature selection. Two other R packages (SDDA and klaR) were therefore tested in addition to MASS, and the three sets of results were compared with those obtained using the same tool as Jarvis *et al.*, i.e. SPSS. For this reason, the LDA methods and results are reported in four parts (in this and the following chapter), under the headings *LDA using R + MASS*, *LDA using R + SDDA*, *LDA using R + klaR* and *LDA using SPSS*.

4.5.1 Statistical assumptions

Before subjecting the data to statistical analysis it was necessary to check that it adhered to the assumptions made by the relevant methods. LDA's assumptions are discussed in §2.2.1.5 on page 21. Of these, the first four are clearly met by the present data:

(1) there are more than two groups ($G = 5$);
(2) there are at least two cases per group ($n_i = 100$);
(3) the number of discriminating variables (from 55 to 58) is less than 498, i.e. the total number of cases minus two (500 − 2);
(4) the discriminating variables are measured at the interval level (the word frequencies range from 0 to 118.8 per 1,000 words).

Given the nature of the data it is highly unlikely that any discriminating variable will be a linear combination of other variables (5), but in any case, this will be automatically checked by the software (see §4.5.4, below). That leaves two assumptions: equal covariance matrices (6) and normally distributed data (7). These are the same assumptions that ANOVA makes, along with the (usual) further assumption of independence of observations.

The assumption of approximately equal covariance matrices for each group could unfortunately not be tested within the scope of the present study. According to Kabacoff (2011), this assumption is usually tested with Box's M, "but that test is very sensitive to violations of normality, leading to rejection in most typical cases." Following Larson-Hall (2010: 273) the assumption of normal distribution was tested for a selection of features by inspecting box plots, density curves and Q-Q plots. Figure 14 shows the results for the overall most frequent word, *det*, and for two words that are discussed in Chapter 6, *en* and *skal*.

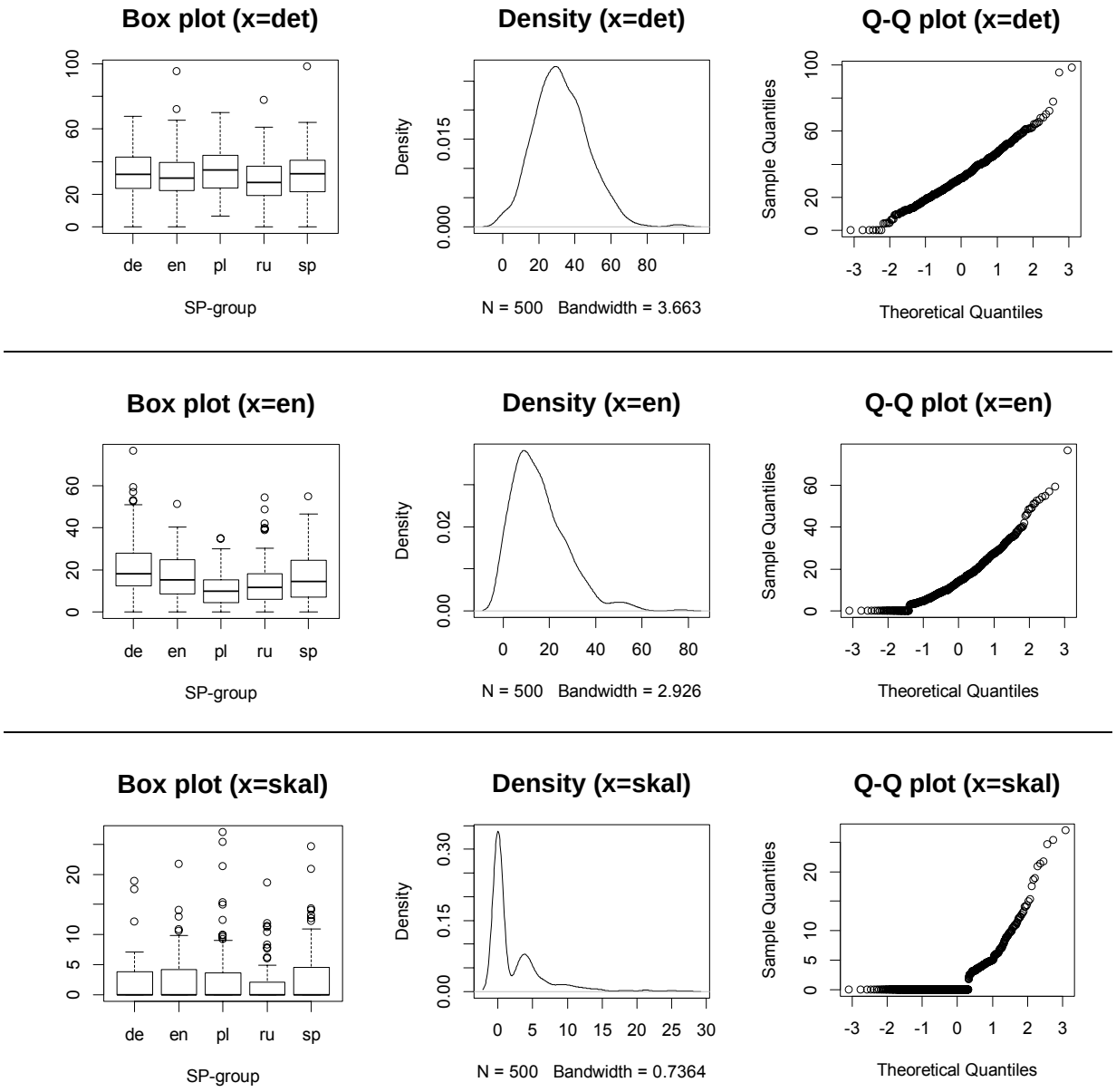


Figure 14: Overall distribution of $det, $en **and** $skal

These plots show the distribution to be at best only approximately normal and at worst considerably non-normal, and this is confirmed by Shapiro-Wilk tests:[1]

```
$det:  W = 0.9836, p = 2.006e-05
$en:   W = 0.9234, p = 2.832e-15
$skal: W = 0.6341, p < 2.2e-16)
```

[1] In order to avoid confusion with the words on which they are based, features are hereafter referred to by prefixing the relevant word with the $ symbol.

Field *et al.* (2012: 183) point out that when the analysis involves comparing groups, "what's important is not the overall distribution but the distribution in each group." Unfortunately it was not possible to test group-wise distribution within the scope of the present project; in any case, it was unlikely to be completely normal for every variable. Fortunately, though, the deviations from normality tend to be of the same kind (i.e. positively skewed). This – and the fact that the samples are all of the same size – serve to counterbalance somewhat the lack of normality (Bård Uri Jensen, pc 2012-11-05).

Some statisticians (Pallant 2001; Weinberg & Abramovitz 2002) assert that if group sizes are 30 or more there is no reason to worry about meeting normality requirements. Larson-Hall & Herrington (2009: 376-377) are sceptical to this claim, but point out that "small deviations from normality in the distribution are fairly robust to Type I errors (rejecting the null hypothesis when in reality it is true, and there actually is no difference between groups)." The real danger is Type 2 errors, in which the null hypothesis is accepted when in reality it is not true and there actually is a difference between groups. This is reassuring for the present study, since it means that observed group differences can be trusted, but for in future work it would be wise to heed Larson-Hall and Herrington's advice to make use of the more modern method of robust statistics (Wilcox 2010) and the techniques of bootstrapping and trimmed means. This would provide even more power to detect group differences.

The question of statistical assumptions was discussed with Jarvis, who commented:

> The question of whether your data are normally distributed would be a big one for me if the purpose of your study were to analyze whether different groups are significantly different from one another. However, your purpose in using ANOVA and DA is simply to construct an optimal model of the similarities within and differences across groups in terms of their word choices. Your "real" results are the classification accuracies arrived at through applying the model to the test cases. The application of the model to the test cases does not involve any inferential statistical tests at all... At worst, a statistician might argue that your violation of assumptions has resulted in a less than optimal model (which would mean that your classification accuracies are lower than they have the potential to be), but there are no Type I errors here to worry about as long as your cross-validation with test cases was sufficiently rigorous (Jarvis, pc 2012-11-07).

4.5.2 Analysis of variance

Following Jarvis *et al.* a series of one-way ANOVAs was performed for each pentagroup prior to performing the multivariate analysis. The purpose of an ANOVA is to determine whether or not the means of several groups are equal, based on the observed variance within each group. The tests were performed using R with the relative frequencies of each feature in turn as one variable and the learners' L1 as the other. Figure 15 shows the output summary for the word *skal* for the NL-group:

```
> summary(aov(skal ~ L1))
             Df Sum Sq Mean Sq F value  Pr(>F)
myData$L1     4   1008  251.95   8.667 9.1e-07 ***
Residuals   495  14389   29.07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 15: R code and output for a one-way ANOVA

The important things to note here are the high value of the F statistic (8.667), and the low probability (9.1e-07) of such a value occurring by chance. This means that the frequency of occurrence of the word *skal* varies statistically across the five L1 groups DE, EN, PL, RU and NL, and could potentially be used to distinguish one or more of these groups from the others. The full ANOVA results are presented in the next chapter (see p.65ff) and the ability of individual features such as the word *skal* to discriminate between L1 groups is discussed in §6.3.

4.5.3 Principal components analysis

Following Baayen (2008: 154ff) the multivariate analysis commenced with an unsupervised exploration of the data was performed using a clustering technique known as principal component analysis (PCA). Unsupervised explorations are so-called because they make no *a priori* assumptions about the structure of the data, i.e. they do not start out from a known set of groupings (such as L1 membership); instead, they attempt to find structure in the data unaided. A `pca` object was created using the R function `prcomp()` as shown in Figure 16. For the data set that included the Spanish data (the SP-group), this resulted in 58 principal components, one for each of the 55 features in the data set. In contrast to Baayen's data, in which the first eight principal components capture almost 80% of the variance, the SP-group data required as many as 32 PCs to capture the same amount. What this shows is that most of the features in the ASK data contribute only a small amount of discriminatory power.

```
# generate pca object
numFeatures = ncol(myData) -1
myData.pca = prcomp(myData[, 1:numFeatures], center=T, scale=T)
summary(myData.pca)
#
# create 3x3 scatter plot matrix
library(lattice)
super.sym = trellis.par.get("superpose.symbol")
splom(~myData.x[ ,1:3], groups = myData$L1,
  panel = panel.superpose,
  key=list(
    title=paste(c(numCases, "texts,", numClasses," L1s,",
                  numFeatures, "1-grams"), collapse=" "),
    text=list(levels(myData$L1)),
    points = list(pch = super.sym$pch[1:numClasses],
                  col = super.sym$col[1:numClasses])
    )
  )
```

Figure 16: R code for the principal components analysis

To explore the structure found in the data, a scatter plot matrix was generated using the function `splom` (Figure 17). Because of the large number of data points (500), the matrix for the first 3 principal components (PCs) is not very easy to interpret, but some group separation is discernible, e.g. of EN, with △s tending toward the South-West in row 2, which indicates negative values for PC2 and positive values for PC1 and PC3, and of RU, with ×s tending towards the West in column 1 (negative values for PC1). The tendency for certain L1 groups to cluster in certain areas strongly suggested that there was some kind of structure in this data

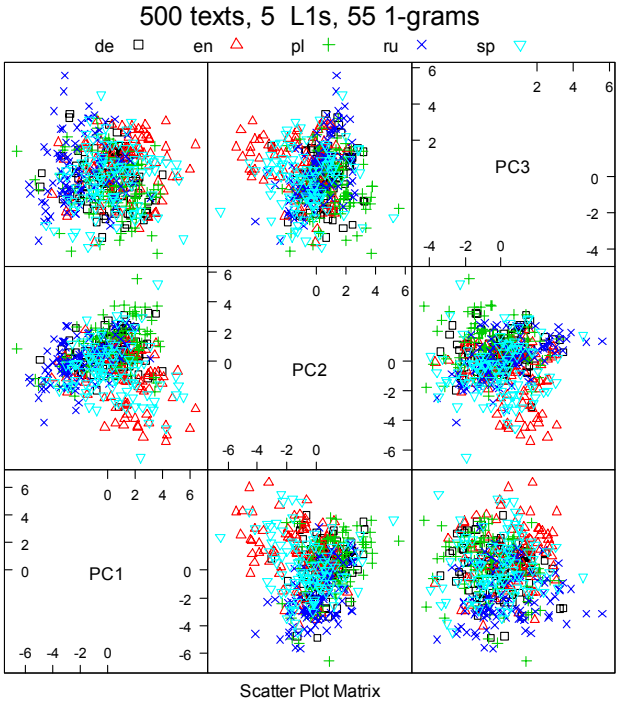


Figure 17: PCA scatter plot matrix for EN, DE, PL, RU + Spanish

which could help distinguish between different L1s. To discover what this structure was, the next step was to harness the predictive and descriptive power of LDA.

4.5.4 LDA using R + MASS

The first attempt to perform LDA follows Baayen in using the `lda()` function from the R package MASS, but with a slightly different approach. In Baayen's example there was too much correlation in the data for `lda()` to work properly. He therefore performed his analysis using principal components, which are uncorrelated by definition. In the case of the ASK data, there is no such problem: creating an `lda` object using the code shown below did not result in any warnings about collinearity, so it was possible to base the analysis directly on the features themselves rather than the principal components, and then to plot the results:

```
myData.lda = lda(myData[ ,1:numFeatures], myData$L1)
plot(myData.lda)
```

Because there are five groupings in the data, one for each of the five L1s ($G = 5$), a total of four ($G - 1$) discriminant functions are created; referred to here as LD1, LD2, LD3 and LD4. In Figure 18 LD1 is plotted against LD2, producing a clear, albeit partial, separation of DE (to the North), RU (West), EN (East) and PL (South), but less separation for SP.[1]

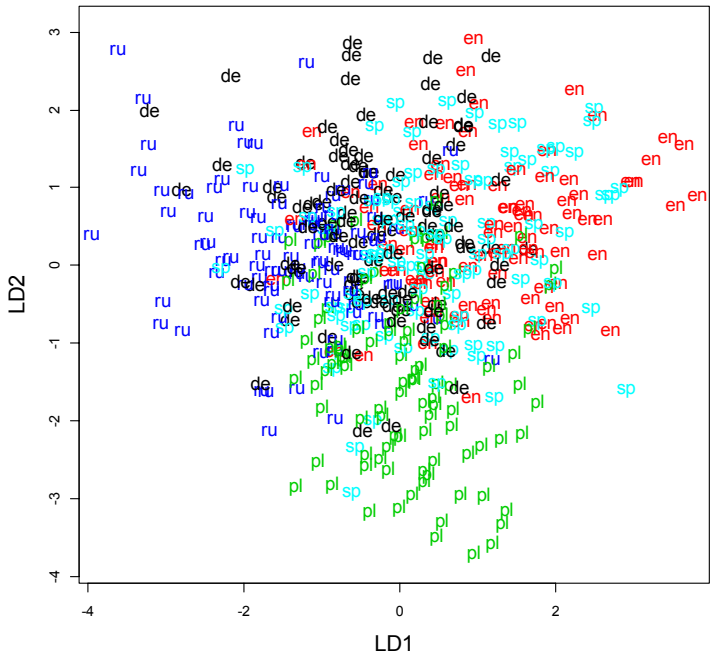


Figure 18: LDA scatter plot for the SP-group (LD1 and LD2)

[1] A matrix plot of all four LDs is given in Appendix D.

Using the function `predict()` this model can be queried for the probability with which it assigns texts to L1s (the results are rounded to two decimals for convenience):

```
round(predict(myData.lda, myData[ , 1:numFeatures])$posterior, 2)
```

Partial results are shown in Table 16, in which an extra column has been inserted to show the actual L1. Shaded cells highlight the L1 that is assigned the highest probability for each text. The L1 is correctly predicted for many texts, with probabilities that range from 36% (`s1005`, PL) to 100% (`s0011`, EN). However several L1s are incorrectly predicted, some of them with a very high degree of (misplaced) confidence, e.g. `s0009` (VI), predicted EN with 87% certainty.

Actual L1	Text	Probabilities				
		DE	EN	PL	RU	VI
VI	**s0005**	**0.00**	**0.01**	**0.01**	**0.00**	**0.97**
VI	**s0007**	**0.01**	**0.02**	**0.68**	**0.02**	**0.26**
VI	**s0008**	**0.00**	**0.06**	**0.13**	**0.00**	**0.80**
VI	**s0009**	**0.01**	**0.87**	**0.11**	**0.00**	**0.01**
EN	**s0010**	**0.20**	**0.73**	**0.03**	**0.01**	**0.03**
EN	**s0011**	**0.00**	**1.00**	**0.00**	**0.00**	**0.00**
	...					
DE	**s1000**	**0.82**	**0.07**	**0.07**	**0.01**	**0.04**
DE	**s1001**	**0.30**	**0.01**	**0.02**	**0.47**	**0.19**
DE	**s1002**	**0.96**	**0.04**	**0.00**	**0.00**	**0.00**
PL	**s1005**	**0.11**	**0.20**	**0.36**	**0.07**	**0.27**
DE	**s1006**	**0.89**	**0.01**	**0.04**	**0.05**	**0.02**
DE	**s1007**	**0.92**	**0.08**	**0.00**	**0.00**	**0.00**

Table 16: Probability table generated by `lda()`

These results may not seem very impressive, but it should be remembered that the purpose of the exercise is not to develop a system for classifying learner texts automatically, but rather to discover L1-related grouping effects and, from them, to identify good L1 predictors that can be subjected to contrastive analysis. Provided the results are statistical, that goal will have been achieved. However, even these results seriously overfit the data, as Baayen points out:

> [The analysis] has done its utmost to find a representation of the data that separates the groups as best as possible. This is fine as a solution for this particular sample of texts, but it does not guarantee that prediction will be accurate for unseen [texts] as well (Baayen 2008: 157).

Following Baayen's example, a multivariate analysis of variance (MANOVA) was performed:

```
myData.manova = manova( as.matrix(myData[,1:numFeatures]) ~ L1, myData )
summary.manova(myData.manova)
```

The summary of the MANOVA results using the default Pillai-Bartlett statistic (Table 17) provides very strong evidence that the mean vectors for the five L1s do indeed differ, which means that the discriminant analysis is finding statistical differences between them.

	Df	Pillai	approx F	num Df	den Df	Pr(>F)	
L1	4	1.4301	4.231	232	1764	< 2.2e-16	***
Residuals	495						

Table 17: Summary of MANOVA test

To gain a more precise impression of the extent to which these results might generalize, a LOOCV was performed using the option built in to the `lda()` function. The results of this analysis are presented in the next chapter (see p.68 ff). A shortcoming of the `lda()` function in the MASS package is that it does not support feature selection, which leads to a risk that the resulting statistical model will overfit the data and give an overly optimistic view of the model's accuracy. To circumvent this problem, two R packages that do offer feature selection were tested, as described in the following sections.

4.5.5 LDA using R + SDDA

SDDA is an R package that offers a "fast algorithm for building multivariate classifiers" using stepwise diagonal discriminant analysis (Clifford 2010). It was chosen for inclusion in this project because of its support for stepwise feature selection, despite the fact that it employs a variant of discriminant analysis called (linear) diagonal discriminant analysis. Stepwise feature selection is performed by adding the variable "that most decreases the (leave-one-out) cross-validated error rate at each step." The addition of features stops "when the cross-validated error rate cannot be decreased" – a parameter over which the user has no control. Feature selection cannot be embedded in the folds of the cross-validation.

SDDA was run on the 40-word data sets for the six pentagroups using the `sdda()` function for feature selection and the `xvalidate()` function for LOOCV. (SDDA has the option of doing 10-fold CV but tests showed that the results of such validation varied quite widely.) The results of the analysis are presented in the next chapter (see p.69 ff).

4.5.6 LDA using R + klaR

klaR is an R package that provides "miscellaneous functions for classification and visualization" (Ligges 2012), including stepwise classification. Although klaR does not support embedded feature selection "out of the box", the feature selection functionality is available in a form that permits embedding to be implemented relatively easily in an R script. klaR's `stepclass()` function allows the user to set a threshold for the "improvement of performance measure desired to include or exclude any variable." The default value is 0.05. However it was found that this value resulted in the selection of very few features, so it was set to the much lower value of 0.001. `stepclass()` was then used in conjunction with `lda()` on the six data sets using 10-fold CV. Interleaved partitioning was performed on data sets that had been sorted by L1, in order to ensure an even distribution of L1 groups in all the training sets and test sets. The results of the analysis are presented in the next chapter (see p.71 ff).

4.5.7 LDA using SPSS

In order for the results of this study to be maximally comparable with those of Jarvis *et al.*, a further analysis of the six data sets was performed using SPSS (version 20). In the time available, it was not possible to implement cross-validation with embedded feature selection, so 'up-front' feature selection was used instead (that is, the same set of features was used in each fold of the cross-validation). Also, LOOCV was used instead of 10-fold CV. Otherwise, the same parameters were used as in the model study, i.e. stepwise method using Wilks' lambda, F to enter $p < 0.05$, F to remove $p > 0.10$, and all available statistics were output. (Screenshots showing the exact settings are given in Appendix E and the results of the analysis are presented in the next chapter, see p.74 ff).

4.5.8 Post-hoc tests

Following their discriminant analysis, Jarvis *et al.* subjected the features chosen during feature selection to post-hoc analysis using the Student-Newman-Keul test with the goal of pinpointing exactly how the selected features served to separate different groups. However, doubt has been cast on the reliability of this test (Hsu 1966; Seaman *et al.* 1991), and for that reason it has not been implemented in R. In the present project Tukey HSD tests were therefore used instead. The details of how this was done and how to interpret the results are presented in §6.3.

5. Findings

The purpose of the tests and analyses described in the preceding chapter was to discover whether the lexical choices of language learners could be used to distinguish between learners from different L1 backgrounds and, if so, which of those choices provided the most discriminating power. This chapter presents the findings of those investigations and sets the scene for the theoretical discussion in the next chapter.

5.1 Analysis of variance

The analysis of variance (§4.5.2) produced F statistics showing whether the mean frequency of each feature differed statistically across L1 groups. Table 18 and Table 19 show the results of the six sets of analyses in order of decreasing value of F for each group of learners. Only words whose mean frequencies differ statistically at the level of $p < 0.001$ across the five L1s are included. For the SO-group, 35 words were statistical at this level. The figures for the other pentagroups are SP 28, VI 31, NL 35, SH 32 and SQ 33. (For a complete list of F values for all six sets of features, see the file **anova.xls**). The words *sted* 'place' and *bo* 'live' exhibit the greatest degree of variability in every one of the six pentagroups. Other words that are among the ten most variable in all six pentagroups are *eller* 'or', *er* 'is/are', *en* and *et* 'a' and *viktig* 'important'. In all, 29 words exhibited a degree of variability that was statistical at $p < 0.001$ in at least five of the six pentagroups. They are listed in Figure 19, glossed in Appendix B, and discussed further in Chapter 6.

sted, bo, et, viktig, en, eller, er, og, man, norge, jeg, å, den, var, barna, også, liker, bare, i, norsk, ikke, veldig, må, da, barn, ha, andre, skal, hvis

Figure 19: L1 predictor candidates (ANOVA)

DE EN PL RU + Somali		
Feature	*F value*	*p*
sted	33.460	< 2e-16
bo	23.080	< 2e-16
et	18.320	4.73E-14
jeg	16.520	1.03E-12
viktig	16.510	1.04E-12
er	15.050	1.27E-11
eller	14.480	3.46E-11
en	13.540	1.75E-10
fordi	12.370	1.36E-09
man	11.430	6.96E-09
skal	10.950	1.62E-08
bare	10.600	2.99E-08
var	9.699	1.47E-07
barna	9.177	3.70E-07
og	8.866	6.40E-07
norge	8.270	1.84E-06
den	7.786	4.32E-06
om	7.147	1.34E-05
også	6.923	1.99E-05
å	6.855	2.24E-05
det	6.717	2.86E-05
barn	6.583	3.62E-05
veldig	6.186	7.30E-05
som	6.119	8.22E-05
være	6.076	8.86E-05
norsk	6.048	9.31E-05
liker	5.940	0.000113
ikke	5.823	0.000138
ha	5.681	0.000178
i	5.570	0.000216
mye	5.565	0.000218
må	5.506	0.000242
da	5.456	0.000264
på	5.366	0.000309
andre	5.361	0.000312

DE EN PL RU + Spanish		
Feature	*F value*	*p*
sted	26.760	< 2e-16
bo	19.360	8.21E-15
å	15.750	3.82E-12
et	15.690	4.27E-12
viktig	13.020	4.34E-10
en	12.810	6.21E-10
eller	11.760	3.90E-09
er	10.590	3.09E-08
og	9.811	1.21E-07
man	9.506	2.07E-07
norge	9.154	3.85E-07
var	8.351	1.59E-06
liker	8.221	2.00E-06
barna	7.535	6.74E-06
barn	7.507	7.09E-06
den	6.674	3.09E-05
da	6.555	3.81E-05
ikke	6.165	7.57E-05
bare	6.157	7.69E-05
også	6.109	8.37E-05
i	5.900	0.000121
veldig	5.864	0.000129
jeg	5.720	0.000166
norsk	5.711	0.000169
må	5.663	0.000184
ha	5.306	0.000343
andre	5.213	0.000404
hvis	4.820	0.000803

DE EN PL RU + Vietnamese		
Feature	*F value*	*p*
sted	31.040	< 2e-16
bo	20.090	2.39E-15
viktig	16.390	1.29E-12
en	15.360	7.48E-12
et	14.870	1.75E-11
eller	13.600	1.59E-10
er	13.590	1.62E-10
og	11.810	3.61E-09
også	9.475	2.18E-07
norge	9.226	3.39E-07
jeg	8.94	5.62E-07
den	8.786	7.38E-07
man	8.506	1.21E-06
skal	7.772	4.43E-06
det	7.504	7.13E-06
bare	7.252	1.11E-05
i	7.153	1.32E-05
barna	7.045	1.60E-05
var	6.998	1.74E-05
å	6.840	2.30E-05
ikke	6.642	3.27E-05
liker	6.571	3.70E-05
må	6.341	5.55E-05
veldig	6.276	6.23E-05
norsk	5.943	0.000112
at	5.866	0.000128
ha	5.571	0.000215
da	5.518	0.000237
barn	5.506	0.000242
hvis	5.207	0.000408
andre	4.792	0.000843

Table 18: One-way ANOVAs of features (SO, SP, VI)

DE EN PL RU + Dutch			DE EN PL RU + Serbo-Croat			DE EN PL RU + Albanian		
Feature	*F value*	*p*	*Feature*	*F value*	*p*	*Feature*	*F value*	*p*
sted	27.74	< 2e-16	**sted**	32.01	< 2e-16	**sted**	29.75	< 2e-16
bo	16.02	2.4E-12	**bo**	18.42	4.02E-14	**bo**	25.14	< 2e-16
en	15.82	3.42E-12	**et**	18.14	6.45E-14	**et**	16.11	2.05E-12
et	15.43	6.67E-12	**en**	15.65	4.58E-12	**viktig**	15.43	6.67E-12
viktig	13.18	3.29E-10	**viktig**	12.4	1.28E-09	**en**	12.94	4.97E-10
er	12.38	1.32E-09	**eller**	12.31	1.50E-09	**eller**	12.79	6.45E-10
eller	12.3	1.53E-09	**den**	12.17	1.92E-09	**skal**	11.25	9.64E-09
og	9.899	1.03E-07	**er**	11.61	5.11E-09	**man**	11.14	1.17E-08
man	9.428	2.37E-07	**skal**	10	8.63E-08	**jeg**	10.94	1.66E-08
norsk	9.27	3.14E-07	**og**	9.633	1.65E-07	**er**	10.9	1.76E-08
norge	8.928	5.74E-07	**man**	9.56	1.88E-07	**var**	10.7	2.51E-08
skal	8.667	9.10E-07	**å**	9.316	2.89E-07	**norge**	10.54	3.34E-08
den	7.821	4.06E-06	**norge**	9.026	4.83E-07	**og**	10.32	4.93E-08
i	7.395	8.64E-06	**barna**	8.599	1.03E-06	**kan**	8.881	6.24E-07
liker	7.273	1.07E-05	**ikke**	7.678	5.24E-06	**men**	8.326	1.67E-06
å	7.257	1.10E-05	**veldig**	7.601	5.99E-06	**liker**	8.325	1.67E-06
også	7.102	1.45E-05	**vi**	7.546	6.61E-06	**i**	8.111	2.43E-06
barna	6.52	4.05E-05	**også**	7.363	9.14E-06	**også**	6.864	2.21E-05
var	6.475	4.39E-05	**må**	7.19	1.24E-05	**ha**	6.661	3.16E-05
ikke	6.314	5.83E-05	**var**	7.022	1.67E-05	**barna**	6.481	4.34E-05
må	6.162	7.62E-05	**liker**	6.746	2.72E-05	**å**	6.443	4.64E-05
bare	6.141	7.91E-05	**da**	6.502	4.18E-05	**da**	6.268	6.32E-05
som	5.93	1.15E-04	**norsk**	6.458	4.52E-05	**ikke**	6.118	8.23E-05
jeg	5.707	1.70E-04	**om**	6.346	5.51E-05	**den**	6.064	9.06E-05
da	5.536	2.29E-04	**jeg**	6.089	8.66E-05	**bare**	5.965	0.000108
om	5.483	2.52E-04	**i**	5.893	0.000122	**fordi**	5.904	0.00012
mer	5.479	0.000253	**andre**	5.876	0.000126	**barn**	5.867	0.000128
ha	5.459	0.000262	**ha**	5.676	0.000179	**veldig**	5.8	0.000144
barn	5.304	0.000344	**hvis**	5.317	0.000337	**må**	5.672	0.00018
veldig	5.267	0.000368	**bare**	5.311	0.00034	**hvis**	5.658	0.000185
mange	5.223	0.000397	**barn**	4.845	0.000769	**norsk**	5.401	0.000291
hvis	5.181	0.000427	**at**	4.755	0.0009	**at**	4.955	0.000635
venner	4.775	0.000868				**andre**	4.796	0.000838
kan	4.706	0.000980						
andre	4.7	0.000990						

Table 19: One-way ANOVAs of features (NL, SH, SQ)

5.2 Discriminant analysis

Discriminant analysis was performed using four different approaches: R + MASS (with LOOCV and no feature selection), R + SDDA (with up-front feature selection and 10-fold CV), R + klaR (10-fold CV and embedded feature selection) and SPSS (up-front feature selection and LOOCV). The results of these analyses are presented in the following sections.

5.2.1 LDA using R + MASS

Table 20 shows confusion matrices generated by the `lda()` function in R's MASS package. There is one confusion matrix per pentagroup and the number of correctly predicted L1s for each set of actual L1 texts is shown in bold. In the SO-group, for example, 56 of the 100 DE texts were correctly predicted. 10 were mistakenly identified as EN, 10 as PL, 15 as RU and 9 as SO. Since there were exactly 100 texts from each L1 group, these numbers translate directly into percentages. Thus, DE was correctly predicted 56% of the time in the SO-group, 49% in the SP-group, 55% in the VI-group, etc.

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SO**
Actual L1	**DE**	**56**	10	10	15	9
	EN	17	**56**	15	6	6
	PL	13	10	**53**	12	12
	RU	18	4	10	**62**	6
	SO	6	10	13	15	**56**
	Accuracy 56.6 % *p* = 2.67e-72					

A. DE EN PL RU + **Somali**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SP**
	DE	**49**	11	15	16	9
	EN	17	**47**	11	7	18
	PL	12	8	**58**	11	11
	RU	14	5	14	**60**	7
	SP	12	19	16	14	**39**
	Accuracy 50.6 % *p* = 2.48e-52					

B. DE EN PL RU + **Spanish**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**VI**
	DE	**55**	9	15	16	5
	EN	18	**53**	13	6	10
	PL	14	6	**50**	11	19
	RU	13	3	11	**60**	13
	VI	12	9	19	12	**48**
	Accuracy 53.2 % *p* = 1.34e-60					

C. DE EN PL RU + **Vietnamese**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**NL**
Actual L1	**DE**	**41**	10	14	15	20
	EN	15	**51**	14	9	11
	PL	13	5	**63**	10	9
	RU	12	4	15	**65**	4
	NL	18	13	10	4	**55**
	Accuracy 55.0 % *p* = 1.14e-66					

D. DE EN PL RU + **Dutch**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SH**
	DE	**56**	8	13	16	7
	EN	19	**53**	14	8	6
	PL	17	6	**47**	11	19
	RU	15	4	9	**61**	11
	SH	8	8	22	13	**49**
	Accuracy 53.2 % p = 1.34e-60					

E. DE EN PL RU + **Serbo-Croat**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SQ**
	DE	**56**	9	14	14	7
	EN	18	**50**	16	7	9
	PL	15	8	**57**	6	14
	RU	8	4	12	**62**	14
	SQ	9	10	11	14	**56**
	Accuracy 56.2 % p = 7.18e-71					

F. DE EN PL RU + **Albanian**

Table 20: Confusion matrices (MASS)

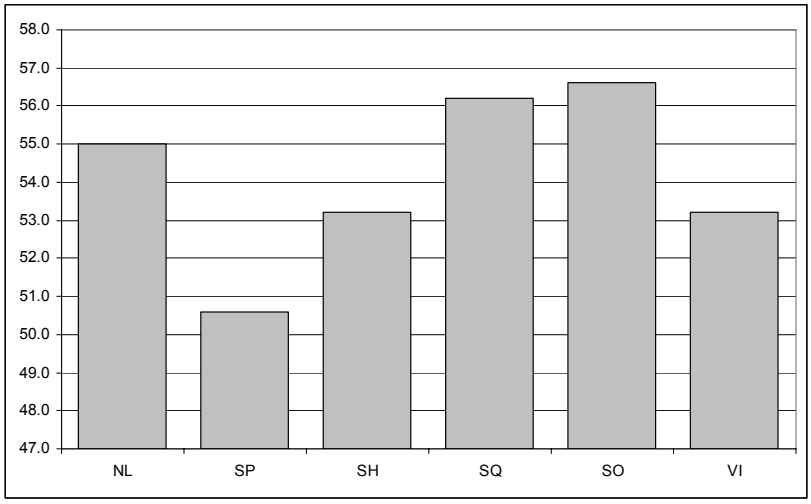


Figure 20: Overall accuracy rates (MASS)

The overall prediction accuracy in these six analyses, plotted in Figure 20, ranged from a low of 50.6% in the SP-group to highs of 56.2 in the SQ-group and 56.6% in the SO-group. These success rates might seem low, but as the *p*-values in Table 20 indicate, they are all statistical: the probability of attaining such results by chance is tantamount to zero.

5.2.2 LDDA using R + SDDA

Table 21 shows confusion matrices generated using the R package SDDA with stepwise feature selection. Comparison with the results from MASS (see Table 20 on page 68) shows a substantially lower degree of accuracy. Whereas the average accuracy across all six pentagroups using MASS was 54.1%, the average accuracy using SDDA is just 42.3%.

Since the data were identical, there are two possible explanations. The first is that the algorithm used by SDDA is less well-suited to this set of data than the plain discriminant analysis algorithm implemented in MASS. A more likely explanation, however, is the tiny number of features selected by SDDA's feature selection algorithm. Table 22 shows which features (*bare, den, eller,* etc.) were selected in each of the six analyses. In the case of the NL-group, a total of nine features was selected (*eller, en, i, liker, mer, og, også, skal* and *sted*). For the SP-group, a different set of nine features was selected (with some overlap), while eight features were selected for the SH-, SO- and SQ-groups, and a mere seven for the VI-group.

It is unclear why SDDA should select so few features. According to the documentation, variables are added to the model until "the cross-validated error rate cannot be decreased"

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SO**
Actual L1	**DE**	**53**	2	11	24	10
	EN	23	**47**	7	13	10
	PL	20	7	**32**	23	18
	RU	16	2	16	**44**	22
	SO	11	10	17	20	**42**
	Accuracy 43.6 % p = 5.63e-33					

A. DE EN PL RU + **Somali**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SP**
	DE	**46**	5	21	20	8
	EN	23	**28**	13	15	21
	PL	17	3	**51**	20	9
	RU	14	4	19	**53**	10
	SP	14	23	18	17	**28**
	Accuracy 41.2 % p = 2.53e-27					

B. DE EN PL RU + **Spanish**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**VI**
	DE	**52**	2	18	14	14
	EN	23	**46**	9	10	12
	PL	14	5	**32**	20	29
	RU	12	2	14	**51**	21
	VI	15	10	25	24	**26**
	Accuracy 41.4 % p = 8.94e-28					

C. DE EN PL RU + **Vietnamese**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**NL**
Actual L1	**DE**	**53**	2	19	15	11
	EN	23	**32**	18	14	13
	PL	15	1	**48**	26	10
	RU	15	3	22	**57**	3
	NL	19	16	18	14	**33**
	Accuracy 44.6 % p = 1.76e-35					

D. DE EN PL RU + **Dutch**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SH**
	DE	**53**	2	17	21	7
	EN	26	**38**	8	14	14
	PL	13	3	**36**	22	26
	RU	14	4	18	**49**	15
	SH	15	11	27	14	**33**
	Accuracy 41.8 % p = 1.09e-28					

E. DE EN PL RU + **Serbo-Croat**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SQ**
	DE	**49**	4	15	17	15
	EN	23	**44**	13	13	7
	PL	14	6	**33**	33	14
	RU	19	4	18	**46**	13
	SQ	20	9	11	25	**35**
	Accuracy 41.4 % p = 8.94e-28					

F. DE EN PL RU + **Albanian**

Table 21: Confusion matrices (SDDA)

– in other words, as long as the model's accuracy continues to improve. Exactly how this works can only be gleaned from the source code which is written in C and not easily accessible to the present researcher. Understanding why SDDA contents itself with so few variables is therefore a task that must be left to future research.

	bare	den	eller	en	er	fra	i	jeg	liker	men	mer	norge	og	også	om	skal	som	sted	viktig	å	#
NL			•	•			•		•		•		•	•		•		•			9
SH		•	•	•								•	•	•		•		•			8
SO	•		•	•				•							•		•	•	•		8
SP			•	•	•	•			•				•	•				•		•	9
SQ			•	•			•			•			•			•		•	•		8
VI			•	•					•				•	•				•	•		7
totals:	1	1	6	6	1	1	2	1	3	1	1	1	5	4	1	3	1	6	3	1	

Table 22: Features selected by SDDA

While SDDA is of less interest than MASS in terms of its predictive powers, its ability to identify features as being good L1 predictors is useful. From Table 22 we observe that three features were picked out in each of the six analyses: *eller*, *en* and *sted*. In addition, *og* was picked in five passes, *også* in four, and *liker*, *skal* and *viktig* in three passes. The total of 20 features, listed in Figure 21, will be revisited in the next chapter.

eller, en, sted; og; også; liker, skal, viktig; i; bare, den, er, fra, jeg, men, mer, norge, om, som, å

Figure 21: L1 predictor candidates (SDDA)

5.2.3 LDA using R + klaR

Table 23 shows confusion matrices resulting from the analysis in which the R package klaR was used to perform stepwise feature selection which was embedded within the folds of a 10-fold cross-validation. The overall predictive accuracy was on a par with SDDA and is further discussed in §6.1.1.

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SO**
Actual L1	**DE**	**42**	5	20	22	11
	EN	20	**46**	13	10	11
	PL	17	4	**37**	29	13
	RU	20	4	13	**52**	11
	SO	11	8	14	20	**47**
	Accuracy 44.8 % p = 5.43e-36					

A. DE EN PL RU + **Somali**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SP**
	DE	**40**	5	21	24	10
	EN	22	**36**	9	14	19
	PL	23	3	**41**	22	11
	RU	19	4	13	**54**	10
	SP	14	25	18	20	**23**
	Accuracy 38.8 % p = 3.50e-22					

B. DE EN PL RU + **Spanish**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**VI**
	DE	**47**	6	19	16	12
	EN	23	**44**	9	11	13
	PL	13	5	**45**	23	14
	RU	15	2	20	**49**	14
	VI	17	9	15	20	**39**
	Accuracy 44.8 % p = 5.43e-36					

C. DE EN PL RU + **Vietnamese**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**NL**
Actual L1	**DE**	**36**	9	28	15	12
	EN	21	**42**	13	12	12
	PL	16	6	**43**	27	8
	RU	10	0	16	**61**	13
	NL	19	14	18	15	**34**
	Accuracy 43.2 % p = 5.35e-32					

D. DE EN PL RU + **Dutch**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SH**
	DE	**50**	9	18	12	11
	EN	22	**43**	17	9	9
	PL	20	3	**29**	23	25
	RU	19	6	12	**51**	12
	SH	16	11	22	16	**35**
	Accuracy 41.6 % p = 3.14e-28					

E. DE EN PL RU + **Serbo-Croat**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SQ**
	DE	**48**	4	22	17	9
	EN	16	**47**	19	10	8
	PL	15	6	**36**	29	14
	RU	13	2	21	**48**	16
	SQ	12	8	13	27	**40**
	Accuracy 43.8 % p = 1.81e-33					

F. DE EN PL RU + **Albanian**

Table 23: Confusion matrices (klaR)

Folds	Word	NL	SH	SO	SP	SQ	VI	5grps
57	viktig	10	10	10	9	10	8	6x
48	sted	6	8	9	5	10	10	6x
43	eller	6	8	7	5	8	9	6x
30	en	5	4	4	5	6	6	6x
29	skal	6	5	4	–	9	5	5x
28	og	6	2	4	1	6	9	6x
24	et	7	2	4	5	3	3	6x
22	barna	6	4	4	2	4	2	6x
21	man	7	4	–	3	4	3	5x
19	bo	3	3	2	6	2	3	6x
19	kan	3	2	3	4	5	2	6x
17	andre	2	1	6	3	4	1	6x
17	vi	4	7	1	1	1	3	6x
17	norsk	6	4	1	3	–	3	5x
16	må	3	5	1	3	1	3	6x
16	i	4	2	1	–	4	5	5x
15	fordi	–	–	8	1	5	1	4x
14	å	–	3	3	8	–	–	3x
13	på	3	1	3	2	3	1	6x
13	fra	–	2	2	3	2	4	5x
12	jeg	1	1	3	–	2	5	5x
12	men	1	–	3	1	4	3	5x
12	mye	2	2	7	–	1	–	4x
11	ikke	3	–	2	1	4	1	5x
11	barn	1	–	2	5	3	–	4x
11	om	–	2	3	–	2	4	4x
10	ha	1	2	4	1	1	1	6x
10	den	1	1	4	2	–	2	5x
9	av	2	–	2	2	1	2	5x
9	bare	1	2	2	–	3	1	5x
9	liker	1	1	2	–	3	2	5x
9	mange	4	2	–	1	–	2	4x
8	er	1	1	3	2	–	1	5x
8	norge	1	2	3	1	–	1	5x
8	du	–	2	1	–	2	3	4x
8	så	–	2	2	1	–	3	4x
7	for.	2	–	–	1	3	1	4x
6	når	–	2	–	2	1	1	4x
6	også	1	2	–	1	–	2	4x
6	være	2	–	2	–	1	1	4x
6	at	1	–	–	3	–	2	3x
6	seg	3	–	1	–	2	–	3x
6	som	3	–	2	–	1	–	3x
6	var	1	1	–	–	4	–	3x
6	det	–	–	–	1	–	5	2x
6	mer	6	–	–	–	–	–	1x
5	alle	2	1	1	–	–	1	4x
5	de	–	2	1	1	1	–	4x
5	han	1	1	–	1	2	–	4x
5	til	1	2	–	1	1	–	4x
5	venner	1	–	–	1	1	2	4x
5	veldig	–	1	–	–	–	4	2x
4	hvis	–	–	–	1	2	1	3x
3	med	–	1	–	2	–	–	2x
3	min	–	–	1	–	–	2	2x
3	hadde	–	–	–	–	3	–	1x
2	da	–	–	1	1	–	–	2x
2	noen	–	–	–	–	–	2	1x
1	har	–	1	–	–	–	–	1x
1	hun	–	–	–	–	1	–	1x
1	meg	–	–	–	1	–	–	1x
1	mennesker	1	–	–	–	–	–	1x
		43	**41**	**41**	**41**	**42**	**45**	

Table 24: Features selected by klaR

Of more interest is the set of features selected, both overall and for each group of five L1s, which is summarized in Table 24. The first column in this table gives the total number of folds in which each feature was selected. There was a total of 60 folds altogether (10 per pentagroup). As the table shows, the feature most often selected was the word *viktig*, which was included in the model in 57 of the 60 folds. It was followed by *sted*, which was included 48 times: 6 times in the NL-group (i.e. DE, EN, PL, RU + NL), 8 times in the SH-group, 9 times in the SO-group, etc. Both of these features were chosen in all 6 pentagroups, as shown in the right-hand column (6x). The 28 features selected in ten or more folds (Figure 22) will be further discussed in the next chapter.

viktig, sted, eller, en, skal, og, et, barna, man, bo, kan, andre, vi, norsk, må, i, fordi, å, på, fra, jeg, men, mye, ikke, barn, om, ha, den

Figure 22: L1 predictor candidates (klaR)

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SO**
Actual L1	**DE**	**55**	8	10	18	9
	EN	17	**53**	18	5	7
	PL	17	6	**58**	8	11
	RU	10	6	11	**63**	10
	SO	7	8	11	15	**59**
Accuracy 57.6 % p = 5.08E-76						

A. DE EN PL RU + **Somali**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SP**
Actual L1	**DE**	**51**	8	13	15	13
	EN	14	**43**	16	10	17
	PL	8	8	**62**	10	12
	RU	15	4	10	**64**	7
	SP	13	21	16	14	**36**
Accuracy 51.2 % p = 2.65E-54						

B. DE EN PL RU + **Spanish**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**VI**
Actual L1	**DE**	**54**	7	11	19	9
	EN	18	**50**	16	8	8
	PL	13	5	**51**	14	17
	RU	10	4	12	**60**	14
	VI	12	8	15	11	**54**
Accuracy 53.8 % p = 1.08E-62						

C. DE EN PL RU + **Vietnamese**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**NL**
Actual L1	**DE**	**43**	11	15	13	18
	EN	16	**51**	14	9	10
	PL	11	6	**57**	14	12
	RU	10	2	18	**66**	4
	NL	17	13	11	3	**56**
Accuracy 54.6 % p = 2.14E-65						

D. DE EN PL RU + **Dutch**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SH**
Actual L1	**DE**	**55**	8	15	18	4
	EN	19	**53**	17	8	3
	PL	14	7	**52**	12	15
	RU	15	5	11	**63**	6
	SH	9	8	19	12	**52**
Accuracy 55.0 % p = 9.13E-67						

E. DE EN PL RU + **Serbo-Croat**

		Predicted L1				
		DE	**EN**	**PL**	**RU**	**SQ**
Actual L1	**DE**	**61**	5	14	13	7
	EN	15	**51**	19	8	7
	PL	16	7	**51**	12	14
	RU	9	4	14	**63**	10
	SQ	7	7	12	13	**61**
Accuracy 57.4 % p = 2.75E-75						

F. DE EN PL RU + **Albanian**

Table 25: Confusion matrices (SPSS)

5.2.4 LDA using SPSS

Table 25 shows confusion matrices resulting from the analysis using SPSS with up-front feature selection and LOOCV. The overall predictive accuracy was on a par with MASS and is further discussed in §6.1.1. Table 26 shows that the number of features selected for each pentagroup ranged from 25 for the SP-group to 34 for the SO-group.

	Features entered (and removed)					
	NL	**SH**	**SO**	**SP**	**SQ**	**VI**
1	sted	sted	sted	sted	sted	sted
2	en	en	jeg	en	en	en
3	eller	eller	en	å	eller	og
4	mer	den	viktig	eller	viktig	eller
5	skal	skal	eller	og	skal	viktig
6	viktig	og	som	norge	og	den
7	og	norge	bare	da	kan	norge
8	i	da	fordi	den	liker	(det)
9	mange	barna	barna	barn	men	barna
10	liker	mange	og	et	i	skal
11	barna	vi	om	viktig	bo	også
12	den	ikke	skal	fordi	barna	bare
13	venner	er	veldig	ikke	da	at
14	norsk	til	meg	veldig	fordi	veldig
15	av	også	på	vi	bare	venner
16	kan	andre	den	man	som	er
17	ikke	fra	er	fra	veldig	på
18	vi	være	min	også	man	av
19	da	med	andre	kan	også	til
20	han	et	han	på	du	norsk
21	veldig	så	vi	er	jeg	som
22	om	om	mange	han	hun	kan
23	men	liker	barn	til	andre	du
24	har	når	venner	du	å	fra
25	er	de	også	venner	til	med
26	bare	min	bo		på	man
27	du	på	de		fra	jeg
28	man		til		den	de
29	for		kan		når	(det)
30	mennesker		du		vi	andre
31			man			bo
32			mye			
33			norge			
34			norsk			

Table 26: Features selected by SPSS

A total of 53 features were selected altogether and these constitute the L1 predictor candidates for SPSS listed in Figure 23, which are analyzed in the next chapter.

andre, at, av, bare, barn, barna, bo, da, de, den, det, du, eller, en, enn, er, et, for, fordi, fra, han, har, hun, i, ikke, jeg, kan, liker, man, mange, med, meg, men, mennesker, mer, min, mye, norge, norsk, når, og, også, om, på, skal, som, sted, så, til, veldig, venner, vi, viktig, være, å

Figure 23: L1 predictor candidates (SPSS)

This concludes the summary of statistical findings from the two kinds of analysis performed on the data: ANOVA and the LDA. The next chapter discusses these findings, comparing them with those of Jarvis *et al.*, providing further statistical analysis of the L1 predictors, and subjecting the latter to contrastive analysis in order to ascertain whether or not they can be ascribed to language transfer.

6. Discussion

The statistical models that are constructed using discriminant analysis can be used for two distinct purposes:

1) predictive, i.e. using classification to "detect" the class membership of a particular individual based on the features that it exhibits; and
2) descriptive, i.e. using feature selection to identify the features that constitute the best predictors of class membership.

In terms of the data used in the present study, these purposes translate to:

1′) detecting the L1 of a Norwegian L2 learner based on his or her lexical choices; and
2′) revealing which lexical features are most typical for different L1 groups.

In this chapter the results of the study are discussed in the light of these two purposes. First, L1 detection rates are compared across the four classifiers used in the study and contrasted with the rates achieved by Jarvis *et al.* (§6.1). This allows the first research question to be addressed: Can automated text-based classification be used to identify the L1 background of Norwegian language learners based on their use of lexical features of the target language?

Sometimes misclassification can be as revealing as successful classification. Some of the findings in this respect are therefore discussed in §6.2. Following this the various sets of L1 predictors are compared and subjected to a post-hoc analysis in §6.3 in order to answer the second research question: What are the best source language (L1) predictors?

LDA, the statistical method that led to the discovery of those L1 predictors, operates in a way that is conceptually very simple: Given a set of individuals that can be divided into a

number of discrete groups and that exhibit certain features, the method creates a model, defined in terms of a set of linear functions that minimizes the within-group variance and maximizes the between-group variance. As was pointed out in §2.2.1.2 on page 18, this naturally provides the first two types of evidence called for by Jarvis in his methodological requirements:

- Minimized within-group variance equates to Jarvis' intragroup homogeneity
- Maximized between-group variance equates to Jarvis' intergroup heterogeneity

However, LDA alone cannot provide the third type of evidence, cross-language congruity. To do so requires a linguistically analysis of the patterns of lexical usage revealed by LDA, in order to determine whether they can be traced back to features of the users' L1 usage. Such contrastive analysis is the topic of §6.4. The closing section of this chapter (§6.5), in summarizing the contrastive analysis, considers certain issues regarding the role of linguistic distance and the kind of lexical transfer that has been detected.

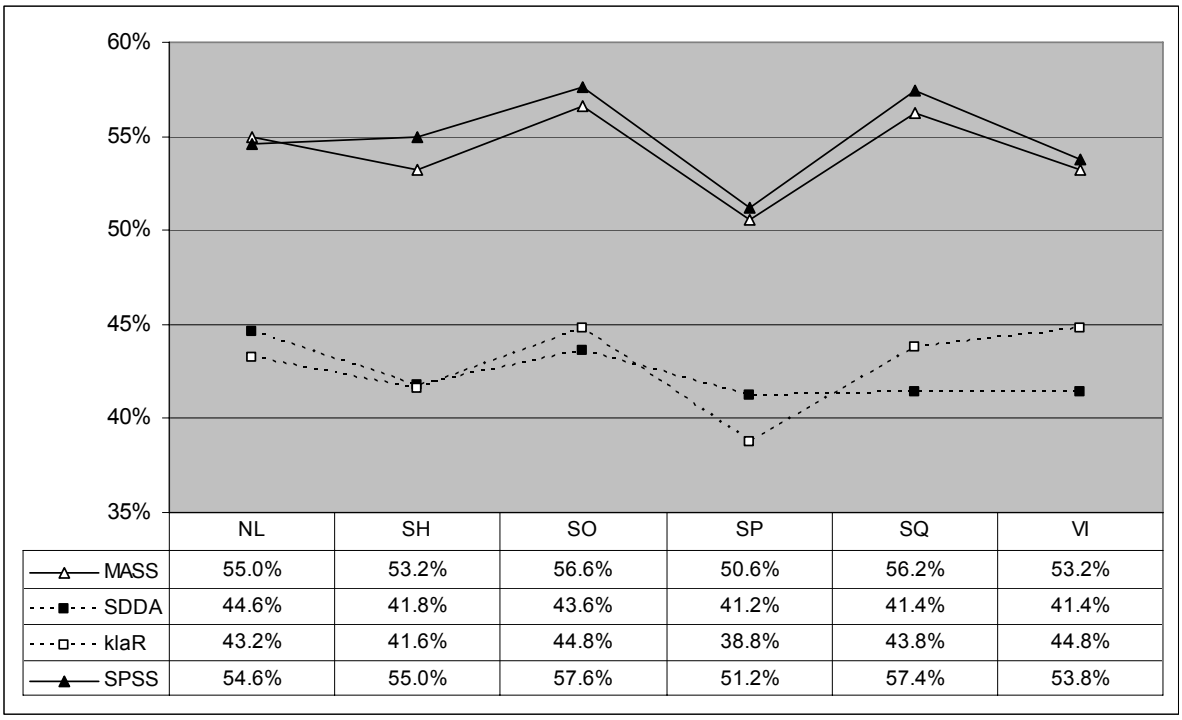
6.1 L1 detection

6.1.1 Comparing classifiers

As a side-effect of the attempt to replicate the study by Jarvis *et al.* using open source software, this project was obliged to test multiple classifiers. However, unlike Jarvis (2011), who compared classifiers that used radically different algorithms (§2.3.1.7), the four classifiers tested here (MASS, SDDA, klaR and SPSS) were all implementations of LDA. Three of these classifiers (MASS, SDDA and klaR) are packages built on the open source environment R. The fourth, SPSS, was included for reasons of comparability (since it was used by Jarvis *et al.*). The `lda()` function in the MASS package is the most widely used R-based implementation of LDA and it is also used in klaR. SDDA, on the other hand, implements a different "flavour" of LDA, diagonal linear discriminant analysis, which makes slightly different assumptions about the data, *viz.* that "the class densities have the same diagonal covariance matrix" (Dudoit *et al.* 2002: 79).

The major shortcoming of the MASS implementation of LDA is that it does not support feature selection, and this was the reason for testing SDDA and klaR in addition. Both of these provide up-front feature selection (in which a set of features is selected once and then used in all folds of the cross-validation), but neither of them offer *embedded* feature

selection. However, it proved possible to implement this fairly easily with klaR. SPSS also provides feature selection but only 'up-front' and not embedded. In contrast to Jarvis *et al.* (and for reasons of scope), no attempt was made to implement the latter in SPSS using scripts.



	NL	SH	SO	SP	SQ	VI
MASS	55.0%	53.2%	56.6%	50.6%	56.2%	53.2%
SDDA	44.6%	41.8%	43.6%	41.2%	41.4%	41.4%
klaR	43.2%	41.6%	44.8%	38.8%	43.8%	44.8%
SPSS	54.6%	55.0%	57.6%	51.2%	57.4%	53.8%

Figure 24: Comparison of prediction accuracy across classifiers

Figure 24 compares the accuracy rates achieved by the four classifiers across the six penta-groups. The percentages are overall accuracies for each pentagroup as a whole, not for the individual languages. MASS (△ without feature selection) and SPSS (▲ with up-front feature selection) performed almost identically, their results averaging 54.1% and 54.9%, respectively, and considerably better than SDDA (■ with up-front feature selection) and klaR (□ with embedded feature selection), whose results averaged 42.3% and 42.8%.

As suggested in §5.2.2, the performance of SDDA can probably be attributed to the tiny number of features selected (from seven to nine depending on the pentagroup), but the poor showing of klaR requires further investigation. Table 24 on page 72 shows that klaR selected between 41 and 45 features for each pentagroup. However, selection took place within the folds of a 10-fold CV and a closer examination of the output from the script `klar.R` reveals

considerable variation in the number of features selected per fold, ranging from six in fold 8 of the SP-group to 21 in fold 1 of the NL-group (Table 27). The average per pentagroup ranges from 10.3 to 13.6, for an overall average of 12.45. This means that the number of features contributing to the model was (on average) little more than with SDDA and much lower than either MASS (the full set of 56-58 features, see p. 55) or SPSS (from 25 to 34 using up-front feature selection, see p. 74).

	fold number										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	(average)
NL	21	8	10	11	20	8	14	10	19	11	13.2
SH	7	17	11	8	12	8	13	15	10	10	11.1
SO	13	9	20	6	18	17	13	8	13	12	12.9
SP	7	10	8	8	15	6	10	12	14	13	10.3
SQ	11	16	9	12	18	9	16	19	10	16	13.6
VI	10	8	13	14	13	17	16	13	14	18	13.6

Table 27: Features selected per fold (klaR)

To sum up, then, the greatest predictive accuracy was achieved with MASS and SPSS using from 25 to 58 features. SDDA and klaR, using from 7 to 21 (average 12.45), achieved much lower accuracy. So is this simply a matter of accuracy correlating positively with the number of features? The more features, the greater the accuracy? In order to answer this question, a further investigation was performed using MASS with feature sets of variable sizes based on word lists that varied from 3 to 200.[1] The results are shown in Table 28 and Figure 25.

	Number of words												
	3	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**42**	**45**	**100**	**200**
NL	30.8	34.4	38.2	40.4	44.2	46.0	47.6	52.4	55.0	56.2	55.2	55.4	56.0
SH	28.2	34.0	40.0	42.6	43.6	48.6	50.0	51.4	53.2	53.4	54.6	54.2	51.4
SO	32.8	36.4	41.8	43.0	46.6	50.6	53.2	55.6	56.6	56.6	59.2	55.4	50.8
SP	29.6	35.6	38.6	40.2	43.4	46.6	47.4	47.6	50.6	51.6	50.4	49.4	51.0
SQ	31.0	36.2	38.4	41.8	44.4	48.6	51.8	55.2	56.2	55.0	55.6	56.4	53.2
VI	35.0	37.0	43.4	43.6	44.8	48.8	51.0	52.6	53.2	53.8	55.2	59.0	53.2

Table 28: Accuracy by number of features

[1] The size of the feature sets themselves varied from 4 to 198 since Jarvis *et al.*'s pooling method (§2.3.2.1) was used and only features for which data points actually existed in the top 200 frequency list were allowed.

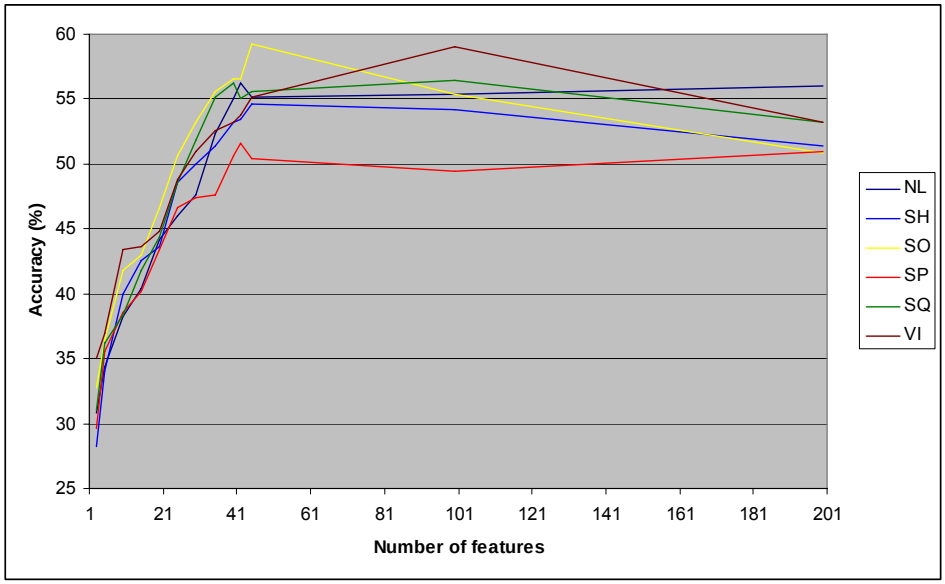


Figure 25: Accuracy by number of features

What this table shows is that improvements in accuracy level off when the number of features reaches that achieved by pooling the 40 or so most frequent words from each L1 group – which happens to be exactly the number used in this study. Further increases in the number of features have no major effect, either positive or negative. Reducing the size of the feature set, on the other hand, leads to a degradation in accuracy, but one that is surprisingly gradual: even pooling just the three most common words in each L1 group still results in statistical levels of predictive accuracy, which is quite remarkable. With the ASK data this produced the same set of four features in all six pentagroups: the words *og*, *det*, *er* and *jeg* ('and, it, is, I'), and produced accuracy rates ranging from 28.2% to 35.0%, which are still statistical (p = 6.95e-06 and p = 4.12e-15, respectively). There is simply no way such results could be achieved by chance, and the evidence for grouping effects based on the L1 is thus overwhelming.

6.1.2 Comparing with Jarvis *et al.*

As far as the predictive part of this study is concerned, the accuracy rates achieved, although statistical, are somewhat lower than those of Jarvis *et al.* The latter achieved an overall rate of 76.9%, with accuracies ranging from 63.6% for DA to 89.7% for SP (see Table 8 on page 34), whereas the results achieved using MASS and SPSS in the present study ranged from 50.6% to 57.6%. There are a number of factors, both data-related and methodological, that could explain this discrepancy.

6.1.2.1 Thematic homogeneity

The ASK data is far less homogeneous in terms of subject matter than the texts used by Jarvis *et al.* Whereas the latter consisted of narrative descriptions of a short (8-minute) segment of the film *Modern Times*, the essays written for the IL test range over a total of 45 different topics, as shown in Table 13 on page 46. As a natural consequence of this, the ASK data is also less homogeneous in terms of vocabulary. This is especially the case for content words, but it is also true of some function words: for example, one can expect few if any future auxiliaries, such as 'will', in a text describing a scene from a film.

The difference in thematic homogeneity is reflected in the set of features used in each study. Compare for instance the list of 53 words used as features in the Jarvis *et al.* study (Figure 8 on page 33) with the list of 55 words used for the SP-group in the present study (Figure 13 on page 55): almost 50% of Jarvis *et al.*'s features are content words, whereas less than 10% of those used in the present study fall into that category. Although there appear to be no studies to date that have compared the degree to which lexical transfer occurs with content words as opposed to function words, it is at the very least safe to say that certain kinds of lexical transfer are less likely to be observed in the ASK data. An example of a likely cross-linguistic effect that would probably go unnoticed in more heterogeneous data is the tendency reported by Jarvis *et al.* (p. 60) for PO speakers to use the word 'woman' in preference to 'girl' – in stark contrast to the other L1 groups. It is therefore reasonable to hypothesize that greater predictive accuracy is can be achieved with data that is thematically homogeneous. In other words, the discrepancy in results obtained here and by Jarvis *et al.* could be due to this factor.[1]

6.1.2.2 Proficiency levels

The IL test data used in this study ranges from A2/B1 to B1/B2 on the CEFR scale, with just a few outliers at A2 and B2 levels (see Figure 12 on page 52); the Chaplin movie data is assumed to cover a wider range, from A2 to C1 (§2.3.2.1). If these assessments are correct,

[1] This could be tested using 10 ASK L1s with 200 features and comparing the results with the Jarvis & Paquot study, which involved 12 L1s and thematically heterogeneous data from ICLE (see §2.3.2.2). However the lack of information regarding possible correlations between topic and L1 group in the Jarvis *et al.* study would mean that the result would be inconclusive.

then the average proficiency levels are similar. Proficiency will then only be an explanatory factor for Jarvis *et al.*'s higher accuracy if there is a higher correlation between proficiency and L1 in the Chaplin movie data than there is in the ASK data. This may be the case, but no information is available on this issue.

6.1.2.3 Sample size

Jarvis *et al.* used 446 texts distributed somewhat unevenly across their five L1s: DA (n = 60), SW (n = 70), PO (n = 60), SP (n = 116) and FI (n = 140). The present study was based on 100 texts from each L1. The difference is not large and any effect it might have would probably tend to favour the study with the more even distribution. Thus it seems unlikely that the sample size can have played a role.

6.1.2.4 Choice of L1s

The same number of L1s was used in both studies, but the L1s themselves were different. In selecting L1s for the present study, DE, EN, PL and RU were chosen in order to mirror the two pairs of genetically related L1s in model study: DA, SW, PO and SP (see §4.3.1). Since there was no obvious candidate to play the role of FI as culturally-close-but-genetically-distant fifth L1, each of the other six L1s has been allowed to play this role in turn. Some of these (NL, SH and SP) are both culturally and genetically close to the base L1s, the others (SO, SQ and VI) are both culturally and genetically distant, and yet none of the six pentagroups, whatever their composition, yield accuracy rates anywhere close to those of Jarvis *et al.* The only conclusion to be drawn is that the choice of L1s did not play a role.

6.1.2.5 Type of language acquisition

The authors of the texts used by Jarvis *et al.* were FL students of EN living in their respective homelands: Brazil (PO), Mexico (SP), Denmark (DA) and Finland (FI and SW). The authors of the ASK texts, by contrast, were all SL learners of NO residing in Norway. It seems likely that the latter fact may have had a homogenising effect on the interlanguage of the NO learners: They will have been subjected to far less variation in educational style (which can cause L1 grouping effects), and the fact that they will have been exposed to the TL on a daily basis may also have had a levelling effect. The contrast between EFL and NOA may thus be an explanatory factor for the difference in results between the two studies.

6.1.2.6 Target language

Another factor that may have influenced the results is the use of different target languages. Because of its genetic affiliation, EN is close to DA and SW in terms of grammar and function word vocabulary, whereas it may be closer to PO and SP as regards content word vocabulary, due to the influence from SP's close relatives, FR and LA. NO, on the other hand, is close to DE, EN (and NL) in terms of grammar and both kinds of lexis, and more distant from PL, RU (and SH) in these respects. How this might affect the ability to detect L1 group membership remains a topic for future research, although it is worth recalling Jarvis' comment (see §2.1.2) that lemmatic transfer (in contrast to lexemic transfer) does not appear to be constrained by language distance.

To summarize: As far as data-related reasons are concerned, it is possible to hypothesize that the discrepancy between the accuracy rates achieved in the two studies can be traced to two main factors: (1) greater thematic heterogeneity in the ASK data, and (2) the fact that this data stems from second language rather than foreign language learners. Testing such a hypothesis is a topic for future research.

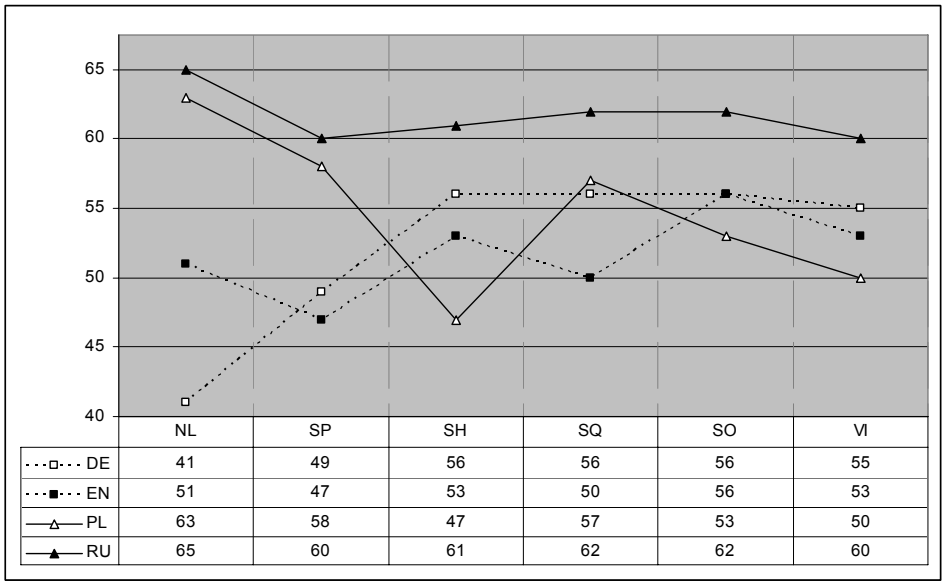
6.1.2.7 Methodological factors

In theory, differences in methodology could also account for the disparity between the results achieved by Jarvis *et al.* and those achieved in the present study; they might include the number of features and cases used, the use of different software, or the failure to use embedded stepwise feature selection. In practice, however, none of these seem likely to be the cause. In the first place, the present study had a slightly higher number of features per case (approx. 8.9 compared to Jarvis *et al.*'s 8.4). Secondly, the results achieved using R + MASS in the present study were comparable to those achieved using SPSS, the tool used by Jarvis *et al.* Thirdly, the use of 'up-front' rather than embedded feature selection should, if anything, lead to "overly optimistic results classification accuracy rates" (Jarvis & Paquot 2012: 81), not lower rates, as here.

In short, none of the methodological differences between the two studies can account for the discrepancy in the accuracy rates achieved.

6.2 Misclassification

The degree to which L1s are correctly predicted is the primary indicator of DA's ability to detect L1-related structure in the data. In addition, certain insights can also be attained by examining the pattern of misclassification, as Jarvis *et al.* do (see p. 34). In the following discussion the focus is on how misclassification varies across pentagroup and across classifiers.



	NL	SP	SH	SQ	SO	VI
DE	41	49	56	56	56	55
EN	51	47	53	50	56	53
PL	63	58	47	57	53	50
RU	65	60	61	62	62	60

Figure 26: Base L1 prediction accuracy as a function of 5th L1

Figure 26 plots the prediction accuracy achieved using R + MASS (see Table 20 on page 68) for each of the four base L1s in each pentagroup as a function of the fifth L1. The purpose of this chart is to see whether prediction accuracy for the four L1s used in all six analyses varies depending on which other L1 is included in the analysis and, if so, how: what might be termed the "5th L1 effect". It may be observed, first of all, that RU is consistently the most accurately predicted L1, with accuracies varying over a range of just 5% (60–65%). The corresponding figures for the other languages are EN 9% (47-56%), DE 15% (41–56%), and PL 16% (47–63%).

Some degree of correlation between changes in prediction accuracy and relatedness may be noted. In the case of DE, the inclusion of NL leads to a drop in DE's success rate to 41% (from an average of 52.2%). DE and NL are, of course, closely related, so one might hypothesize that similar kinds of transfer effects from the two L1s are causing difficulty in disentangling them from one another. Such a hypothesis is backed up by the observation (see Table 20D)

that the L1 most often mistaken for DE is NL, and that the L1 most often mistaken for NL is DE. A similar situation exists between PL and SH. With SH in the analysis, PL's success rate drops to 47% (from an average of 54.7%). The L1 most often mistaken for PL is SH, and *vice versa*. This would appear to be another instance of misclassification due to relatedness, like that observed by Jarvis *et al.* between DA and SW and between PO and SP (§2.3.2.1).

The same does *not* hold between EN and NL, which are also closely related. In fact, the success rate for EN is most affected by the inclusion of SP, dropping to 47% (from an average of 51.7%). Moreover, the L1 most often mistaken for EN is SP, while the L1 most often mistaken for SP is EN (see Table 20B). If transfer effects are at work here, a possible explanation for the high degree of confusion between EN and SP – despite their more distant genetic affiliation – is the high degree of language contact between EN and the Romance languages (primarily FR) and the influence of LA, and the consequently greater similarities in the lexicon of EN and SP. However, testing such a hypothesis requires a more detailed examination of the particular features underlying the EN-SP confusion.

The case of RU throws doubt on whether language transfer can explain the 5th L1 effect revealed in the analysis so far. As a Slavic language one would expect it to be affected in the same way as PL by the inclusion of SH, but it is not. A possible explanation for this is that one of the best L1 predictors (discussed below in §6.4.6) is the word *er* 'is', i.e. the present tense form of the copula, one of the most frequent forms in NO (and in all the learner texts), for which there are counterparts in PL and SH, but not in RU. This one typological difference could explain why the inclusion of SH does not lead to lower accuracy for RU as it does for PL.

In summary, the results from R + MASS suggests that misclassification can often be traced to a combination of genetic relatedness and/or typological similarity.

Comparisons may also be made across the different classifiers. In Table 29, each quadrant contains a sub-table based on results from one of the four classifiers. Each sub-table contains six columns, representing the six pentagroups, and five rows – one for each of the four base L1s, and one for the 5th L1 (labelled #5). Each cell specifies an L1 and should be read as follows:

- **First four rows** – for example, **RU** in column **SO** of row **DE** in sub-table MASS:
  In the SO-group MASS most often misclassifies DE as RU

- **Fifth row** – for example, **RU** in column **SO** of row **#5** in sub-table MASS: *MASS most often misclassifies SO as RU* [SO is of course only classified once per classifier]

MASS

	SO	SP	VI	NL	SH	SQ
DE	**RU**	RU	RU	NL	RU	RU/PL
EN	DE	SP	DE	DE	DE	DE
PL	DE	DE	VI	DE	SH	DE
RU	DE	DE/PL	DE/VI	PL	DE	SQ
#5	**RU**	EN	PL	DE	PL	RU

klaR

	SO	SP	VI	NL	SH	SQ
DE	RU	RU	PL	PL	PL	PL
EN	DE	DE	DE	DE	DE	PL
PL	RU	DE	RU	RU	SH	RU
RU	DE	DE	PL	PL	DE	PL
#5	RU	EN	*RU*	DE	PL	RU

SDDA

	SO	SP	VI	NL	SH	SQ
DE	RU	PL	PL	PL	RU	RU
EN	DE	DE	DE	DE	DE	DE
PL	RU	RU	VI	RU	SH	RU
RU	SO	PL	VI	PL	PL	DE
#5	RU	EN	PL	DE	PL	RU

SPSS

	SO	SP	VI	NL	SH	SQ
DE	RU	RU	RU	NL	RU	PL
EN	PL	SP	DE	DE	DE	PL
PL	DE	SP	VI	RU	SH	DE
RU	PL	DE	VI	PL	DE	PL
#5	RU	EN	PL	DE	PL	RU

Table 29: Misclassification rates

The MASS results show that DE tends to be confused with RU, except when NL is included, and this is confirmed by the SPSS results. All the classifiers show a strong tendency for EN to be confused with DE – except when SP is part of the mix. For PL the picture is less clear: MASS has it most often misclassified as DE (not RU, as one might expect based on genetic affiliation), unless VI or SH are included. The SPSS results, on the other hand, show it being variously misclassified as DE, SP, VI, RU or SH, depending on the pentagroup. RU, for its part, tends generally to be confused with either DE or PL.

The pattern concerning 5th L1s is quite consistent: (1) SO is most often confused with RU, (2) SP with EN, (3) VI with PL, (4) NL with DE, (5) SH with PL, and (6) SQ with RU. Three of these (2, 4 and 5) clearly implicate relatedness; the remainder (1, 3 and 6), on the other hand, involve the most distantly related language (SQ) and the two completely unrelated languages (SO and VI). Why SO and SQ should be most often confused with RU, and why VI should be most often misclassified as PL are questions for future work, but the roles of the copula and the indefinite article would be a good place to start.

6.3 L1 predictors

Four of the analyses in this study resulted in sets of "good" L1 predictors. None of these sets were identical, nor even of the same size:

(1) the ANOVA tests identified from 28 to 35 features, depending on pentagroup, whose means differed statistically across L1s; 29 (listed in Figure 19 on page 65) occurred in at least five of the six pentagroups;
(2) SDDA selected from seven to nine features, again depending on the pentagroup, and pooling those that were selected in three or more groups resulted in a list of 20 features (see Figure 21 on page 71);
(3) the tally using klaR and embedded feature selection varied from 41 to 45, with 28 features occurring in ten or more folds of the six 10-fold cross-validations (see Figure 22 on page 73); finally,
(4) SPSS selected from 25 to 34 features, with 23 being chosen for four or more groups of five L1s (see Figure 23 on page 75).

Despite all this variation there was a good deal of overlap in the sets of features that were chosen. Table 30 provides a consolidated list of features that were L1 predictor candidates in at least one of the four above-mentioned analyses, listed according to the number of analyses in which they were selected.

4	den, eller, en, og, skal, sted, viktig
3	andre, bare, barna, er, fra, i, jeg, man, norge, også, å
2	barn, bo, da, et, ha, ikke, kan, liker, må, men, norsk, om, på, veldig, vi
1	du, fordi, hvis, mer, mye, som, til, var, venner

Table 30: Consolidated set of L1 predictors

All but four of the words in this table (*ha*, *hvis*, *må* and *var*) are also to be found in the list of 53 features selected by SPSS (see Figure 23 on page 75). Since it contains the largest number of words, it is the latter set of features that was chosen for the remainder of the analysis in this project, with the addition of the two words *det* and *enn* (which were included in order to permit comparisons with the other definite article *den* and another conjunction, *som*).

andre, at, av, bare, ***barn****,* ***barna****,* ***bo****, da, de, den, det, du, eller, en, enn, er, et, for, fordi, fra, han, har, hun, i, ikke, jeg, kan,* ***liker****, man, mange, med, meg, men,* ***mennesker****, mer, min, mye,* ***norge****,* ***norsk****, når, og, også, om, på, skal, som,* ***sted****, så, til, veldig,* ***venner****, vi,* ***viktig****, være, å*

Figure 27: Final set of 55 L1 predictors

The complete set of L1 predictors is shown in Figure 27 with content words shown in bold in order to distinguish them from function words. These, then, are the words whose usage among learners of Norwegian varies the most across groups with different L1 backgrounds. It may be observed that over 80% (45/55) of them are function words. The significance of this fact will be discussed in §6.5.2.

Having arrived at these results, the final question to be answered is which, if any, of these predictors can be traced to language transfer. To approach this question it helps to know not only which features are good predictors of L1 group membership, but also which particular L1 groups each of them serves to discriminate. Jarvis *et al.* obtained that information via SNK tests (Jarvis *et al.* 2012: 59). As noted above, this test is not implemented in R, so for the present study Tukey HSD tests were used instead. A series of such tests, one for each feature, was applied to the set of 55 features using the R code shown in Figure 28.

```
for (X in myFeatures) {
  cat("Feature: ", X, "\n")
  print(summary(a1 <- aov(myData[,X] ~ myData$L1)))
  print(TukeyHSD(a1))
  print(sort(tapply(myData[,X], myData$L1, mean)))
}
```

Figure 28: R code for the Tukey HSD tests

Tukey's HSD is a single-step multiple comparison procedure and statistical test that is used in conjunction with an ANOVA to find means that are differ statistically from each other. The output that it generates can be interpreted to discover which groups show statistical separation on which feature. As an example, Figure 29 shows the output from R for the NL-group with the feature *skal*. The figures for adjusted probability in the column headed **`p adj`** show that the difference in frequency of use of *skal* between EN and DE learners (`en-de`) is not statistical ($p = 0.9979029$), while that between NL and DE learners (`nl-de`) very clearly is ($p = 0.0000570 << 0.001$).

```
Feature:  skal
              Df Sum Sq Mean Sq F value  Pr(>F)
myData$L1      4   1008  251.95   8.667 9.1e-07 ***
Residuals    495  14389   29.07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = myData[, X] ~ myData$L1)

$`myData$L1`
         diff       lwr        upr     p adj
en-de  0.239 -1.848562  2.3265618 0.9979029
nl-de  3.494  1.406438  5.5815618 0.0000570
pl-de  0.663 -1.424562  2.7505618 0.9079474
ru-de -0.561 -2.648562  1.5265618 0.9480543
nl-en  3.255  1.167438  5.3425618 0.0002276
pl-en  0.424 -1.663562  2.5115618 0.9811612
ru-en -0.800 -2.887562  1.2875618 0.8321757
pl-nl -2.831 -4.918562 -0.7434382 0.0021187
ru-nl -4.055 -6.142562 -1.9674382 0.0000016
ru-pl -1.224 -3.311562  0.8635618 0.4946648

   ru    de    en    pl    nl
1.562 2.123 2.362 2.786 5.617
```

Figure 29: Sample output from Tukey HSD test

The last two lines of the output list the five L1s and their means (for the feature *skal*) in ascending order of mean values. The mean for the RU-group is 1.562 occurrences per 1,000 words, that for the DE-group is 2.123, etc. In other words, RU learners of Norwegian use the word *skal* less often than DE learners, who themselves use it less often than EN and PL learners, etc., and the word is most often used by learners with an NL background:

```
RU < DE < EN < PL < NL
```

A more convenient way to view this information is in the form of a ‘homogeneity table’, such as that in Figure 30. This shows the five L1s in (ascending) order of mean value and displays a bar for each group of L1s whose means do not differ statistically from one another. It shows that learners whose L1 is RU, DE, EN or PL do not differ statistically in their frequency of use of the word *skal*; they thus constitute what is termed a ‘homogeneity group’ (shown as a bar connecting those four groups). In contrast, NL learners stand out as constituting a ‘homogeneity group’ of their own, since they use the word statistically more often than even the next most frequent group of users, PL. (The probability of the difference between PL and NL being due to chance, $p = 0.0021187$, can be read off from the `pl-nl` line in Figure 29.)

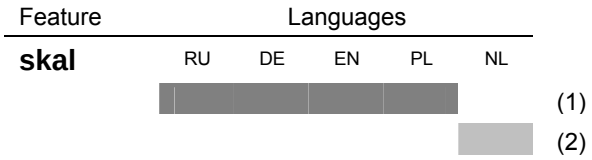


Figure 30: 5 L1 homogeneity table for $skal

Discovering that learners from NL backgrounds use the word *skal* statistically more often than users from RU, DE, EN and PL backgrounds, prompts the question, can this fact be traced to cross-linguistic influence from the learners' L1s? Similar questions may be posed regarding the patterns revealed for all the features selected as good L1 predictors. Approaching these questions using contrastive analysis (§2.1.4) has the potential to provide the third type of evidence called for by Jarvis: that of cross-language congruity. If it can be shown that the more frequent use of *skal* by NL speakers mirrors some aspect of the Dutch language, the case for cross-linguistic influence will be strengthened.[1]

6.4 Contrastive analysis

This section considers in more detail some of the L1 predictors and patterns uncovered so far. The aim is not to provide a complete analysis or a definitive answer to the question of whether or not they are instances of lexical transfer: that would require a separate study for each featural pattern. The purpose is rather to demonstrate how the results obtained through discriminant analysis can provide further analyzed and how they might offer new insights into lexical transfer. Before embarking on the contrastive analysis, however, it is important to consider mediating variables that might provide alternative explanations for the patterns in question and thus confound the analysis.

6.4.1 Mediating variables

When evaluating evidence for transfer it is important to bear in mind that there are many other factors that might potentially contribute to L1 grouping effects and lead to Type I errors, i.e. false positives. This section provides a brief overview of alternative explanatory factors that need to be controlled for (i.e. eliminated, held constant or stratified) in the case

[1] Jarvis himself considers CA that compares two languages in the abstract to produce only secondary evidence relevant to transfer. Primary evidence requires the analysis of actual language performance, since transfer is ultimately a phenomenon that occurs at the level of individuals. "That doesn't mean that we always need to examine individuals' L1 tendencies, but … we should always acknowledge their importance" (pc 2012-11-22).

of the ASK data. Jarvis (2000), elaborating Ringbom (1987), Biskup (1992), Ellis (1994) and Sjöholm (1995), provides a list of nine factors that can influence L2 acquisition: (1) age; (2) personality, motivation, and language aptitude; (3) social, educational, and cultural background; (4) language background (all previous L1s and L2s); (5) type and amount of target language exposure; (6) target language proficiency; (7) language distance between the L1 and target language; (8) task type and area of language use; and (9) prototypicality and markedness of the linguistic feature.

Personality, motivation, and language aptitude (2) cannot be assessed on the basis of the ASK data. Some aspects of the learners' social, educational, and cultural background (3) has been touched earlier in relation to the choice of L1s (§4.3.1) and proficiency level (§4.3.3), and gender is discussed below. Language background (4), at least as regards the L1, was the topic of §3.2 and §4.3.1 and is the variable under investigation in this study. Language distance (7) has also been discussed, both in connection with L1 background and elsewhere.

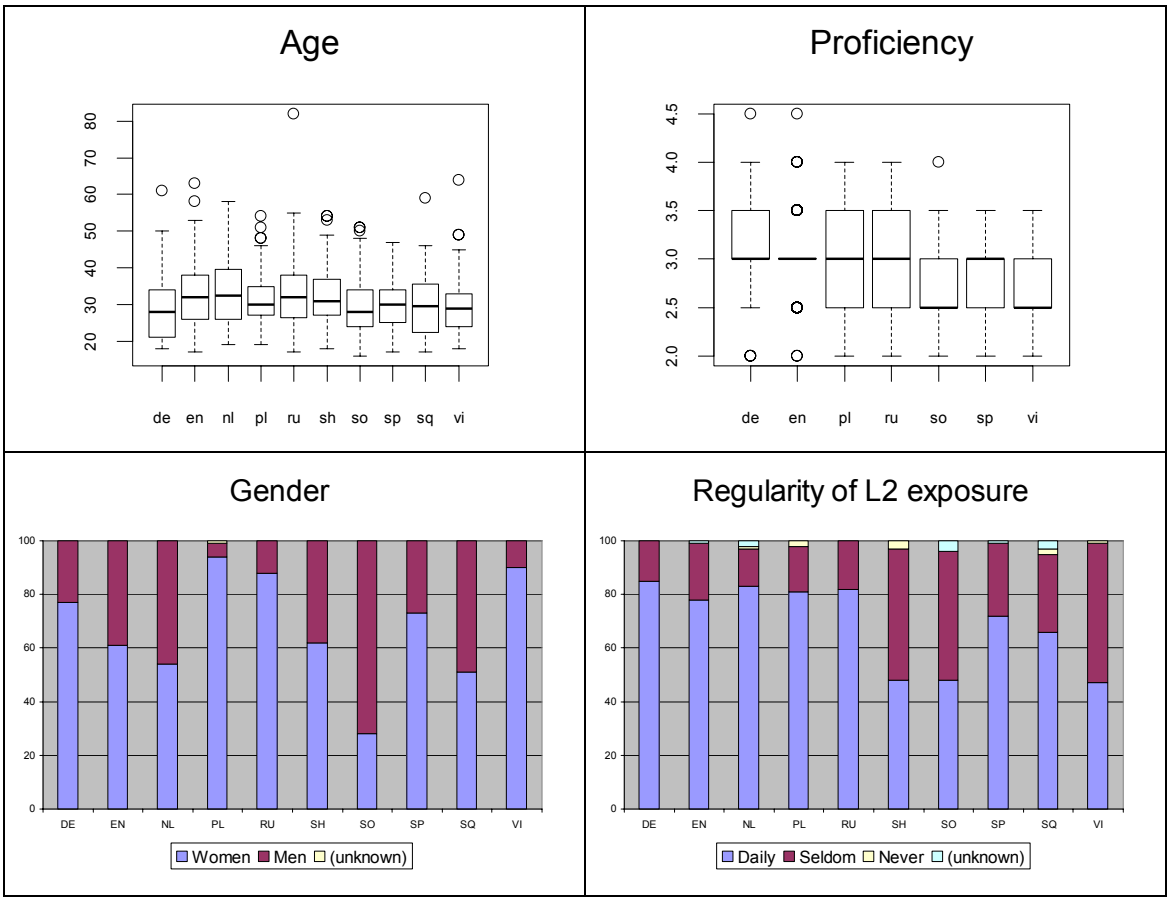


Figure 31: Mediating variables

In this section, age (1), gender (3) and type and amount of target language exposure (5) are covered, along with some further comments on proficiency (6). Figure 31 shows box plots for age and proficiency, and bar charts for gender and regularity of L2 exposure, all based on the extensive metadata available in ASK. The diagrams reveal a number of differences between L1 groups.

The learner's ages vary relatively little with the median age in the range 28-33 and there are relatively few outliers (with the honourable exception of an 82-year old RU woman); it thus seems to be unlikely to play a major role in predicting L1 group membership.

The box plot of proficiency provides an alternative presentation of the facts, this time in terms of a numeric scale arrived at by numbering the CEFR levels, as in Carlsen (2010b: 142).[1] It shows how five of the seven L1 groups for which CEFR data is available (see §3.4 on page 43), DE, EN, PL, RU and SP, all have median values of 3.0 (i.e. level B1), whereas the median value for SO and VI is lower at 2.5 (level A2/B1). Of the L1 groups whose median is 3.0, EN, PL and RU are evenly distributed (with EN speakers heavily concentrated around the median), while SP is weighted below the median and DE above it. Thus proficiency may play a role in separating SO and VI from the rest, but is unlikely to do so otherwise.

As for gender, the figures show a predominance of women in all L1 groups, with the exception of SO and SQ, both of which are overwhelmingly Muslim communities.[2] Men are particularly underrepresented in the PL, RU and VI groups. Whether or not these facts can have influenced the results is beyond the scope of this study.

Finally, regularity of exposure to the L1 shows a good deal of variation with the SH, SO and VI groups professing a lower level of interaction with native speakers than others. This clearly relates to cultural distance and also correlates with proficiency level, however the implications of these observation must be left to future research.

Mediating variables such as these will be taken into consideration as appropriate in the analyses to follow, in which selected features (those that seem to provide most separation between L1 groups) are discussed in terms of contrastive analysis.

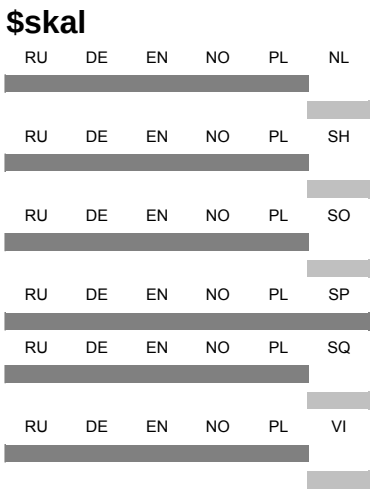
---

[1] 2=A2, 2.5=A2/B1, 3=B1, 3.5=B1/B2, etc.

[2] The SH L1 group consist in large part of Kosovo Albanian refugees.

6.4.2 Future tense: $skal

As observed in §6.3 and confirmed by the homogeneity table to the right,[1] the NL-group uses *skal* statistically more often than RU, DE, EN or PL. So too do speakers of SH, SO, SQ and VI. While there is a recognized general tendency for learners of Norwegian to overuse *skal* when forming the future tense (Mac Donald 1990: 27),[2] this does not explain the large discrepancy between RU, DE, EN or PL, on the one hand, and five of the other six L1 groups on the other.



One might hypothesize a correlation between proficiency level and overuse of *skal*, which would account for the pattern exhibited by SO and VI, two L1 groups that score below average in terms of CEFR ratings (see §3.4). Furthermore, since proficiency level tends to correlate negatively with linguistic and cultural distance, the same hypothesis could be extended to SH and SQ (two L1 groups for which CEFR data is not available). But that explanation seems unlikely to hold for NL speakers, who share many linguistic, cultural and socio-economic traits with EN and DE speakers, and can be assumed to be on a par with them in terms of proficiency.

One possible explanation is thematic bias. Certain topics are more concerned with future events than others. If there should be a preponderance of essays on such topics by NL speakers, this could explain the higher frequency of *skal*, a word that is mainly used to talk about future events. The word *skal* occurs 1,096 times in the data and is distributed very unevenly across texts that cover 44 of the 46 topics shown in Table 13 (see page 44 ff): a mere six topics account for more than 50% of the occurrences of *skal*. The topic with by far the most occurrences (269, or 24.5%) is – not surprisingly – the one entitled *Framtida* ('The future'). There are 63 essays on this topic, but only 8 of them are by NL speakers, as compared to 22, 10, 10 and 13 by speakers of SO, SP, SQ and VI, respectively.

---

[1] Extracts of the homogeneity tables given in the appendices (p. 131ff) are shown here for ease of reference.

[2] One reason for this may be the oversimplified account found in many Norwegian learner grammars. For example, Greftegreff (1985: 27) states categorically and without reservation: "Futurum lager vi av **skal** + infinitiv" ('The future is formed using **skal** + infinitive'). In actual fact the main temporal opposition in Norwegian is between past and non-past; the future is very often unmarked and the present tense is often used with future meaning, as in *Han reiser neste uke* lit. 'He travels next week' (Næss 2011a: 157).

Table 31 compares the number of occurrences per topic for DE, EN, NL and SP speakers for the six topics that have the most occurrences of *skal* amongst NL speakers. The word count (wc), number of texts (tc) and the ratio between them is shown for each L1 group. Thus, NL speakers produced 39 occurrences of *skal* in the 8 texts entitled *Framtida*, a ratio of 4.9 occurrences per text. This figure can be compared with that for SP speakers writing on the same topic, i.e. 2.9 – a considerable difference. Comparisons can also be made between EN and SP speakers for the topic *Bomiljø* ('Residential environment'), and between DE and EN speakers for the topic *Nyheter* ('News'). In each case the ratio of occurrences of *skal* per text is consistently higher for NL speakers. In other words, even when choice of topic is held constant, the predilection of NL speakers for the word *skal* is still very clear.

	DE			**EN**			**NL**			**SP**		
	wc	tc	ratio	wc	tc	ratio	wc	tc	ratio	wc	tc	ratio
Framtida	-	-	-	-	-	-	**39**	8	4.9	**29**	10	2.9
Bomiljø	-	-	-	**20**	38	0.5	**21**	16	1.3	**14**	23	0.6
Bolig og bosted	-	-	-	-	-	-	13	9	1.4	-	-	-
Frivillig hjelp i organisasjoner	2	5	0.4	-	-	-	9	2	4.5	-	-	-
Nyheter	**7**	10	0.7	**4**	9	0.4	**8**	7	1.1	2		-
Reise	-	-	-	-	-	-	8	14	0.6	-	-	-

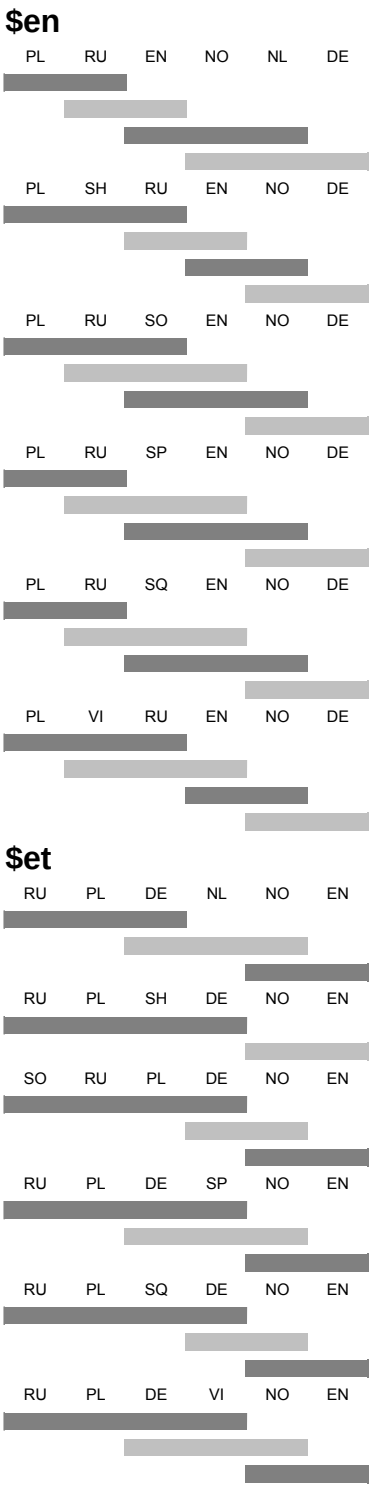
Table 31: Number of occurrences of 'skal' (by topic)

While neither proficiency level nor thematic bias can fully explain the pattern exhibited by NL speakers, there is an alternative explanation: In Dutch, the future tenses "are formed with the auxiliary *zullen*, shall, in all persons" (Koolhoven 1961: 44). This verb is cognate with NO *skulle* (the infinitive form of *skal*) and its present tense form *zal* closely resembles NO *skal* – in contrast to the various forms of the DE future tense marker, *werden* (i.e. *werde*, *wirst*, *wird*, *werden* and *werdet*). The EN form 'shall' is also cognate with *skal* and bears a strong formal resemblance to it, but it is much less frequent in EN than the alternatives, i.e. the enclitic ''ll' 'will' and the 'going to' construction,[1] which may explain why it does not trigger transfer to the same extent as NL *zal*. In conclusion, the strong tendency for NL speakers to overuse *skal* would appear to be a rather clear-cut case of formal lexical transfer.

[1] E.g. Wiktionary's TV/movie frequency lists give the following rankings: 'I'll' 82, 'you'll' 248, 'he'll' 540, 'she'll' 673, 'it'll' 543, 'we'll' 221, 'they'll' 700, 'will' 76, 'gonna' 108, 'shall' 849.

6.4.3 Indefinite articles: $en and $et

Speakers of the Slavic languages use the indefinite articles *en* and *et* much less frequently than learners from other L1 backgrounds. The relatively uncontroversial cross-linguistic explanation for this well-known fact is the lack of articles in the Slavic languages. The analysis of the ASK data confirms this fact (PL, RU and SH are all at the lower, left-hand end of the scale) and reveals the same tendency among speakers of VI and SO, which also lack articles. More interestingly, the analysis reveals the fact that DE speakers use the masculine form *en* statistically more often than all other L1 groups except NL. On the other hand, EN speakers use the neuter form *et* statistically more than every other L1 group.



The DE forms *ein* (m., n.) and *eine* (f.) bear a close formal resemblance to NO *en* (and even more to the form *ein* found in many varieties of NO), and the NL form *en* is identical. In the absence of alternative explanations, it seems rather likely that formal lexical transfer is responsible for the observed pattern.

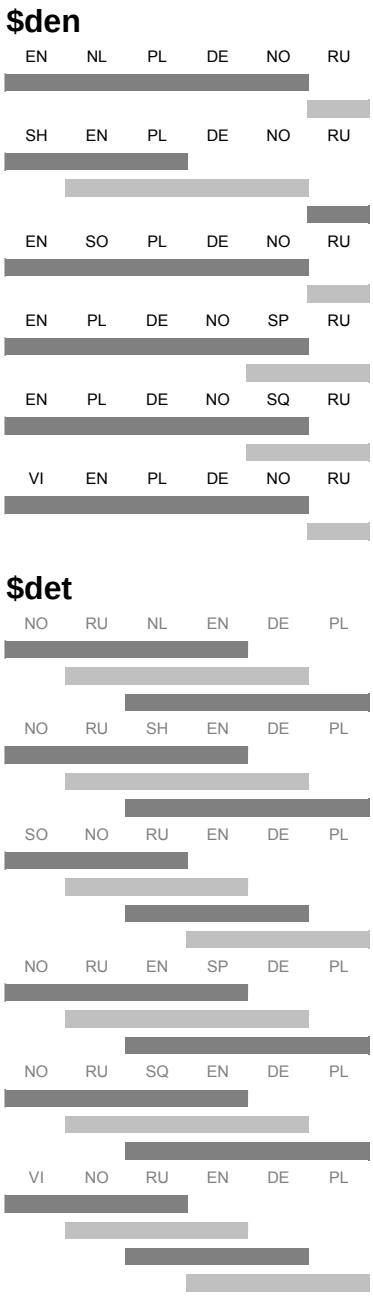
The greater formal similarity between EN 'a' (pronounced [ə]) and NO *et* (with its short vowel and relatively unobtrusive unvoiced plosive) than between [ə] and NO *en* may also suggest formal lexical transfer. Of course, the similarity is greater between NO *en* and EN 'an', but 'an' is much less frequent in EN than 'a'.[1] so the fact that the latter doesn't trigger transfer could be taken as evidence that frequency in the L1 plays a constraining role in lexical transfer.

Thus, L1 transfer appears to be at work in both the tendency for some speakers to underuse the indefinite article, for DE speakers to overuse *en*, and for EN speakers to overuse *et*.

[1] The Wiktionary rankings are 102 and 5 respectively, with 'a' used more than 11 times more often than 'an'.

6.4.4 Personal and demonstrative pronouns: $den and $det

The words *den* (m.) and *det* (n.) have a number of functions in Norwegian, as 3sg pronouns for non-human referents ('it'), as demonstrative pronouns ('that') and as preposed definite articles ('the') . *Det* is also used as the formal subject, as in *det regner* 'it's raining' (Golden *et al.* 2008: 133; Næss 2011a). The analysis shows that the form *den* is used statistically more often by RU speakers than by all the other L1 groups, with the exception of SQ and SP, and furthermore that SH speakers use it statistically less often than those three L1 groups. The form *det*, on the other hand, is used most frequently by PL speakers, and the difference compared to VI and SO speakers is again statistical.[1]



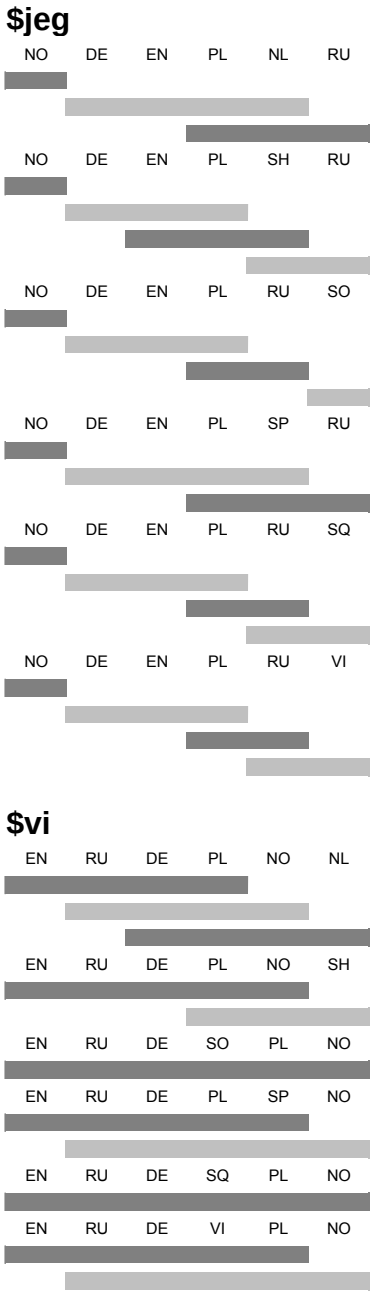
The most striking thing about this pattern is the difference between RU and PL: the absolute usage figures for *den* are PL 122, RU 166 (almost a 40:60 ratio), whereas the figures for *det* are PL 668, RU 496 (almost a 60:40 ratio). Why the RU and PL groups should differ so radically in their preferences is something of a mystery, since the 3sg personal pronouns are the same in PL (*on*, *ona*, *ono*) and RU (он, она, оно), so there is no phonological reason for one L1 group to prefer *den* and the other to prefer *det*. As regards the demonstrative pronouns (PL *ten*, *ta*, *to*, RU этот, это, эта), if the formal difference were to have any cross-linguistic effect, one would expect PL *ten* to be identified with NO *den*, which would produce the exact opposite effect to the one which is in fact found.

An explanation for this particular pattern must be sought in a detailed examination of how the particular usages break down across different functions, constructions and (perhaps) error types. However, such an analysis is beyond the scope of the present work.

[1] Could VI speakers underuse of *det* also be a sign of avoidance, since they otherwise tend to overgeneralize the use of the formal subject (Næss 2011b)?

6.4.5 Pronouns: $jeg and $vi

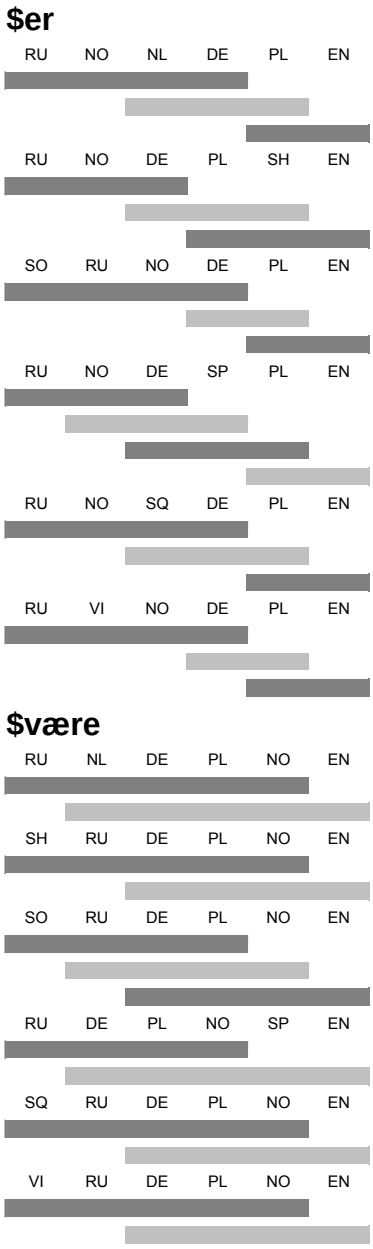
The usage pattern for the 1sg pronoun *jeg* separates DE and EN from RU, SH, SO, SQ, VI. The latter use *jeg* statistically more often than the former, with other L1 groups in-between. Interestingly, all the learner groups, including DE and EN, use the word statistically more often than native speakers.[1] The differences are less marked when it comes to the 1pl pronoun *vi*, but some of them are still statistical. The word is more frequent in native speaker texts than in those of the four base L1s, DE, EN, PL and RU. The EN speakers use it least of all, closely followed by RU speakers, and only NL and SH speakers use it more often than native speakers.



It is tempting to see an inverse correlation between overuse of *jeg* and proficiency level, since it is least used by native speakers and those L1 groups that exhibit the highest CEFR scores. However, Golden & Kulbrandstad (to appear), in an in-depth examination of SP and VI speakers' use of *jeg* and *vi* find no such correlation. They analyse their data in terms of two proficiency levels and find that both L1 groups use *jeg* more often at the upper level and *vi* more often at the lower level, a pattern which is generally upheld across different essay topics. Discussing cross-linguistic influence, Golden and Kulbrandstad observe that the only widely used personal pronoun in Vietnamese is in fact *tôi* 'I' (Rosén 2001: 31), which they believe might explain why VI speakers have a tendency to use *jeg* rather more than other L1 groups. Such a hypothesis would be confirmed by the fact that VI speakers use 2sg *du* 'you' far less frequently than other L1 groups (see Appendix H).

[1] Could this be a manifestation of the 'Law of Jante'? Used to describe an attitude towards individuality and success common in Scandinavia, the term refers to a mentality which de-emphasizes individual effort and places all emphasis on the collective, while discouraging those who stand out as achievers (Wikipedia).

6.4.6 Copula: $er, $være and $å



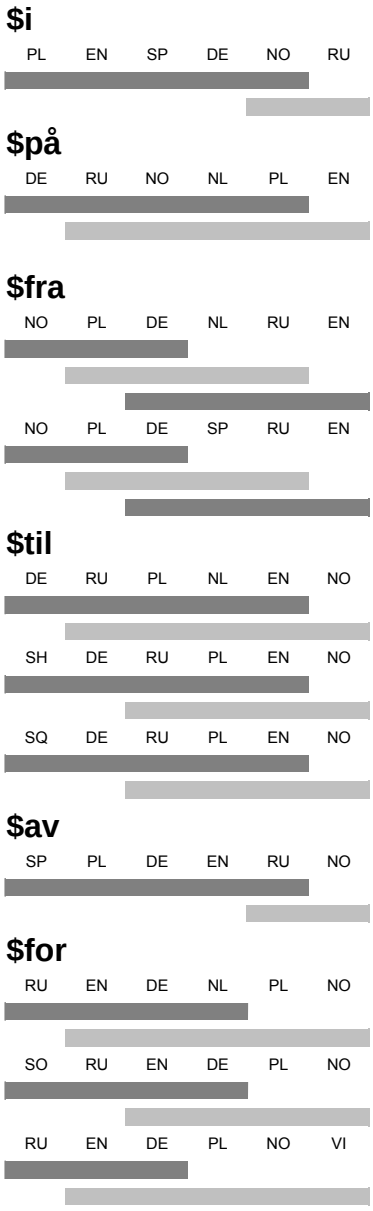
Speakers of EN use *er* 'is, are' more than all the other L1 groups (including native speakers), and statistically more so than every group with the exception of PL and SH. It is unlikely that such a pattern can be due to any of the mediating variables discussed in §6.4.1. A more likely explanation may be lexical transfer due to the strong formal resemblance between NO *er* and EN 'are' and the relatively high frequency of the verb 'to be' in EN (e.g. in the present continuous construction). Also of note is that PL and SH differ from fellow Slavs RU in their more frequent use of *er*. This most likely reflects the fact that PL and SH have a copula in the present tense (*być* and *биmи* / *biti*, respectively) whereas RU no longer does (cf. RU *дом там* vs. PL *dom jest tam*, SH *кућа је тамо* / *kuća je tamo* 'the house is there').

EN speakers also use *være* 'to be' more than any other group: statistically more so than RU, SH, SO, SQ and VI. In this case there is no formal resemblance, so the explanation must lie elsewhere, perhaps frequency in the L1. The contrast with RU is again explained by the absence of a copula in the present tense. The other L1s do have copulas, but some at least of them are used less often than EN 'to be'. For example, the VI copula *là* is not used with adjectives, since the latter are verbal, hence *tôi sợ chó* 'I am afraid of dogs', lit. 'I fear dog' (Wiull 2007: 46). In SO the verb *yahay* 'to be' contracts with adjectives, losing its root (*-ah-*) in the process (Saeed 1993: 185ff; Husby 2001: 51). There is no infinitive at all in SQ, so while there is an equivalent of NO *er* (*është*), there is no equivalent of *være* (Husby 1999: 23), and while SH does have a copula, it tends to be little used as a result of influence from other Balkan languages (Mønnesland 1990: 45; Lie 2005: 87).

A final, related observation is that the infinitive marker *å* is also used most frequently by EN speakers and least frequently by speakers of RU and SH, both of which lack an equivalent infinitive marker (see Appendix I).

6.4.7 Prepositions: $i and $på, $fra and $til, $av and $for

Prepositions, especially spatial prepositions, are renowned for being "among the hardest expressions to acquire when learning a second language" (Coventry & Garrod 2004: 4) and they have already been the subject of some interesting work based on ASK (Szymanska 2010; Malcher 2011). The homogeneity tables to the right show some of the patterns uncovered by the analysis, which can be briefly summarized as follows:



- *i* 'in' is used substantially more by RU speakers than by all other L1 groups;
- *på* 'on' is used the most by EN speakers and the least by DE speakers;
- *fra* 'from' is used statistically more often by EN speakers than by PL or native speakers
- *til* 'to' is underused by all L1 groups, especially DE, SH and SQ;
- *av* 'of' is used statistically more by native speakers than by all L1 groups, with the exception of RU.
- *for* 'for' is used statistically less by speakers of RU (and SO) than by speakers of PL (and VI).

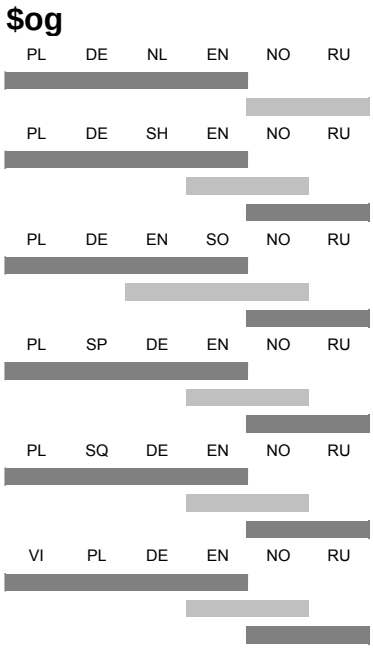
Explaining these patterns would require a detailed examination the individual usages, which is beyond the scope of the current work. For example, to understand the discrepancy between EN and DE use of *på* it would be necessary to check the token frequency of constructions in which the corresponding (prototypical) L1 forms (*on* and *auf*) are congruent in one of the L1s but not in the other, e.g.:

NO *på søndag* = EN 'on Sunday' but ≠ DE 'am Sonntag', whereas
NO *på engelsk* = DE 'auf Englisch' but ≠ EN 'in English'.

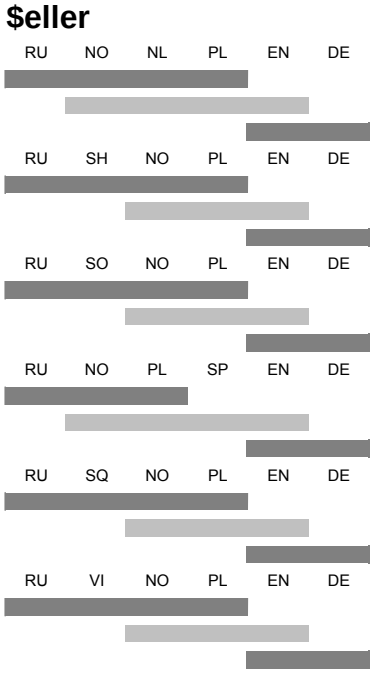
If shown to have a cross-linguistic basis, the patterns revealed here would be regarded as instances of what Jarvis calls collocational and subcategorization transfer (see §2.1.2).

6.4.8 Conjunctions: $og and $eller

Interesting patterns also emerge for some of the conjunctions: *Og* 'and' is used statistically more often by RU speakers than any other L1 group (but not native speakers), whereas *eller* 'or' is used less often by RU speakers than anyone else, and statistically more than EN and DE speakers, who use it most.



Why EN and DE speakers should overuse *eller* is unclear, as is the statistical difference between DE and NL in this regard. None of these patterns seem likely to be due to any of the mediating variables, so a contrastive explanation should probably be sought.



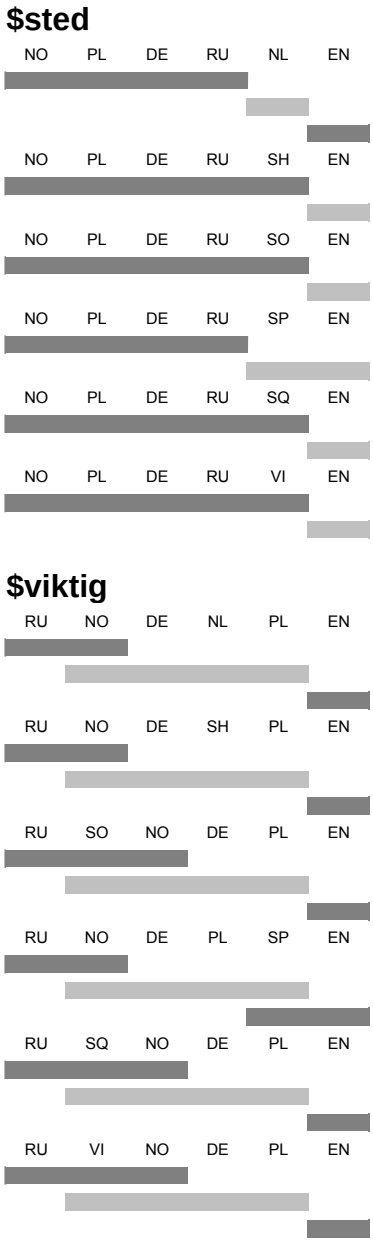
As for *og*, the striking contrast between PL and RU cannot be related to formal transfer, since PL *i* and RU и are phonologically identical. Furthermore, RU или 'or' closely resembles the NO form and yet is underused rather than overused, as might have been predicted. Since it has nothing to do with formal resemblance, a possible explanation for this pattern could be different token frequencies for the corresponding forms in the L1s. However, an initial attempt to investigate this route using the Wiktionary frequency lists[1] proved inconclusive: RU и is ranked as 1, whereas PL *i* is ranked as 2 (after *w* 'in'), and RU или is ranked as 54 whereas PL *lub* 'or' is ranked as 49. The differences, in other words, appear to be minimal.

Connectives are thus something of a mystery and there is clearly scope for research on their use across different L1 groups to complement that of Carlsen (2010a) on connectives across proficiency levels.[2]

[1] http://en.wiktionary.org/wiki/Wiktionary:FREQ

[2] Such research would also need to investigate why *men* 'but' is used significantly less often by EN speakers than by speakers of DE and NL (see Appendix I).

6.4.9 Content words: $viktig



Only 10 of the 53 words selected as L1 predictors by SPSS are content words, viz. *barn, barna, bo, liker, mennesker, norge, norsk, sted, venner, viktig*. Some of these are very likely to be prompt-induced, as discussed in §4.4.3 on page 53. For example, *barn* and *barna* 'child, children' (ind. sg/pl and def. pl, respectively) would be expected to occur particularly frequently in essays entitled *Barneoppdragelse* 'Parenting'. There are 53 essays on this particular topic and 23 of them (43%) were written by PL speakers. This clearly explains why the latter use these words statistically more often than most other L1 group (excepting only SH and SQ, which account for a further 34% of essays on this topic). Similar considerations probably apply to *bo* 'live'*, norge* 'Norway'*, norsk* 'Norwegian' and *venner* 'friends'; this could be easily checked by querying ASK for the distribution of these words (or their lemmas) by L1 relative to essay title.

Two content words that are not obviously prompt-related are *sted* 'place' and *viktig* 'important'. The former is generally evenly distributed across essay titles, except that 30% of the occurrences are found in essays entitled *Bomiljø* 'Residential environment'. Most of the essays on this topic (87%) were written by EN, SP and NL speakers, which explains the pattern shown in the homogeneity table (above). However, in ASK the word *sted* is used to replace specific place names in order to meet the requirements of the Norwegian Data Protection Act. It is thus overrepresented by definition and any observed differences across L1s are almost certainly topic-related.

The case of *viktig* 'important' is less clear. It is used statistically more by EN speakers than any other L1 group (except SP) and least of all by RU speakers. The explanation may again lie in token frequencies in the L1 (consider, for example, the frequency of *Important!* vs. *Achtung!*), but more work is needed to verify this.

6.5 Concluding remarks

6.5.1 Linguistic distance

The L1 predictors discussed in the preceding section are those that provide the highest degree of separation between one (or more) L1s and the others. What is apparent is that it tends to be the L1s that are most closely related to the target language that are easiest to separate. Thus, *skal* serves to separate NL from the rest; *en* and *et* separate DE and EN from each other, as does *på*, while *eller* separates the two of them from the rest; *er* and *være* separate EN, and so too does *viktig*.

There are some counterexamples to this tendency, for example the ability of the indefinite articles *en* and *et* to separate Slavic languages (but also SO and VI) from other Indo-European languages. However, it is rather striking that no good L1 predictors have emerged for the more distantly related languages SQ, SO and VI. This would appear to confirm the suggestion made by Ringbom that cross-linguistic similarity is an important factor in second language acquisition: "If you learn a language closely related to your L1, prior knowledge will be consistently useful, but if the languages are very distant, not much prior knowledge is relevant" (Ringbom 2007: 1). If the TL is perceived as being very different from the L1, there will be less reason to draw on the L1 in making hypotheses about the TL.

Ringbom cites the example of L3 EN learners who exhibit more transfer effects from their L2 SW than from their L1 FI. A corresponding example from the data in the present study may be the text `s0009`, written by a VI speaker but confidently predicted as EN in the LDA analysis using R + MASS (see Table 16 on page 62). The metadata in ASK shows this learner to have stated her homeland to be the UK and her level of English as "intermediate" (only four of the 105 VI speakers are more proficient in English than this). It appears that this learner's knowledge of L2 EN is causing transfer effects in her L3 production, thus confirming that language distance plays a role in transfer. This is also confirmed by misclassification rates in general, as discussed in §2.3.2.1 and §6.2.[1]

[1] Since the more distantly related L1s are also those which exhibit the lowest proficiency levels in the ASK corpus (see §3.4), an alternative hypothesis might be that good L1 predictors tend to emerge once a certain level of proficiency has been attained. However, this is a topic for future research.

6.5.2 Types of transfer

The cursory discussion of 20 or so L1 predictors in §6.4 showed fairly conclusive evidence of transfer in a few cases (in particular, *skal*, *en*, *et* and *er*) and tentative evidence in several more (*den*, *det*, *jeg*, *vi*, *være*, *å*, *i*, *på*, *fra*, *til*, *og*, *eller*, *viktig*). Assuming for the sake of argument that all of these really are examples of cross-linguistic influence on the learner's lexicon, it is instructive to try to classify them according to the taxonomy given in §2.1.2. That taxonomy defines three types of form-related transfer (false friends, code-switching and cross-linguistic blends) and four types of meaning-related transfer (semantic extension, calquing, collocational transfer and subcategorization transfer). The L1 predictors revealed in the present study seem to represent instances of four of these.

The false friends exhibit a formal resemblance to the corresponding word in the source language and are easy to identify: *skal* ↔ *zal*; *en* ↔ *ein / eine / en*; *et* ↔ 'a'; and *er* ↔ 'are'. It is important to recognize that these friends' are only "false" in the sense that their semantic or pragmatic ranges do not overlap exactly with those of the SL word. Thus it is not always an error to use *skal* in situations where NL would require *zal*, nor of course is it by any means always wrong for a DE speaker to use the indefinite article *en*. In other words, such cognates can occasion both positive and negative transfer. They are not always "false friends", as is the case with SW *offer* 'victim', the example given in §2.1.2.

Most of the other examples of transfer listed above (if such they be) could be classified as instances of semantic (or pragmatic) extension, although none of them as obviously as the "slips of the kieli" used to exemplify the category (see page 10). For example, the overuse of *jeg* by VI speakers may well result from the (correct) interlingual identification of *jeg* with *tôi* and the inappropriate use of the former in contexts that would be appropriate for the latter. The prepositions, on the other hand, could be classed as examples of what Jarvis terms collocational and subcategorization transfer, both of which would be regarded from a cognitive linguistic perspective as the transfer of "constructions" (Ellis 2003). And how might one classify PL, RU, SH, SO and VI speakers' low use of articles? Is this lexical transfer (from nowhere to nowhere) or grammatical transfer? Or is there in reality no meaningful distinction between lexicon and grammar, since "lexicon, morphology and syntax form a continuum of symbolic structures" (Langacker 1987: 3) – otherwise known as constructions (Croft 2007: 489)? That, indeed, is a topic for future research.

7. Conclusion

7.1 Research questions

This thesis represents the first attempt to apply data mining methods to the investigation of cross-linguistic influence on a language other than English, the first to be based on open source software, and the first to fully demonstrate the importance of contrastive analysis in complementing the statistical technique of discriminant analysis. It has raised more questions than it has answered, but the research questions themselves have been addressed.

Q1 *Can data mining techniques be used to identify the L1 background of Norwegian language learners based on their use of lexical features of the target language?*

The study confirmed the hypothesis that data mining techniques can indeed be used (up to a point) to identify the L1 background of a Norwegian language learner on the basis of his or her use of lexical features of the target language. The models constructed using the method of discriminant analysis were able to achieve prediction rates of up to 57.6% accuracy in when using 500 texts from five different L1 backgrounds. The probability of achieving such success rates by chance is infinitesimally small ($p < 2.2e\text{-}16$) and this alone proves that the lexical choices made by learners can be related to their L1 backgrounds.

Q2 *What are the best (lexical) source language predictors?*

The best source language predictors uncovered by the analysis are mostly function words, such as auxiliaries, articles, pronouns and prepositions. Some of them, such as the ability of indefinite articles to predict a Slavic L1 background, come as no surprise. Others, such as the tendency for Dutch speakers in particular to overuse *skal*, for German speakers to display a predilection for *en* rather than *et*, and for English speakers to do the opposite, were not at all

expected, which serves to confirm hypothesis H2, that data mining techniques are capable of revealing subtle patterns of learner language that might otherwise go undetected.

Q3 *Can those predictors be traced to cross-linguistic influence?*

The study also showed fairly conclusively that some predictors at least (e.g. *skal*, *en*, *et*, *er*) most likely are the result of language transfer, and it thus provided partial confirmation of hypothesis H3, that many (but not all) of the L1 predictors are traceable to cross-linguistic influence. Others are due to thematic bias and some have yet to be explained.

7.2 Methodology

The methods applied in this study provide three of the four types of evidence called for by Jarvis in his methodological requirements for transfer research. Two of these are provided "out of the box" when using discriminant analysis, because of its very design. When the rate of predictive accuracy it achieves is statistical – as it has been throughout this study – then the fact that it constructs its model by attempting to minimize within-group variance and maximize between-group variance automatically provides evidence of intragroup homogeneity and intergroup heterogeneity. The third type of evidence, cross-language congruity, is then provided through the application of contrastive analysis.

7.3 Contribution to knowledge

The study has therefore contributed to our knowledge of lexical transfer in a number of ways: It has validated the techniques developed by Jarvis & Crossley (2012) with a target language other than English, it has revealed some very interesting and subtle aspects of Norwegian interlanguage that have not been recognized earlier, and it has gone considerably beyond Jarvis & Crossley (2012) in applying contrastive analysis to provide more compelling cross-linguistic explanations of the patterns revealed by the statistical analysis.

The aspects of Norwegian interlanguage that have been uncovered provide confirmatory evidence, some of it more nuanced, for various types of lexical transfer and supply abundant material for further research based on ASK. In addition, the comparison with the results obtained by Jarvis *et al.* has raised interesting questions about the possible differences between second language acquisition and foreign language learning when it comes to cross-linguistic influence.

More importantly, perhaps, the techniques used here have the potential to open up whole new avenues of exploration within Norwegian SLA research using the wonderful data available in ASK, and also more broadly within Norwegian corpus linguistics. I also hope that the careful presentation of how discriminant analysis works (and its many areas of application) can convince even the most statistically challenged linguists that it is actually quite easy to harness its power in every area of empirical language research, and I will be more than happy to support any of my colleagues who wish to try it out.

7.4 Future directions

As for the future, this study has raised a lot of questions that could be followed up in new projects at the master's and doctoral level, and well beyond. An obvious starting point is to delve deeper into those aspects of L2 usage uncovered so far: testing the hypothesis in §6.4 experimentally, studying the unexplained predictor patterns, investigating specific errors connected with the individual lexical items, and following up areas of research ruled out of scope in this study. Some examples:

- Why did SDDA and klaR perform so badly (pp. 70 and 78) and how should embedded stepwise feature selection be implemented in R?
- Why are SO and SQ most often confused with RU, and why is VI most often misclassified as PL (p. 86)?
- How often (if at all) do DE speakers use *en* incorrectly and in what circumstances – and ditto with EN speakers' use of *et* (p. 95)?
- Why do PL and RU speakers differ so radically in their use of *den* and *det* (p. 96)?
- What are the precise forms of cross-language congruity that underlie the widely varying use of prepositions (p. 99)?
- What accounts for the different ways connectives are used across L1 groups (p. 100)?
- Do good L1 predictors tend to emerge once a certain level of proficiency has been attained (page 102)?
- Why have no good L1 predictors emerged for the more distantly related languages SQ, SO and VI (p. 102)?

Much work also remains to be done in terms of investigating possible correlations with proficiency level and cultural influence, and finding ways to eliminate, hold constant or stratify the many complex factors that influence L1 acquisition and that so easily can confound the study of transfer. If a common platform for handling such matters could be established within the framework of ASK, it would save each researcher from having to reinvent the wheel and lead to even more fruitful ways of exploiting this rich seam of data.

As for myself, I am drawn to the ideas of cognitive linguistics and I hope to be able to continue my work with data mining by investigating language transfer within that framework. The notions of constructions, chunking and frequency effects are central to cognitive linguists' view of language (Bybee 2010) and they therefore also play a major role in second language acquisition (Ellis 2003, 2012). It makes sense for me to regard the various kinds of 'gram' investigated by Jarvis and his colleagues as points along a continuum of constructions, from simple, contentful and fully specified 1-grams and *n*-grams, through the partly specified (and thus also partly schematic), such as those subjected to fleeting investigation by Mayfield Tomokiyo & Jones, through to fully schematic POS *n*-grams and beyond. That, at least, is the direction in which I would like to direct my energy in the coming years.

8. References

Andenæs, Ellen. 1984. Vietnamesisk og norsk språkstruktur. In Anne Hvenekilde & Else Ryen (eds.) Kan jeg få ordene dine, lærer? Oslo: Cappelen.

Arabski, Janusz (ed.) 2006. Cross-linguistic Influences in the Second Language Lexicon. Clevedon: Multilingual Matters.

Baayen, R.H. 2008. *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bellinger, David. 1980. Consistency in the pattern of change in mothers' speech: some discriminant analyses. *Journal of Child Language* 7, 469-487.

Bestgen, Yves, Sylviane Granger & Jennifer Thewissen. 2012. Error patterns and automatic L1 identification. In Jarvis & Crossley, 127-153.

Biskup, Danuta. 1992. L1 influence on learners' renderings of English collocations: A Polish/German empirical study. In Pierre J.L. Arnaud & Henri Béjoint (eds.) *Vocabulary and Applied Linguistics*. London: Macmillan, 85-93.

Bruland, Johan, Nguyên Thiên Co & Øivin Andersen. 1979. *Norsk-vietnamesisk kontrastiv grammatikk*. Bergen: Friundervisningens forlag.

Burns, Robert B. & Richard A. Burns. 2008. *Business Research Methods and Statistics Using SPSS*. London: SAGE Publications. (Additional chapters 22-25 available online at http://www.uk.sagepub.com/burns/chapters.htm.)

Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Campbell, Lyle & William J. Poser. 2008. *Language Classification: History and method*. Cambridge: Cambridge University Press.

Carlsen, Cecilie. 2010a. Discourse connectives across CEFR-levels: A corpus based study. In Inge Baartning, Maisa Martin & Ineka Vedder (eds.) *Communicative proficiency and*

linguistic development: intersections between SLA and language testing research. Eurosla monographs series 1.

Carlsen, Cecilie. 2010b. Å knytte ASK til Rammeverket – hvorfor og hvordan. In Hilde Johansen, Anne Golden, Jon Erik Hagen & Ann-Kristin Helland (eds.) *Systematisk, variert, men ikke tilfeldig. Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag.* Oslo: Novus forlag, 133-147.

Carlsen, Cecilie. 2012. Proficiency Level: A fuzzy variable in computer learner corpora. *Applied Linguistics Advanced Access*.

Clifford, David. 2010. *Package 'SDDA'* (2010-02-02), http://cran.r-project.org/.

Council of Europe (CoE). 2001. *Common European Framework of Reference for Languages: Learning, teaching assessment*. Cambridge: Cambridge University Press.

Croft, William. 2007. Construction Grammar. In Dirk Geeraerts & Herbert Cuyckens (eds.) *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press, 463-508.

Crossley, Scott & Danielle S. McNamara. 2012. Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. In Jarvis & Crossley, 106-126.

Dodge, Yadolah (ed.) 2003. *The Oxford Dictionary of Statistical Terms.* Oxford: Oxford University Press.

Dudoit, Sandrine, Jane Fridlyand & Terence P. Speed. 2002. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data, *Journal of the American Statistical Association* 97:457, 77–87.

Ellis, Nick C. 2003. Constructions, chunking, and connectionism: The emergence of second language structure. In Catherine J. Doughty & Michael H. Long (eds.) *The Handbook of Second Language Acquisition*. Oxford: Blackwell, 63-103.

Ellis, Nick C. 2012. Frequency-based accounts of second language acquisition. In Susan M. Gass & Alison Mackey (eds.) *The Routledge Handbook of Second Language Acquisition*. London: Routledge, 193-210.

Ellis, Rod. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Estival, Dominique, Tanja Gaustad, Son Bao Pham, Will Radford & Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific*

Association for Computational Linguistics (PACLING 2007), Melbourne, Australia, 31-39.

Everitt, Brian S. 2005. *An R and S-PLUS Companion to Multivariate Analysis*. London: Springer.

Field, Andy, Jeremy Miles & Zoë Field. 2012. Discovering Statistics Using R. London: SAGE.

Fisher, Ronald A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.

Fisiak, Jacek (ed.) 1981. *Contrastive linguistics and the language teacher*. Oxford: Pergamon Press.

Fletcher, Paul & Jo Peters. 1984. Characterizing language impairment in children: An exploratory study. *Language Testing* 1, 33-49.

Fox, Cynthia A. 1991. A discriminant analysis of yes-no questions in Quebec French. *Word* 42(3), 277-293.

Fries, Charles C. 1945. *Teaching & Learning English as a Foreign Language*. Ann Arbor, MI: The University of Michigan Press.

Ganschow, Leonore & Richard Sparks. 1996. Anxiety about Foreign Language Learning among High School Women. *The Modern Language Journal* 80(2), 199-212.

Gass, Susan M. 1996. Second language acquisition and linguistic theory: the role of language transfer. In William C. Ritchie and Tej K. Bhatia (eds.) *Handbook of Second Language Acquisition*. San Diego, CA: Academic Press, 317-45.

Gast, Volker (to appear). Contrastive Analysis. In Michael Byram & Adelheid Hu (eds.) *The Routledge Encyclopedia of Language Teaching and Learning*, 2nd Edition. London: Routledge.

Golden, Anne, Kirsti Mac Donald, Bjørg A. Michalsen & Else Ryen. 1990. *Hva er vanskelig i grammatikken? Sentrale emner i norsk som andrespråk.* Oslo: Universitetsforlaget.

Golden, Anne, Kirsti Mac Donald & Else Ryen. 2008. *Norsk som fremmedspråk. Grammatikk*. 3. utgave. Oslo: Universitetsforlaget.

Golden, Anne, Lise Iversen Kulbrandstad & Kari Tenfjord. 2007. Norsk andrespråks-forskning: Utviklingslinjer fra 1980 til 2005. *Nordand* 1(2), 5-37.

Golden, Anne & Lars Anders Kulbrandstad (to appear). *Peilinger i pronomen: 1. person entall & flertall: Aspekter ved pronomenvalg i tekster skrevet av voksne innlærere av norsk*

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse & Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2), 193-202.

Granger, Sylviane. 2008. Learner corpora in foreign language education. In N. Van Deusen-Scholl & N. H. Hornberger (eds.) *Second and Foreign Language Education*. New York: Springer, 337-351.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier & Magali Paquot (eds.) 2009. *The International Corpus of Learner English*. Version 2. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Greftegreff, Liv Astrid. 1985. *Enkel norsk grammatikk*. Oslo: NKS-Forlaget.

Hand, David J. 2005. Discriminant Analysis, Linear. In Peter Armitage & Theodore Colton (eds.) *Encyclopedia of Biostatistics*, 2nd Edition. Chichester: Wiley.

Haugen, Einar. 1976. *The Scandinavian Languages: An introduction to their history*. London: Faber & Faber.

Hsu, Jason C. 1996. *Multiple comparisons: Theory and methods*. London: Chapman & Hall.

Huberty, Carl J. 1994. *Applied Discriminant Analysis*. New York: Wiley.

Husby, Olaf. 1989. *Norsk-persisk kontrastiv grammatikk*. Oslo: Friundervisningens forlag.

Husby, Olaf. 1991. *Vietnamesisk grammatikk: en innføring i vietnamesisk grammatikk sett med norske øyne*. Oslo: Friundervisningens forlag.

Husby, Olaf. 1999. *En kort innføring i albansk*. Trondheim: Tapir.

Husby, Olaf. 2000. *En kort innføring i filipino*. Trondheim: Tapir.

Husby, Olaf. 2001. *En kort innføring i somali*. Trondheim: Tapir.

Hvenekilde, Anne (ed.) 1980. *Mellom to språk: 4 kontrastive språkstudier for lærere*. Oslo: Cappelen.

Hvenekilde, Anne (ed.) 1990. *Med to språk: Fem kontrastive språkstudier for lærere*. Oslo: Cappelen.

Hyltenstam, Kenneth (ed.) 1979. *Svenska i innvandrarperspektiv.* Lund: Liber.

James, Carl. 1980. *Contrastive Analysis*. Harlow: Longman.

Jarvis, Scott. 2000. Methodological rigour in the study of transfer: Identifying L1 influence in the interlanguage lexicon. *Language Learning* 50(2), 245-309.

Jarvis, Scott. 2009. Lexical transfer. In Aneta Pavlenko (ed.) *The Bilingual Mental Lexicon. Interdisciplinary approaches.* Bristol: Multilingual Matters.

Jarvis, Scott. 2010. Comparison-based and detection-based approaches to transfer research. *EUROSLA Yearbook* 10, 169-192.

Jarvis, Scott. 2011. Data mining with learner corpora: Choosing classifiers for L1 detection. In Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin & Magali Paquot (eds.) *A Taste for Corpora. In honour of Sylviane Granger*. Amsterdam: John Benjamins, 131-158.

Jarvis, Scott. 2012. The detection-based approach: An overview. In Jarvis & Crossley, 1-33.

Jarvis, Scott, Gabriela Castañeda-Jiménez & Rasmus Nielsen. 2004. Investigating L1 lexical transfer through learners' wordprints. Paper presented at the 2004 Second Language Research Forum. State College, Pennsylvania.

Jarvis, Scott & Aneta Pavlenko. 2008. *Crosslinguistic Influence in Language and Cognition*. London: Routledge.

Jarvis, Scott & Scott A. Crossley (eds.) 2012. *Approaching Language Transfer through Text Classification. Explorations in the detection-based approach*. Bristol: Multilingual Matters.

Jarvis, Scott & Magali Paquot. 2012. Exploring the role of n-grams in L1 identification. In Jarvis & Crossley, 71-105.

Jarvis, Scott, Yves Bestgen, Scott A. Crossley, Sylviane Granger, Magali Paquot, Jennifer Thewissen & Danielle S. McNamara [Jarvis, Bestgen *et al.*] 2012. The comparative and combined contributions of n-Grams, Coh-Metrix indices and error types in L1 classification of learner texts. In Jarvis & Crossley, 154-177.

Jarvis, Scott, Gabriela Castañeda-Jiménez & Rasmus Nielsen [Jarvis *et al.*] 2012. Detecting L2 writers' L1s on the basis of their lexical styles. In Jarvis & Crossley, 34-70.

Johansen, Hilde, Anne Golden, Jan Erik Hagen & Ann-Kristin Helland (eds.). 2010. *Systematisk, variert, men ikke tilfeldig. Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag*. Oslo: Novus.

Kabacoff, Robert. 2011. *R in Action: Data Analysis and Graphics with R*. New York: Manning.

Karlsson, Fred. 2008. *Finnish: An essential grammar*, 2nd Edition. Abingdon: Routledge.

Klecka, William R. 1980. *Discriminant Analysis*. Quantitative Applications in the Social Sciences 19. London: Sage Publications.

Kline, Rex B. 2004. *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

König, Ekkehard & Volker Gast. 2009. *Understanding English-German Contrasts*, 2nd Edition. Berlin: Erich Schmidt.

Koolhoven, H. 1961. *Teach yourself Dutch*. London: The English Universities Press.

Koppel, Moshe, Jonathan Schler & Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. Chicago: Association for Computing Machinery, 624-628.

Lachenbruch, Peter A. 1975. *Discriminant Analysis*. New York: Hafner.

Lado, Robert. 1957. *Linguistics Across Cultures. Applied linguistics for language teachers.* Ann Arbor, MI: University of Michigan Press.

Langacker, Ronald. 1987. *Foundations of Cognitive Grammar*. Vol. 1. Stanford, CA: Stanford University Press.

Larson-Hall, Jenifer. 2010. *A Guide to Doing Statistics in Second Language Research Using SPSS*. New York: Routledge.

Larson-Hall, Jenifer & Richard Herrington. 2010. Improving Data Analysis in Second Language Acquisition by Utilizing Modern Developments in Applied Statistics. *Applied Linguistics* 31(3), 368–390.

Lewis, M. Paul (ed.) 2009. *Ethnologue: Languages of the World*, 16th Edition. Dallas, TX: SIL International. Online version: http://www.ethnologue.com/.

Lie, Svein. 2005. *Kontrastiv grammatikk – med norsk i sentrum*, 3rd Edition. Oslo: Novus.

Ligges, Uwe. 2012. *Package 'klaR'* (2012-08.28), http://cran.r-project.org/.

Llach, María Pilar Agustín. 2010. An Overview of Variables Affecting Lexical Transfer in Writing: A Review Study. *International Journal of Linguistics* 2(1).

Loewen, Shawn & Hayo Reinders. 2011. *Key Concepts in Second Language Acquisition*. Basingstoke: Palgrave Macmillan.

Mac Donald, Kirsti. 1990. Uttrykk for framtid i norsk. In Golden *et al.*, 27-37.

Malcher, Jenny. 2011. *Jeg liker å treffe folk i café. Man må nyter de fine tingene på verden! Preposisjoner og morsmålstransfer – en korpusbasert studie med* i *og* på *i fokus.* Masters thesis, Department of Linguistics and Scandinavian Studies, University of Oslo.

Marckworth, Mary Lois & William J. Baker. 1980. A discriminant function analysis of co-variation of a number of syntactic devices in five prose genres. *Experimental Linguistics*, 231-246.

Mayfield Tomokiyo, Laura & Rosie Jones. 2001. You're not from 'round here, are you? Naive Bayes detection of non-native utterance text. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (NAACL '01). Cambridge, MA: Association for Computational Linguistics.

McNamara, Danielle S. & Arthur C. Graesser. 2011. Coh-Metrix: An automated tool for theoretical and applied natural language processing. In Philip M. McCarthy and Chutima Boonthum-Denecke (eds.) *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution*. Hershey, PA: IGI Global, 188-205.

Meurer, Paul. 2012. Corpuscle – a new corpus management platform for annotated corpora. In Gisle Andersen. *Exploring Newspaper Language. Using the web to create and investigate a large corpus of modern Norwegian*. Amsterdam: John Benjamins, 29-50.

Meyers-Scotton, Carol. 2006. *Multiple Voices: An introduction to bilingualism*. Oxford: Blackwell.

Mitchell, Rosamond & Florence Myles. 2004. *Second language learning theories*, 2nd Edition. London: Hodder Arnold.

Mustonen, Seppo. 1965. Multiple discriminant analysis in linguistic problems. *Statistical Methods in Linguistics* 4, 37-44.

Mønnesland, Svein. 1990. Serbokroatisk-norsk kontrastiv grammatikk. In Anne Hvenekilde (ed.).

Nicoladis, Elena. 1994. *Code-mixing in bilingual children*. PhD thesis, McGill University, Montreal.

Næss, Åshild. 2011a. *Global grammatikk: Språktypologi for språklærere*. Oslo : Gyldendal.

Næss, Åshild. 2011b. "Natur i Norge er det litt forskjellig": *Det*-setninger som topikaliserings-strategi i tekster skrevet av vietnamesiske innlærere. *NOA norsk som andrespråk* 2, 5-23.

Odlin, Terence. 1989. *Language Transfer. Cross-linguistic influence in language learning.* Cambridge: Cambridge University Press.

Odlin, Terence. 2003. Cross-Linguistic Influence. In Doughty & Long, 436-486.

Pallant, Julie. 2001. *SPSS Survival Manual*. Buckingham: Open University Press.

Peng, Roger D. & Nicolas W. Hengartner. 2002. Quantitative Analysis of Literary Styles. *The American Statistician* 56(3), 175-185.

Poulisse, Nanda. 1999. *Slips of the Tongue: Speech errors in first and second language production*. Philadelphia, PA: Benjamins.

R Development Core Team (RDCT) 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ringbom, Håkan. 1987. *The Role of the First Language in Second Language Learning*. Clevedon: Multilingual Matters.

Ringbom, Håkan. 2001. Lexical transfer in L3 production. In Jasone Cenoz, Britta Hufeisen and Ulrike Jessner (eds.) *Cross-linguistic Influence in Third Language Acquisition: Psycholinguistic perspectives*. Clevedon: Multilingual Matters, 59-68.

Ringbom, Håkan. 2007. *Cross-linguistic Similarity in Foreign Language Learning*. Clevedon: Multilingual Matters.

Rosén, Victoria. 1999. *Vietnamesisk - en kontrastiv og typologisk introduksjon*. Trondheim: Tapir.

Rosén, Victoria. 2001. *Vietnamesisk - en kontrastiv og typologisk introduksjon*, 2. utgave. Trondheim: Tapir.

Saaed, John Ibrahim. 1993. *Somali Reference Grammar*, 2nd Edition. Kensington, MD: Dunwoody Press.

Seaman, Michael A., Joel R. Levin & Ronald C. Serlin. 1991. New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin* 110(3), 577-586.

Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics*, 10, 209-241.

Slobin, Dan. 2004. The Many Ways to Search for a Frog: Linguistic typology and the expression of motion events. In Sven Strömqvist and Ludo Verhoven (eds.) *Relating Events in Narrative: Typological and contextual perspectives* (Vol. 2). Mahway, NJ: Lawrence Erlbaum Associates, 219-257.

Swan, Michael. 2007. History is not what happened: The case of contrastive analysis. *International Journal of Applied Linguistics* 17(3), 391-396. Accessed online 2012-08-17 at http://www.mikeswan.co.uk/elt-applied-linguistics/contrastive-analysis.htm.

Swan, Michael & Bernard Smith. 2001. *Learner English: A teacher's guide to interference and other problems*, 2nd Edition. Cambridge: Cambridge University Press.

Szymanska, Oliwia. 2010. A conceptual approach towards the use of prepositional phrases in Norwegian - the case of *i* and *på*. *Folia Scandinavica* 11, 173-183.

Talmy, Leonard. 1985. Lexicalization patterns: Semantic structure in lexical form. In Timothy Shopen (ed.) *Language Typology and Syntactic Description: Vol. 3 Grammatical categories and the lexicon*. Cambridge: Cambridge University Press, 57-149.

Talmy, Leonard. 2000. Toward a Cognitive Semantics: Typology and Process in Concept Structuring (Vol. 2). Cambridge, MA: MIT Press.

Tenfjord, Kari. 2007. ASK and you will find what you seek. In Cecilie Carlsen & Eli Moe (eds.) *A Human Touch to Language Testing*. Oslo: Novus Forlag 2007, 198-208.

Tenfjord, Kari, Jon Erik Hagen & Hilde Johansen. 2009. Norsk andrespråkskorpus (ASK): Design og metodologiske forutsetninger. *NOA* 25(1), 52-81.

Tomasello, Michael. 2003. *Constructing a Language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tono, Yukio. 2000. A corpus-based analysis of interlanguage development: Analysing part-of-speech tag sequences of EFL learner corpora. In Barbara Lewandowska-Tomaszczyk & Patricia James Melia (eds.) *PALC'99: Practical Applications in Language Corpora*. Frankfurt am Main: Peter Lang, 323-340.

Tsur, Oren & Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, Prague, Czech Republic, June 2007. Association for Computational Linguistics, 9-16.

Venables, William N. & Brian D. Ripley. 2002. *Modern Applied Statistics with S*, 4th Edition. New York: Springer.

Weinberg, Sharon Lawner & Sarah Knapp Abramovitz. 2002. *Data Analysis for the Behavioral Sciences Using SPSS*. Cambridge: Cambridge University Press.

Wilcox, Rand R. 2010. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*, 2nd Edition. New York: Springer.

Witten, Ian H. & Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition. San Francisco: Morgan Kaufmann.

Wiull, Hans Olaf. 2006a. *Bli bedre i norsk - se forskjellene mellom norsk og hindi*. Oslo: VOX.

Wiull, Hans Olaf. 2006b. *Bli bedre i norsk - se forskjellene mellom norsk og somali*. Oslo: VOX.

Wiull, Hans Olaf. 2007. *Bli bedre i norsk - se forskjellene mellom norsk og vietnamesisk*. Oslo: VOX.

Wiull, Hans Olaf. 2008. *Bli bedre i norsk - se forskjellene mellom norsk og persisk*. Oslo: VOX.

Wiull, Hans Olaf. 2009. *Bli bedre i norsk - se forskjellene mellom norsk og thai*. Oslo: VOX.

Wiull, Hans Olaf. 2010. *Bli bedre i norsk - se forskjellene mellom norsk og swahili*. Oslo: VOX.

Wong, Sze-Meng Jojo & Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association*. Cambridge: MA: Association for Computational Linguistics, 53-61.

Appendices

Appendix A. Glossary of terms

This appendix contains a short glossary of statistical and mathematical terms used in this dissertation. For those that describe objects in the model of discriminant analysis, their meaning in the context of the present study is ➔ also given.

10-fold CV cross-validation (q.v.) that is performed ten times, each time using one-tenth of the data as the test set and the remaining cases as the training set

ANOVA (analysis of variance) a statistical test of whether or not the means of several groups are all equal, see §4.5.2

case an individual or unit that belongs to a particular group (or class) and that exhibits certain features for which numeric values are available ➔ learner texts from the Norwegian Second Language Corpus (ASK)

coefficient a number or symbol multiplied with a variable (or unknown quantity) in an algebraic term, as 4 in the term $4x$, or a_1 in the term a_1x_1

cross-validation (CV) test method in which the the data is partioned into a training set and a test set, see §2.2.1.3

data set the set of data for the features and group membership of the cases that are used in an analysis ➔ word frequencies per 1,000 words and learner text metadata extracted from ASK

discriminating variable a variable that helps discriminate one group from another ➔ L1 predictors

F statistic the statistic that is calculated during an ANOVA (q.v.) and expresses the ratio of the between-groups variance and the within-groups variance

feature a numerical independent (or predictor) variable from which discriminant functions are constructed ➔ the words for which frequencies are used in the analysis

feature selection the process of selecting those features (q.v.) which contribute the most discriminating power to the statistical model

group a class into which cases are divided, expressed as a dependent, categorical variable ➔ the L1s of the learners who wrote the texts used in the analysis

homogeneity group a grouping of two or more groups (as defined above), each of whose mean value for some feature does not differ statistically from any of the others' mean values for the same feature, see §6.3

LOOCV (leave-one-out cross-validation) cross-validation (q.v.) that is performed N times (N = number of cases in the data set), each time using one case as the test set and the remaining cases as the training set

multivariate analysis the analysis of data which is 'multivariate' (i.e. in which each case exhibits multiple variables); includes classification, discriminant analysis, canonical correlation, factor analysis, component analysis, and various generalizations of homogeneity tests (Dodge 2003: 276)

MANOVA (multivariate analysis of variance) statistical test procedure for comparing multivariate (population) means of several groups

pentagroup (term invented for disambiguation purposes in the present study) ➔ a grouping of five L1 groups, consisting of data from the four base L1s (DE, EN, PL and RU) and one of the other six L1s (NL, SH, SO, SP, SQ and VI)

predictor see *discriminating variable*

variance a measure of how far a set of numbers is spread out

Appendix B. Word list with glosses

The following table contains the 200-word list of most frequently occurring words across both tests for the ten source languages of ASK, along with ranking, overall frequency counts, word class and English gloss.

word (10)	rank	count	pos	English	word (10)	rank	count	pos	English
alle	39	2152	pron	*all*	eksempel	137	629	noun	*example*
alltid	162	520	adv	*always*	eller	28	3176	conj	*or*
alt	80	1005	adj	*all*	en	9	9612	det	*a/an/one*
andre	38	2168	adj	*other/second*	enn	71	1091	prep	*than*
annet	196	414	adj	*other/second*	er	2	20322	aux	*is/are*
at	8	9787	pron	*that*	et	23	3881	det	*a/an/one*
av	19	5909	prep	*of*	etter	68	1192	prep	*after*
bare	45	1911	adv	*only*	familie	178	459	noun	*family*
barn	43	1941	noun	*child/children*	familien	153	538	noun	*family*
barna	58	1476	noun	*children*	finne	134	648	verb	*find*
bedre	101	812	adj	*better*	finnes	108	791	verb	*exist*
betyr	165	512	verb	*mean*	flere	73	1078	adj	*more*
ble	128	698	aux	*became*	fleste	180	457	adj	*most*
bli	64	1321	aux	*become*	folk	53	1718	noun	*people*
blir	51	1766	aux	*become*	for	11	9525	prep	*for*
blitt	181	456	aux	*become*	fordi	55	1529	conj	*because*
bo	131	679	verb	*live*	foreldrene	158	526	noun	*parents*
bor	116	775	verb	*live*	forskjellige	109	788	adj	*different*
bra	111	787	adj	*good*	fra	36	2417	prep	*from*
bruke	151	542	verb	*use*	før	112	786	prep	*before*
bruker	166	511	verb	*use*	først	164	514	adv	*first*
bør	133	652	aux	*should*	første	172	472	adj	*first*
både	110	788	conj	*both*	få	59	1451	verb	*get*
da	57	1492	adv	*when*	får	78	1017	verb	*get*
dag	81	983	noun	*day*	gang	192	417	noun	*time*
de	10	9529	pron	*the/they*	gjelder	170	494	verb	*concern*
del	173	471	noun	*part*	gjør	99	818	verb	*do*
dem	60	1411	pron	*them*	gjøre	76	1044	verb	*do*
den	26	3353	det	*the*	god	98	832	adj	*good*
denne	105	799	det	*this/that*	gode	187	447	adj	*good*
der	90	913	adv	*there*	godt	123	736	adj	*good*
deres	177	462	pron	*their*	grunn	148	559	noun	*reason*
derfor	100	818	adv	*therefore*	gå	89	913	verb	*go*
det	3	19971	det	*the*	går	130	682	verb	*go*
dette	49	1781	det	*this/that*	ha	48	1782	aux	*have*
disse	145	584	det	*these/those*	hadde	77	1036	aux	*had*
du	50	1778	pron	*you*	han	56	1528	pron	*he*

word (10)	rank	count	pos	English
har	15	7588	aux	*has/have*
hele	122	750	adj	*whole*
helt	156	529	adj	*whole*
her	115	777	adv	*here*
hjelpe	193	415	verb	*help*
hjemme	179	458	adv	*home*
hun	113	783	pron	*she*
hva	63	1377	pron	*what*
hver	142	612	det	*each*
hverandre	127	710	pron	*each other*
hvis	44	1930	conj	*if*
hvor	66	1255	pron	*where*
hvordan	119	763	adv	*how*
i	5	15130	prep	*in*
ikke	14	8062	adv	*not*
ingen	146	570	pron	*none*
jeg	6	13077	pron	*I*
jobb	104	800	noun	*job*
jobbe	185	452	verb	*work*
jobber	191	423	verb	*work*
kan	17	6247	aux	*can*
kanskje	117	774	adv	*perhaps*
kom	194	415	verb	*come*
kommer	74	1072	verb	*come*
kultur	200	399	noun	*culture*
kunne	87	921	aux	*could*
kvinner	95	857	noun	*women*
land	69	1146	noun	*country*
landet	157	526	noun	*country*
lese	171	484	verb	*read*
lett	174	466	adj	*easy/light*
liker	106	793	verb	*like*
litt	93	883	adv	*little*
liv	136	638	noun	*life*
livet	85	955	noun	*life*
lære	97	836	verb	*learn*
man	20	4465	pron	*one*
mange	25	3371	adj	*many*
mat	182	455	noun	*food*
med	16	6297	prep	*with*
meg	52	1763	pron	*me*
mellom	139	617	prep	*between*
men	21	4353	conj	*but*
mener	188	433	verb	*think*
menn	149	551	noun	*men*
mennesker	47	1797	noun	*person*
mer	40	2009	adj	*more*
min	61	1399	pron	*my*
mindre	169	495	adj	*less*
mine	198	401	pron	*my*
mitt	138	621	pron	*my*
mulig	190	427	adj	*possible*
mye	32	2725	adj	*much*
må	27	3324	aux	*must*
måte	168	499	noun	*way*
nesten	161	522	adv	*almost*
noe	62	1395	pron	*some/any*
noen	42	1948	pron	*some*
nok	152	541	adv	*enough*
nordmenn	184	453	noun	*Norwegians*
norge	37	2372	noun	*Norway*
norsk	107	792	adj	*Norwegian*
norske	120	754	adj	*Norwegian*
nye	154	533	adj	*new*
nå	72	1086	adv	*now*
når	35	2638	adv	*when*
ofte	92	892	adv	*often*
og	1	20406	conj	*and*
også	31	2875	adv	*also*
om	22	4011	prep	*whether*
opp	160	523	adv	*up*
oss	70	1136	pron	*us*
over	175	464	prep	*over*
penger	132	669	noun	*money*
problemer	189	431	noun	*problems*
på	12	9226	prep	*on*
samfunnet	126	713	noun	*society*
samme	143	607	adj	*same*
sammen	102	810	adv	*together*
se	103	810	verb	*see*
seg	29	3079	pron	*self*
selv	67	1218	pron	*self*
ser	135	641	verb	*see*
si	121	752	verb	*say*
siden	118	771	prep	*since*
sin	125	718	pron	*his/hers/its/their*
sine	91	906	pron	*their*
sitt	176	462	pron	*his/hers/its/their*

word (10)	rank	count	pos	English
skal	41	1967	aux	*shall*
skolen	167	501	noun	*school*
skulle	155	530	aux	*should*
slik	140	617	adv	*such*
snakke	159	525	verb	*talk*
som	7	10821	pron	*who/which/that*
sted	144	600	noun	*place*
stor	79	1009	adj	*big*
store	141	612	adj	*big*
synes	65	1294	verb	*think*
så	24	3560	adv	*so*
ta	94	865	verb	*take*
tenke	197	408	verb	*think*
tid	83	972	noun	*time*
tiden	183	455	noun	*time*
til	13	8609	prep	*to*
ting	84	965	noun	*thing*
to	147	563	det	*two*
trenger	88	921	verb	*need*
tror	82	977	verb	*believe*
ut	96	848	prep	*out*
uten	129	688	prep	*without*
vanskelig	114	779	adj	*difficult*
var	30	2936	aux	*was/were*
ved	150	544	prep	*by*
veldig	34	2704	adj	*very*
venner	86	944	noun	*friends*
verden	124	732	noun	*world*
vet	163	515	verb	*know*
vi	18	6028	pron	*we*
viktig	54	1641	adj	*important*
vil	46	1892	aux	*will/want*
ville	195	415	aux	*wanted*
være	33	2717	aux	*be*
vært	199	399	aux	*been*
ønsker	186	449	verb	*wish*
å	4	15857	inf	*to*
år	75	1064	noun	*year*

Appendix C. CEFR proficiency level descriptors

The two Norwegian language tests, the IL test and the AL test, correspond roughly to levels B1 and B2/C1 on the CEFR scale (see §3.4). For ease of reference, this appendix contains the descriptors for those three levels (CoE 2001: 24ff).

B1: Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.

B2: Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

C1: Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.

Appendix D. Linear discriminant plots

The figure below shows an LDA scatter plot matrix of all four discriminant functions for the SP-group. The LD1:LD2 plot in the first cell of row 2 (R2C1) is identical to Figure 18 on page 61 and shows substantial separation of DE (North), RU (West), EN (East) and PL (South), but little separation of SP (note that R1C2 is an inverted and rotated version of this plot). More separation (to the South-East) is found for SP in the LD3:LD4 plot (R3C4). RU has negative values for LD1 and thus tends to the South in all cells of row 1, whereas pl has negative values for LD2 (South in row 2), etc.

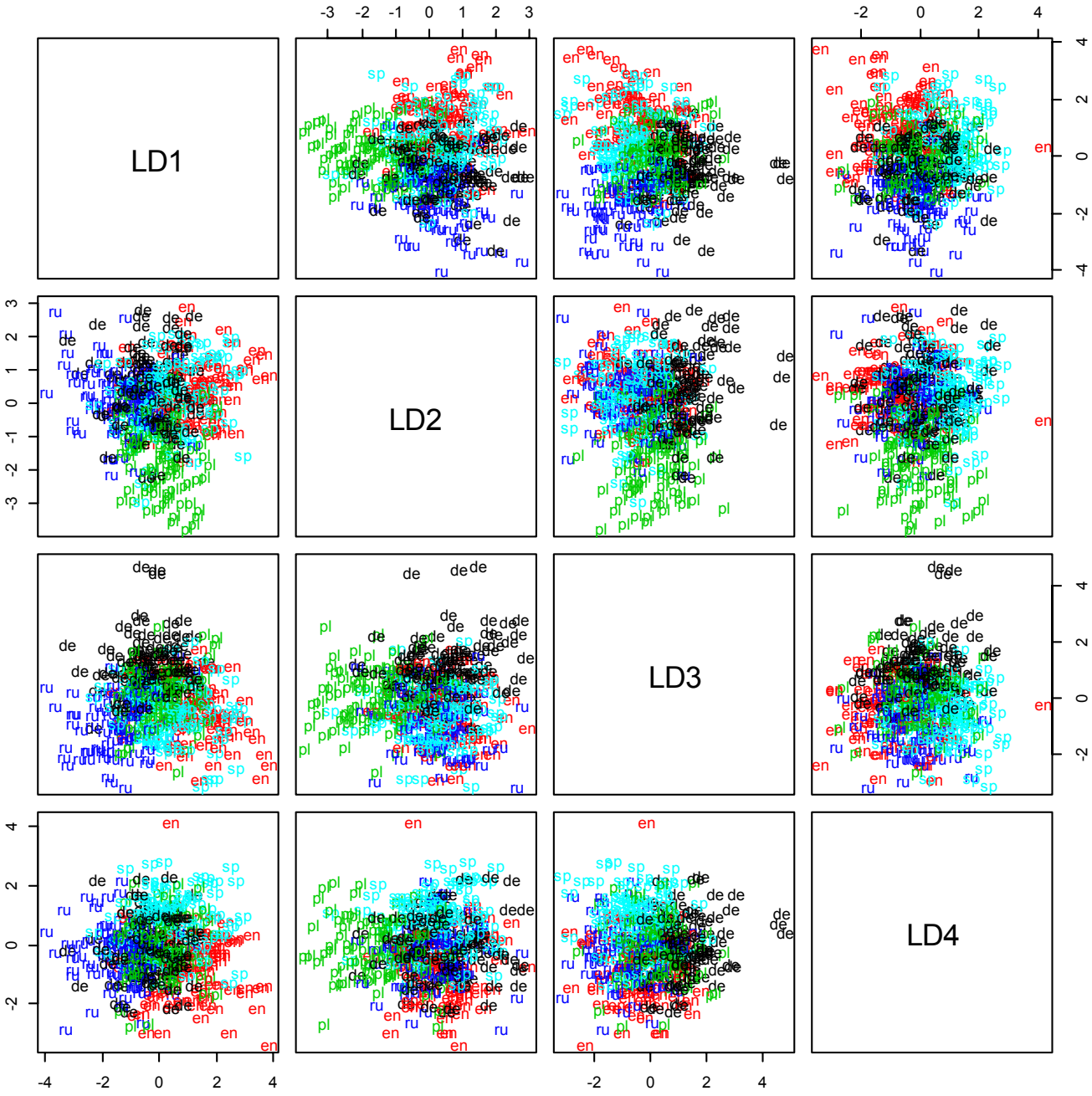


Figure 32: LDA scatter plot matrix for the SP-group (LD1, LD2, LD3 and LD4)

Appendix E. SPSS settings

This appendix contains screen shots showing the settings used for the discriminant analysis performed using SPSS, as discussed in §4.5.7 (page 64).

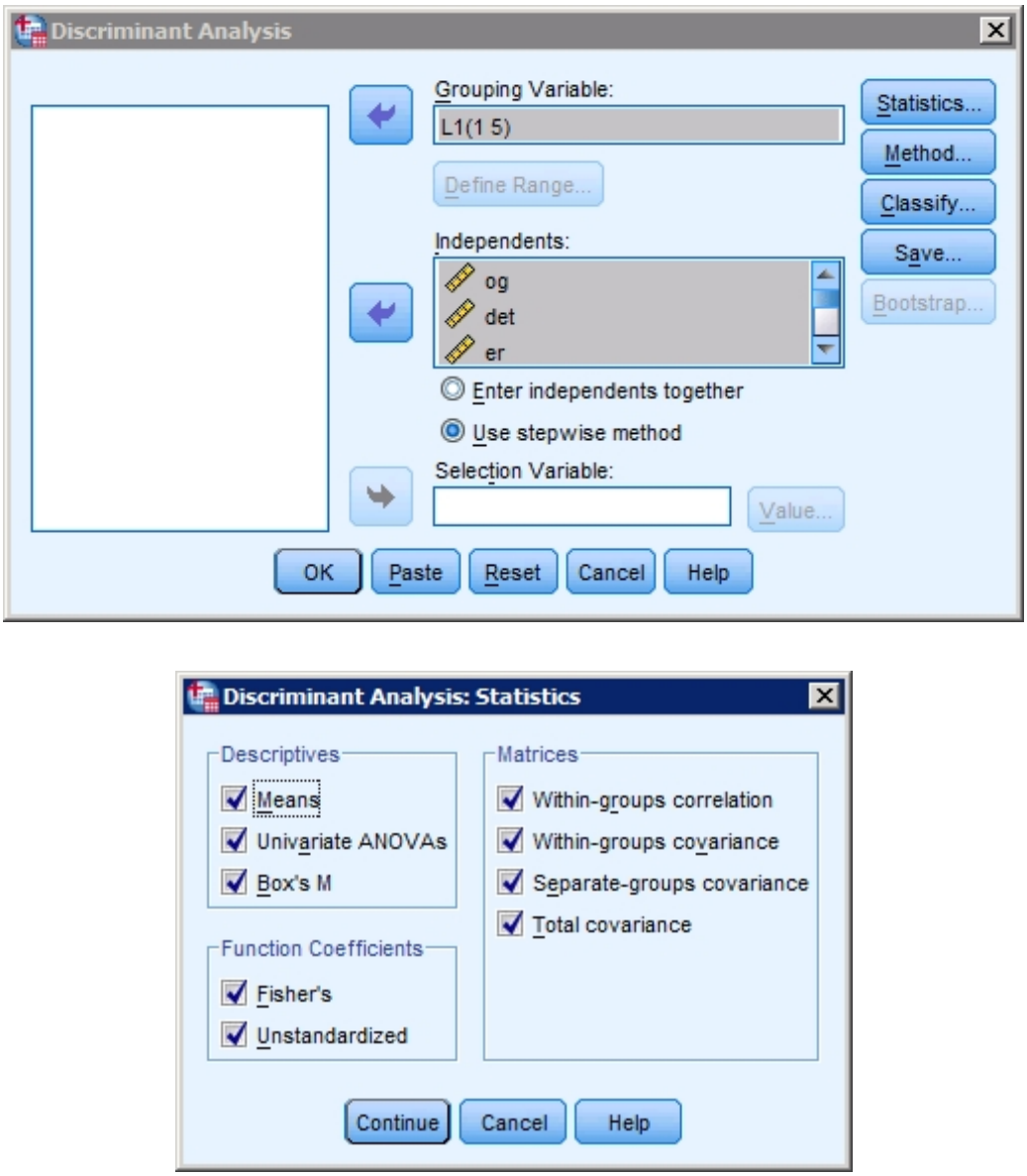
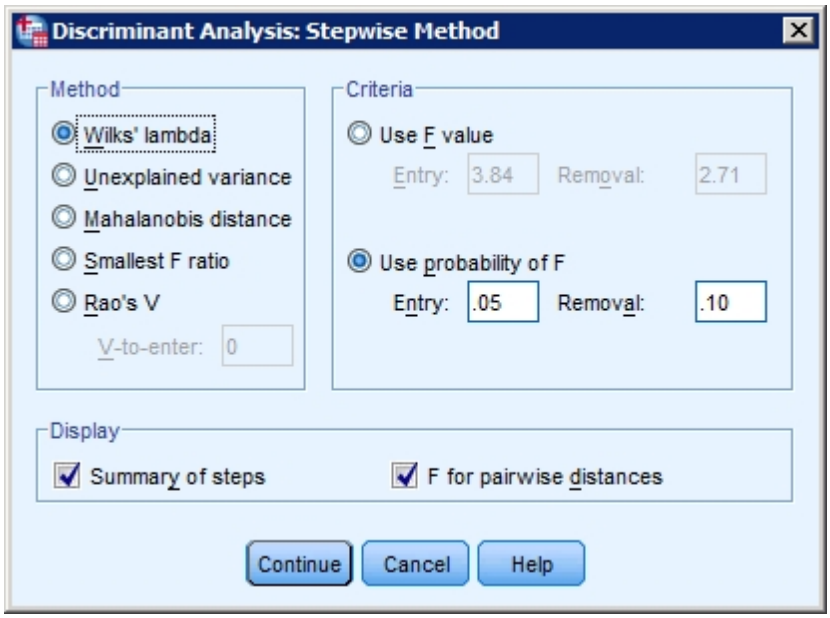


Figure 33: SPSS Discriminant Analysis settings



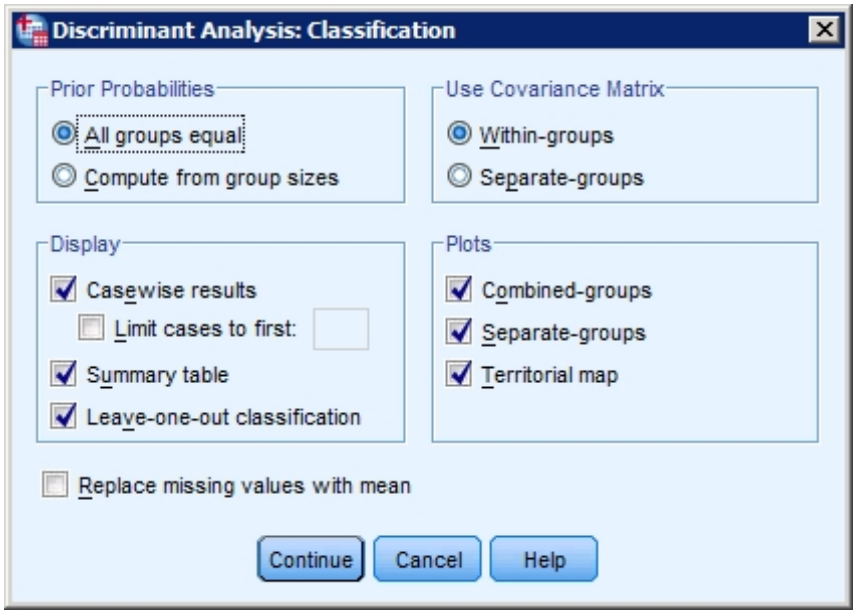


Figure 33: SPSS Discriminant Analysis settings (cont.)

Appendix F. Sample Omnimark script

This is the **mktop200** script that was used to modify the word frequency (distribution) files generated by ASK by removing the case sensitivity. Certain extensive comments have been removed. For the complete script, see the project web site.

```
cross-translate

macro /* arg comment */ is macro-end

/*
  NAME: mktop200.xom
  ------------------

  PURPOSE: Fix case-sensitivity in frequency (distribution) files
  ---------------------------------------------------------------

  Input:   1. List of words (e.g. 200 most frequent words) in LC only
           2. List of input file names (given on command line)
           3. Set of case-sensitive files containing frequency counts

  Output:  Consolidated file containing case-insensitive data (top200-data.txt)

*/

declare data-letters "Ã¥¦¸" "Ã…†˜"  ; for lower-casing of ÆØÅ in UTF-8

global counter wordfreqs variable initial-size 0
global counter row-ctr
global counter col-ctr

; At the start of processing, read the file "/data/Study 2/top200.txt"...
; Set up a shelf of counters (wordfreqs) for accumulating frequencies.

find-start
  local counter row-ctr initial {0}
  ; scan top200.txt row by row
  repeat scan file "t:\data\Study 2\top200.txt"
    match [any-text except "%t"]* => x.rank  "%t"
          [any-text except "%t"]* => x.word  "%t"
          [any-text except "%t"]* => x.count "%t"
          [any-text except "%t"]* => x.pos "%n"
      do when row-ctr = 0
        ; ignore header
      else
        ; create new counter item on wordfreqs shelf
        new wordfreqs key x.word
      done
      increment row-ctr
    match any+
      put error 'Error scanning file t:\data\Study 2\top200.txt%n'
      halt with 1
  again
  ; output header row immediately
  output "#"
  repeat over wordfreqs
    output "%t" || key of wordfreqs
  again
  output "%n"
  next group is start

; output statistics to screen
find-end
  local counter i
  set i to number of wordfreqs
  put #error "Size of shelf: %d(i)%n"
```

```
GROUP START ; ========================================================================

; Read list of files line by line, extract the name of the file and submit
; the contents of that file for further processing.

global stream column-headers variable initial-size 0

find any-text+ => x.infile '%n'*
  put #error x.infile || '%n'
  set row-ctr to 0
  clear column-headers
  using group main
   submit file "T:\data\Study 2\distribution\case-insensitive\%x(x.infile)"

GROUP MAIN ; =========================================================================

; Ignore summary lines at end of input file
find ('(Unweighted means)'|'(Total)') any-text+ '%n'*
  ; ignore

; Read the data file one line at a time
find any-text+ => x.line '%n'*
  local stream text-id
  ; reset values of word frequency counters
  repeat over wordfreqs
    set wordfreqs to 0
  again
  increment row-ctr
  set col-ctr to 1
  ; process each line
  repeat scan x.line
    ; first column
    match [any-text except "%t"]* =>x.cell "%t" when col-ctr = 1
      ; first column of first row
      do when row-ctr = 1
        ; value should be '#'
        put #error 'Expected "#" in R1C1, got [%x(x.cell)]...%n'
          unless x.cell = "#"
      else
        ; first column in rows other than the first contains the text ID
        set text-id to x.cell
      done
      increment col-ctr
    ; pairs of columns (excluding first column)
    match [any-text except "%t"]* =>x.cell "%t" [any-text except "%t"]* "%t"?
      ; any pair of subsequent columns (we're not interested in the second
      ; of these columns because it represents the percentage value)
      do when row-ctr = 1
        ; record which words are in which column
        set new column-headers key "%d(col-ctr)" to "%lx(x.cell)"
      else
        ; increment the wordfreqs item whose key is the col header for this col
        do unless column-headers key "%d(col-ctr)" = "(sum)"
         using wordfreqs key column-headers key "%d(col-ctr)"
          set wordfreqs to wordfreqs + x.cell
        done
      done
      increment col-ctr by 2
  again
  do unless row-ctr = 1
    ; output consolidated frequency counts for this line (text)
    output "%g(text-id)"
    repeat over wordfreqs
      output "%t%d(wordfreqs)"
    again
    output "%n"
  done
```

Appendix G. Sample R script

This is the **tukey.r** script that was used to run Tukey HSD tests producing output that was post-processed by Omnimark and prepared for import to Excel (where conditional formatting was used to create the bar effects in the homogeneity tables).

```
# tukey.r: post-hoc Tukey HSD tests
# ------------------------------------------------------
#
# This script runs one-way ANOVA tests followed by Tukey HSD tests
# The output is written to file in a format designed to be post-processed
# by the Omnimark script tukey.xom in order to generate tab-delimited
# tables of homogeneity groups, which in turn is input to the Excel sheet
# tukey.xls where some clever conditional formatting is applied to create
# the barred homogeneity tables.

# This script has been extended over time and now offers four variants of
# the analysis. The one used in the appendices is from #4 and includes
# Norwegian data in the pentagroups. It should really be rewritten, but it
# ain't broke, so why bother to fix it?
#

start_time <- Sys.time()
setwd("T:/data/Study 1/tukey")

# Which analysis to perform?
#
# First:  42 features across 10 languages (WS Predictor candidates in Study 1.xls)
# Second: 6x 10 features across 5 languages (< spss-features2.txt)
# Third:  6x 53 features across 5 languages (< spss-features3.txt)
# Fourth: -ditto- plus NO in each pentagroup
analysis <- 1
analysis <- 2
analysis <- 2
analysis <- 3
analysis <- 4

# Set up some common variables
base_langs <- c("de", "en", "pl","ru")  # set of base L1s
add_langs <- c("sp", "so", "vi", "nl", "sh", "sq")

if ( analysis == 1 ) {

  # Analysis #1
  # Uses consolidated list of 42 features from four analysis methods
  # (see WS Predictor candidates in Study 1.xls)
  #
  data_set = read.table("T:/data/Study 2/Study 2 data.txt", header=TRUE)
  extraCols = 3   # non-numeric columns (currently L1, Language & CEFR)
  # These are the original 42 features (Table 30)
  myFeatures <- c("den", "eller", "en", "og", "skal", "sted", "viktig", "å",
"andre",
                  "bare", "barna", "er", "fra", "i", "jeg", "man", "norge", "også",
                  "barn", "bo", "da", "et", "ha", "ikke", "kan", "liker", "må",
"men",
                  "norsk", "om", "på", "veldig", "vi", "du", "fordi", "hvis",
"mer",
                  "mye", "som", "til", "var", "venner")
  myLangs <- c(base_langs, add_langs)
  # Strip rows where L1 not in myLangs (e.g. Norwegian)
  myData = droplevels(subset( data_set, L1 %in% myLangs))
  cols = ncol(myData) - extraCols

  # Redirect output to a file
  sink("10L1.txt")

  # Output list of column names (just for reference)
  print(myFeatures)
```

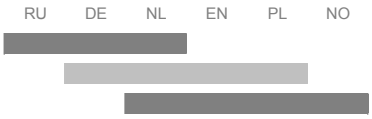
```
  # Output results of one-way ANOVA, Tukey HSD and (sorted) means
  for (X in myFeatures) {
    cat("\n===========\n\nFeature: ", X, "\n")
    print(summary(a1 <- aov(myData[,X] ~ myData$L1)))
    print(TukeyHSD(a1))
    print(sort(tapply(myData[,X], myData$L1, mean)))
  }

  # Stop redirecting output
  sink()
} else {

  # 2. Analysis #2, #3 and #4, which are essentially the same except for using
different
  # sets of features and, in the case of #4, including NO data in the homogeneity
tables.
  #
  # Get the list of features from a text file
  allFeatures = read.table("T:/data/Study 1/spss/spss-features2.txt", header=T)
  if ( analysis >= 3 )
    allFeatures = read.table("T:/data/Study 1/spss/spss-features3.txt", header=T)

  # Redirect output to a file
  if ( analysis == 3 ) {
    sink("6x5L1.txt") } else {
    sink("6x6L1.txt")
    base_langs <- c("no", base_langs)  # add Norwegian to base L1s
  }

  # Get complete data set
  data_set = read.table("/data/Study 2/Study 2 data.txt", header=TRUE)

  # Iterate over each of the six additional languages
  for (add_lang in names(allFeatures)) {

    cat("Languages: de en pl ru ", add_lang, "\n", sep="")
    myLangs <- c(base_langs, add_lang)
    myFeatures <- as.character(allFeatures[add_lang][,1])
    # Create subset of data for these L1s and features (Språkprøven only)
    myData = subset( data_set,
        L1 %in% myLangs & grepl('^s',rownames(data_set)), select =
c(myFeatures,"L1"))

    # Output list of column names (just for reference)
    print(myFeatures)
    # Output results of one-way ANOVA, Tukey HSD and (sorted) means
    for (X in myFeatures) {
      if ( X %in% names(myData) ) {
        cat("\n===========\n\nFeature: ", X, "\n")
        print(summary(a1 <- aov(myData[,X] ~ myData$L1)))
        print(TukeyHSD(a1))
        print(sort(tapply(myData[,X], myData$L1, mean)))
      } #end if
    } #end for(X in myFeatures)
  } #end for(add_lang in names(allFeatures))

  # Stop redirecting output
  sink()

} #end if(analysis == 1)

elapsed_time <- round(Sys.time() - start_time, 1)
print(elapsed_time)
```

Appendix H. Homogeneity tables (by pentagroup)

This appendix contains the results of the Tukey HSD tests described in §6.46.3, ordered by pentagroup.

DE EN PL RU + Dutch

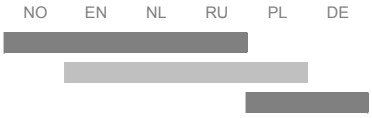
$andre



$at

$av

$bare

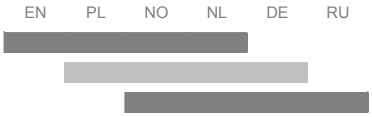


$barn

$barna

$bo

$da



$de

$den

$det

$du

$eller

$en

$er

$set

$for

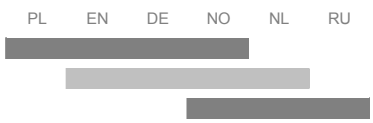
$fordi

$fra

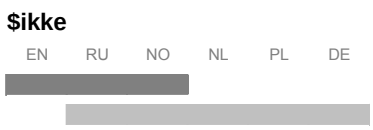
$han

$har

$hun



$i



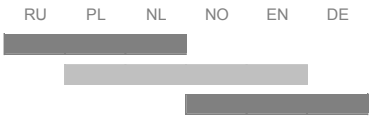
$ikke

$jeg

$kan

$liker

$man

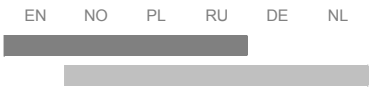


$mange

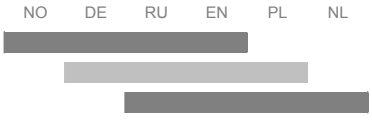
$med

$meg

$men



$mennesker



$mer

$min

$mye

$når

$norge

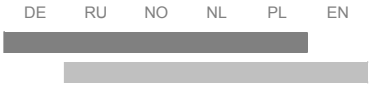
$norsk

$og

$også

$om

$på



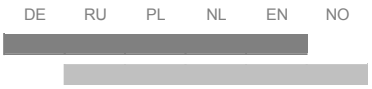
$så

$skal

$som

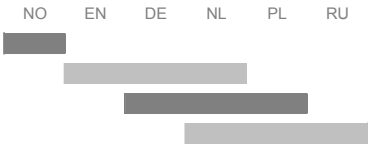
$sted

$til



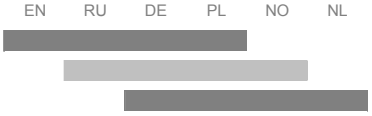
$være

$veldig



$venner

$vi

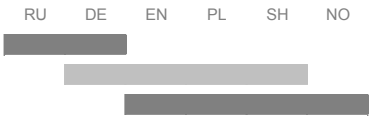


$viktig

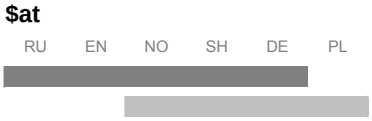
$å

DE EN PL RU + Serbo-Croat

$andre

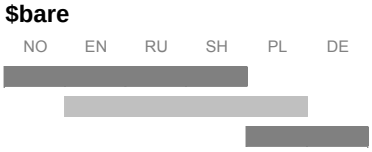


$at

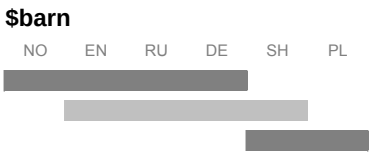


$av

$bare



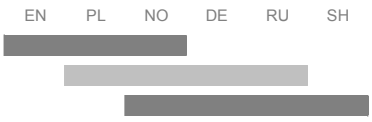
$barn



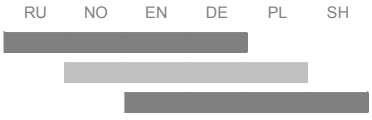
$barna

$bo

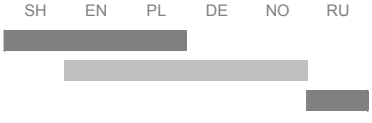
$da



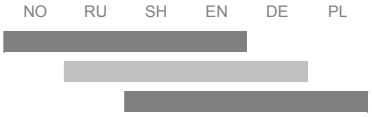
$de



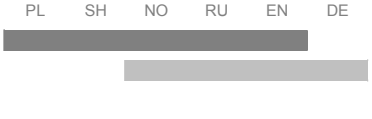
$den



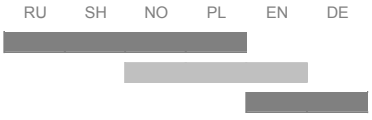
$det



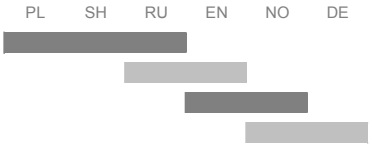
$du



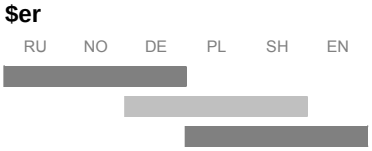
$eller



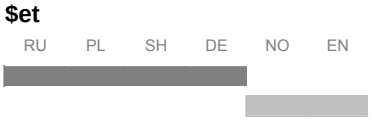
$en



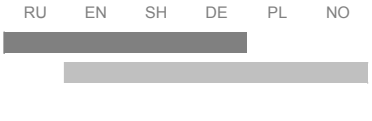
$er



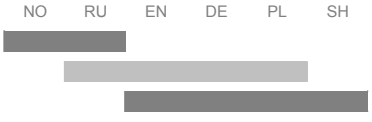
$et



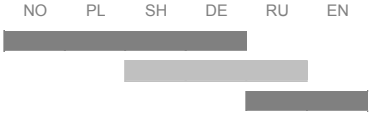
$for



$fordi



$fra



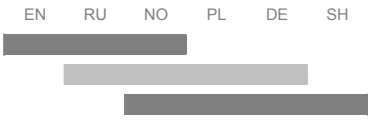
$han

$har

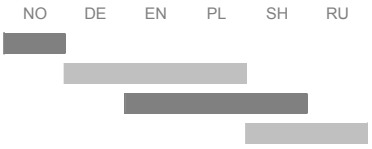
$hun

$i

$ikke

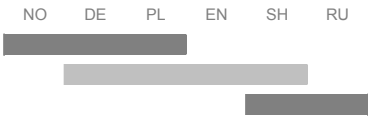


$jeg

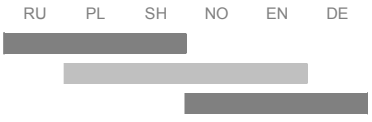


$kan

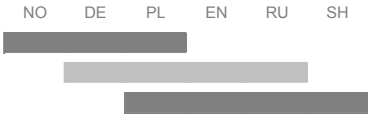
$liker



$man



$mange



$med

$meg

$men

$mennesker

$mer

$min

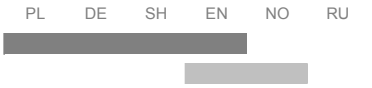
$mye

$når

$norge

$norsk

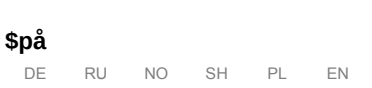
$og



$også

$om

$på



$så

$skal

$som

$sted

$til

$være

$veldig

$venner

$vi

$viktig

$å

DE EN PL RU + Somali

$andre

$at

$av

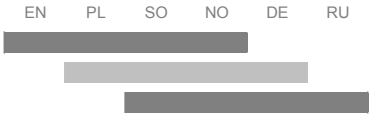
$bare

$barn

$barna

$bo

$da



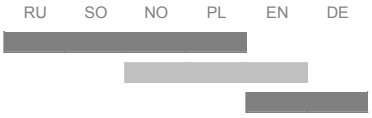
$de

$den

$det

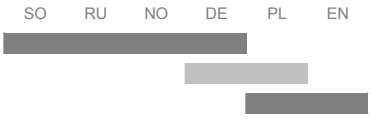
$du

$eller

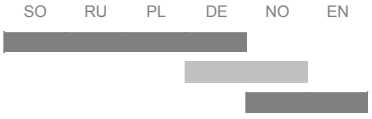


$en

$er



$et



$for

$fordi

$fra

$han

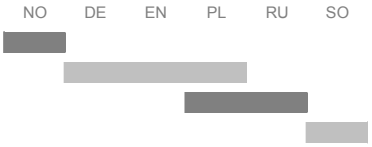
$har

$hun

$i

$ikke

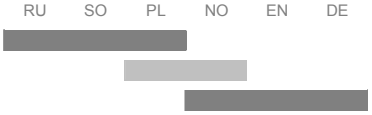
$jeg



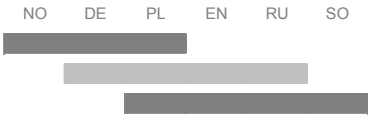
$kan

$liker

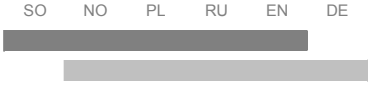
$man



$mange



$med



$meg

$men

EN NO PL RU SO DE

$mennesker

NO DE RU EN SO PL

$mer

RU SO PL EN DE NO

$min

NO DE RU PL EN SO

$mye

SO DE EN NO RU PL

$når

NO SO RU PL EN DE

$norge

NO PL DE EN SO RU

X X X X

$norsk

NO DE EN SO PL RU

$og

PL DE EN SO NO RU

$også

RU NO SO DE EN PL

$om

SO EN NO PL DE RU

$på

SO DE RU NO PL EN

$så

PL NO SO EN DE RU

$skal

RU DE EN NO PL SO

$som

DE RU EN NO PL SO

$sted

NO PL DE RU SO EN

$til

DE RU SO PL EN NO

$være

SO RU DE PL NO EN

$veldig

NO EN SO DE PL RU

$venner

NO PL SO EN DE RU

$vi

EN RU DE SO PL NO

$viktig

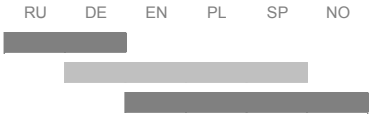
RU SO NO DE PL EN

$å

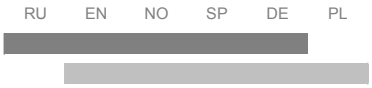
RU DE SO NO PL EN

DE EN PL RU + Spanish

$andre

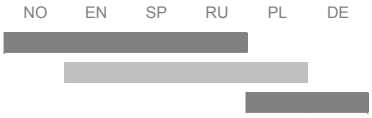


$at



$av

$bare

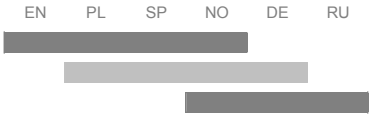


$barn

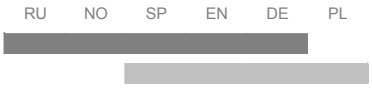
$barna

$bo

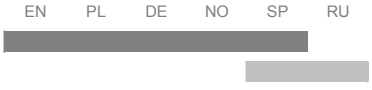
$da



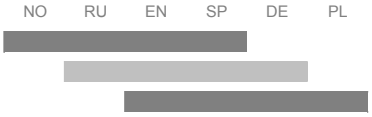
$de



$den

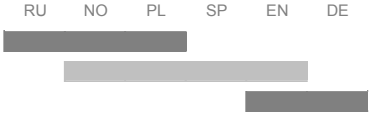


$det

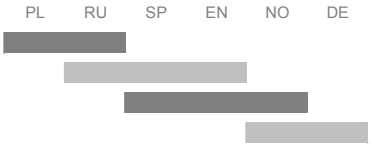


$du

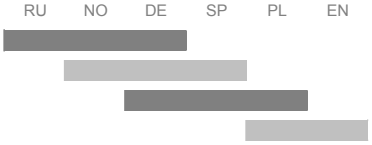
$eller



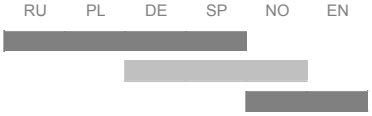
$en



$er



$et



$for

$fordi

$fra

$han

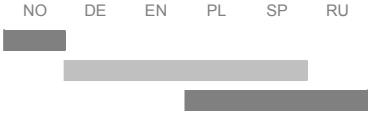
$har

$hun

$i

$ikke

$jeg

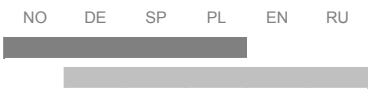


$kan

$liker

$man

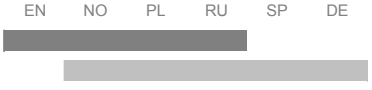
$mange



$med

$meg

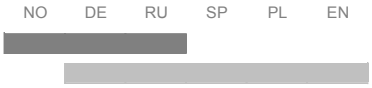
$men



$mennesker

$mer

$min



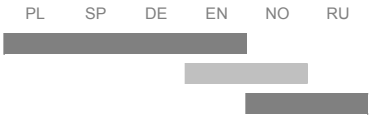
$mye

$når

$norge

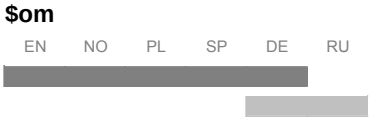
$norsk

$og

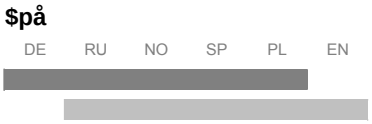


$også

$om



$på



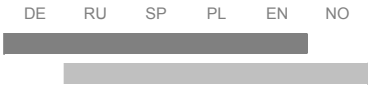
$så

$skal

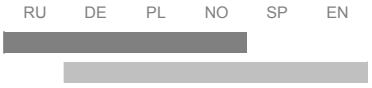
$som

$sted

$til



$være



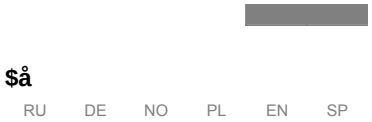
$veldig

$venner

$vi

$viktig

$å



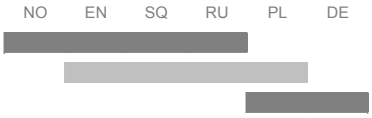
DE EN PL RU + Albanian

$andre

$at

$av

$bare

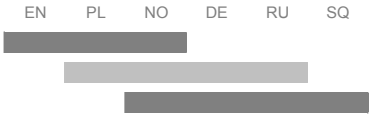


$barn

$barna

$bo

$da



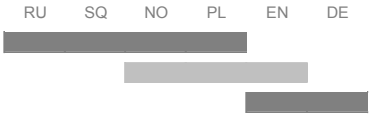
$de

$den

$det

$du

$eller



$en

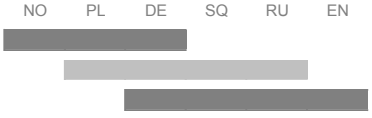
$er

$et

$for

$fordi

$fra

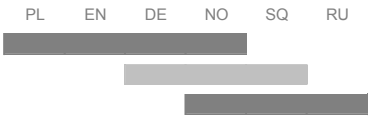


$han

$har

$hun

$i

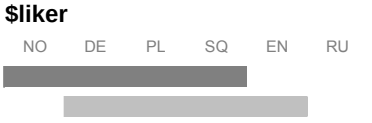


$ikke

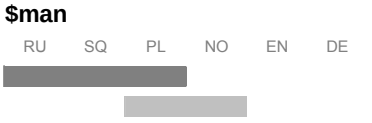
$jeg

$kan

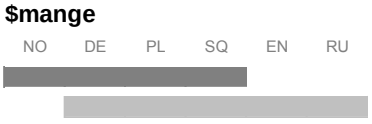
$liker



$man



$mange

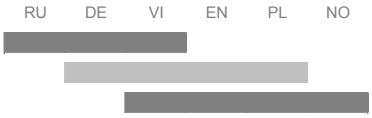


$med


$meg
NO EN RU PL DE SQ
$men
EN NO PL RU DE SQ
$mennesker
NO DE RU EN SQ PL
$mer
RU SQ PL EN DE NO
$min
NO DE RU PL SQ EN
$mye
DE EN SQ NO RU PL
$når
NO RU PL EN SQ DE
$norge
NO PL DE EN SQ RU
$norsk
NO DE EN PL SQ RU
$og
PL SQ DE EN NO RU
$også
RU NO SQ DE EN PL
$om
EN NO PL SQ DE RU
$på
DE SQ RU NO PL EN
$så
PL NO EN DE RU SQ
$skal
RU DE EN NO PL SQ
$som
DE RU EN NO PL SQ
$sted
NO PL DE RU SQ EN
$til
SQ DE RU PL EN NO
$være
SQ RU DE PL NO EN
$veldig
NO EN DE PL SQ RU
$venner
NO PL EN SQ DE RU
$vi
EN RU DE SQ PL NO
$viktig
RU SQ NO DE PL EN
$å
RU DE SQ NO PL EN

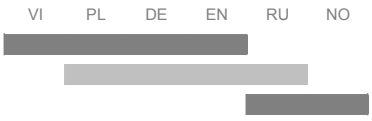
DE EN PL RU + Vietnamese

$andre

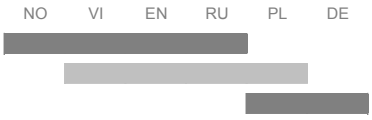


$at

$av



$bare

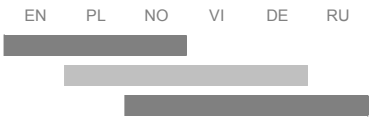


$barn

$barna

$bo

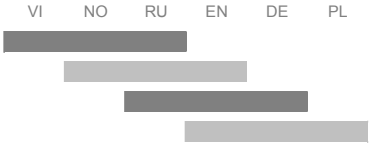
$da



$de

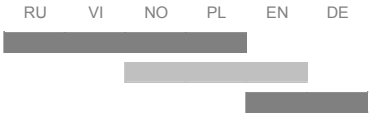
$den

$det

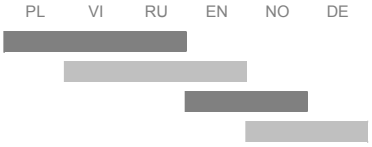


$du

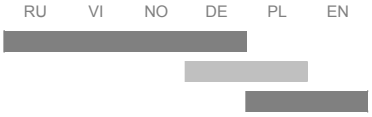
$eller



$en



$er



$et

$for

$fordi

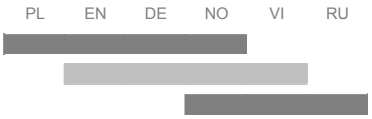
$fra

$han

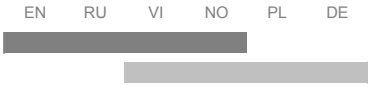
$har

$hun

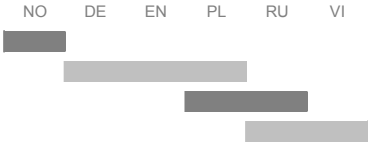
$i



$ikke



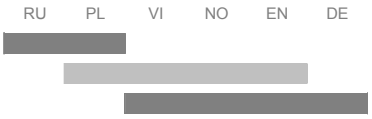
$jeg



$kan

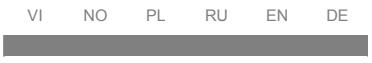
$liker

$man



$mange

$med



$meg

$men

$mennesker

$mer

$min

$mye

$når

$norge

$norsk

$og

$også

$om

$på

$så

$skal

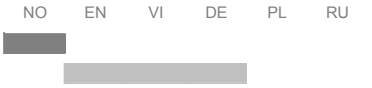
$som

$sted

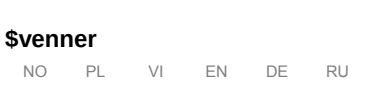
$til

$være

$veldig

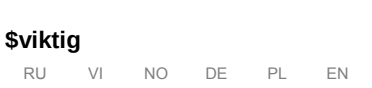


$venner



$vi

$viktig

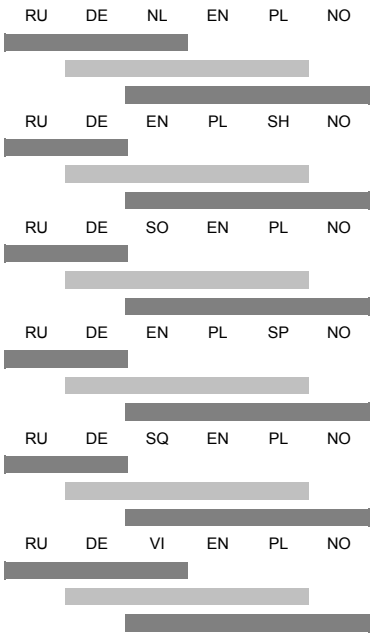


$å

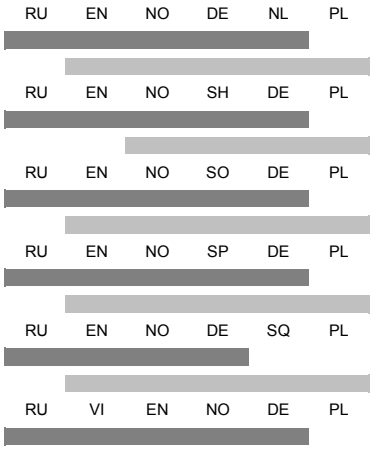
Appendix I. Homogeneity tables (by feature)

This appendix contains the results of the Tukey HSD tests described in §6.3, ordered here by feature.

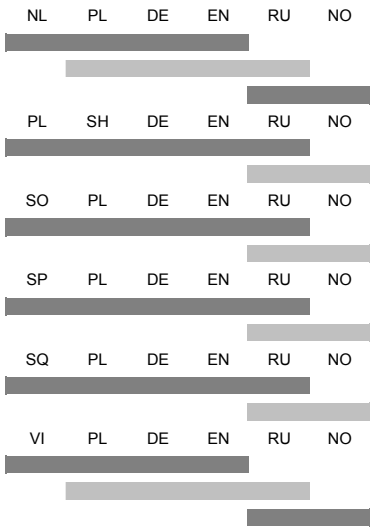
$andre



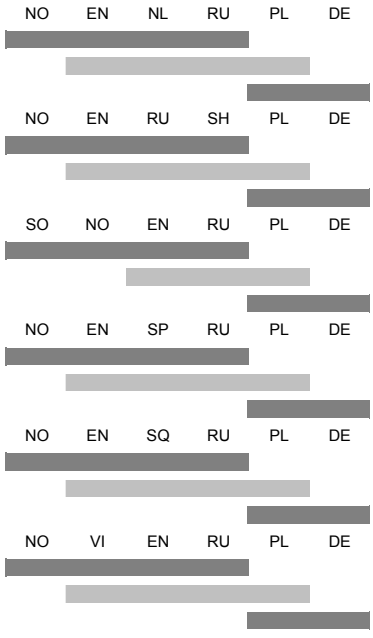
$at



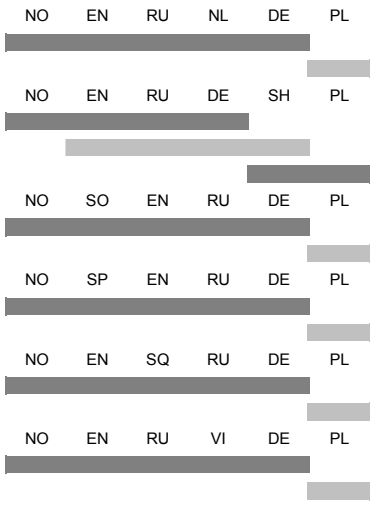
$av



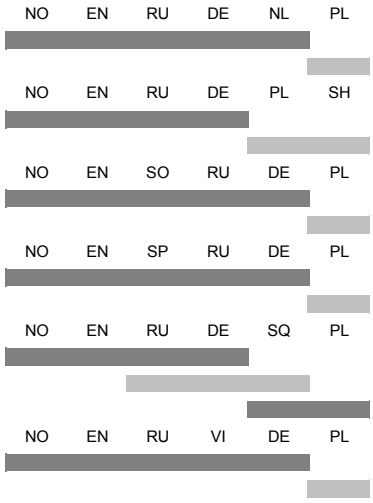
$bare



$barn

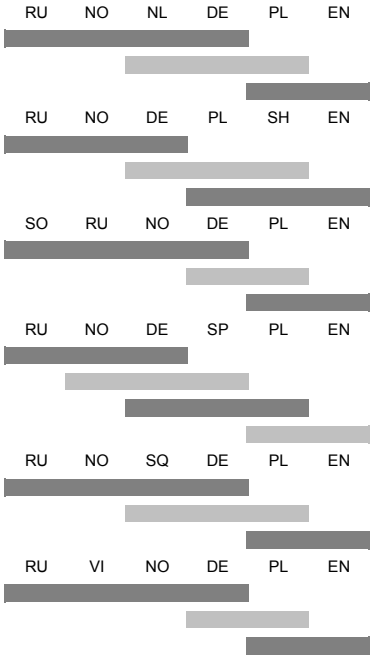


$barna

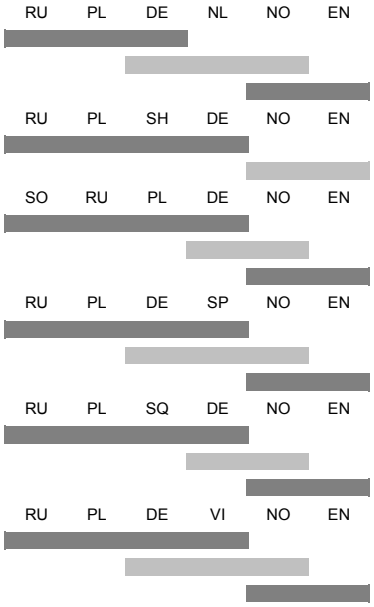



$bo
NO PL RU DE NL EN
NO PL RU DE SH EN
NO PL SO RU DE EN
NO PL RU DE SP EN
NO SQ PL RU DE EN
NO PL RU DE VI EN
$da
EN PL NO NL DE RU
EN PL NO DE RU SH
EN PL SO NO DE RU
EN PL SP NO DE RU
EN PL NO DE RU SQ
EN PL NO VI DE RU
$de
RU NO NL EN DE PL
RU NO EN DE PL SH
RU NO EN SO DE PL
RU NO SP EN DE PL
RU NO EN DE SQ PL
RU NO EN DE VI PL
$den
EN NL PL DE NO RU
SH EN PL DE NO RU
EN SO PL DE NO RU
EN PL DE NO SP RU
EN PL DE NO SQ RU
VI EN PL DE NO RU
$det
NO RU NL EN DE PL
NO RU SH EN DE PL
SO NO RU EN DE PL
NO RU EN SP DE PL
NO RU SQ EN DE PL
VI NO RU EN DE PL
$du
PL NO RU EN DE NL
PL SH NO RU EN DE
PL NO RU EN SO DE
PL NO RU SP EN DE
SQ PL NO RU EN DE
VI PL NO RU EN DE
$eller
RU NO NL PL EN DE
RU SH NO PL EN DE
RU SO NO PL EN DE
RU NO PL SP EN DE
RU SQ NO PL EN DE
RU VI NO PL EN DE
$en
PL RU EN NO NL DE
PL SH RU EN NO DE
PL RU SO EN NO DE
PL RU SP EN NO DE
PL RU SQ EN NO DE
PL VI RU EN NO DE

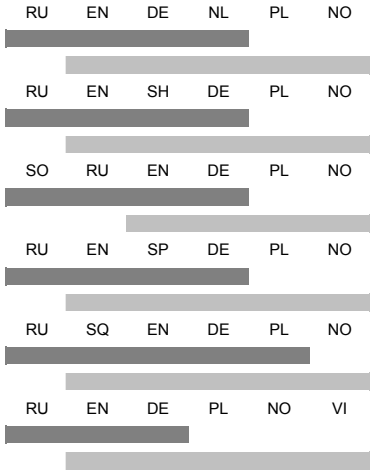
$er



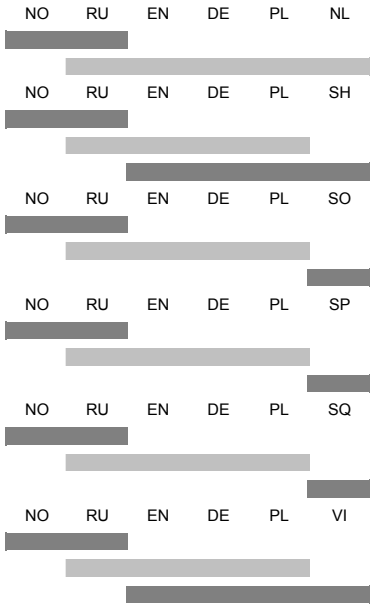
$set



$for

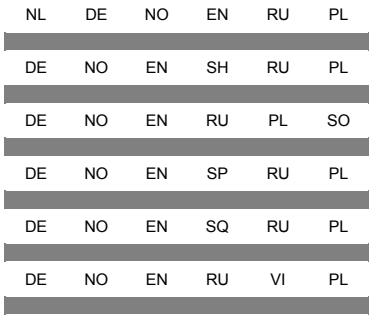


$fordi

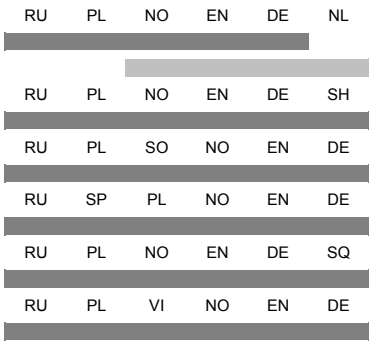


$fra

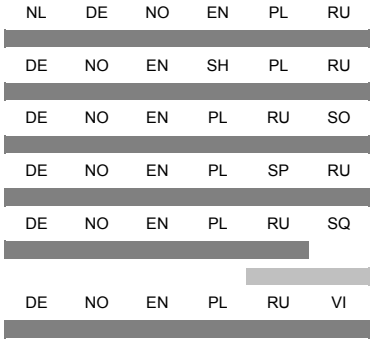
$han



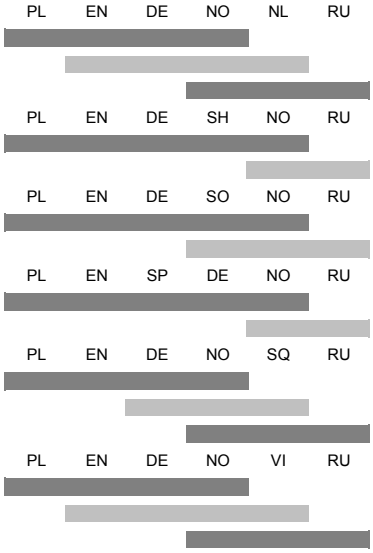
$har



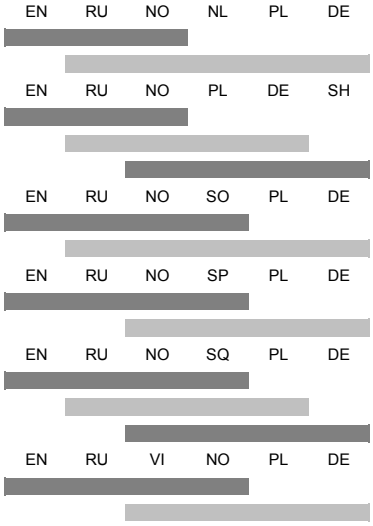
$hun



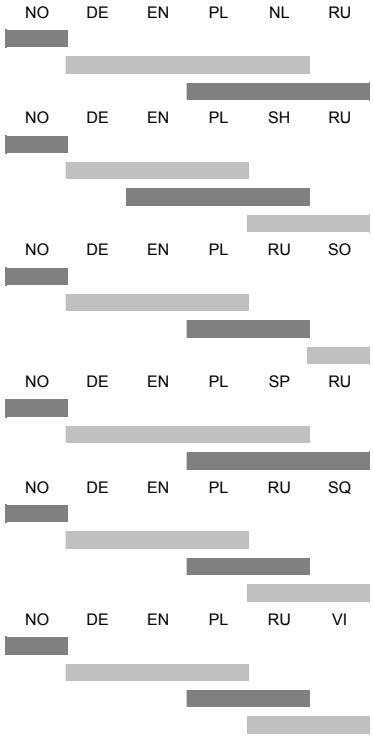
$i



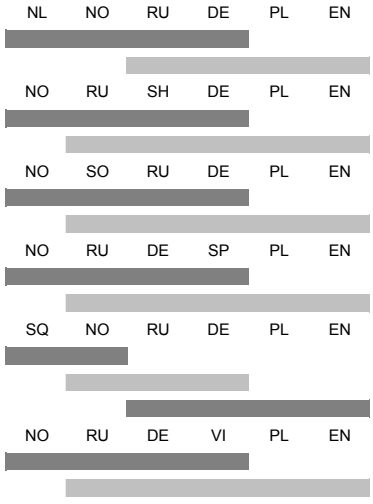
$ikke



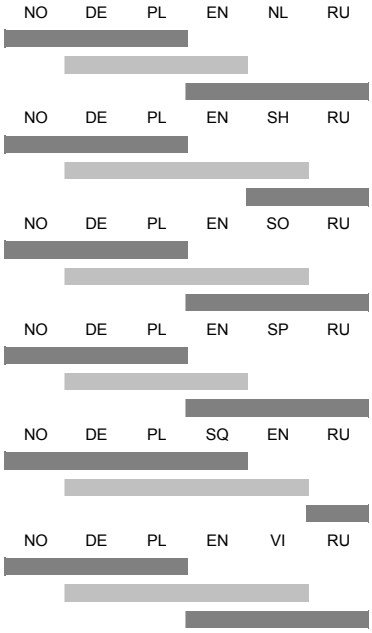
$jeg



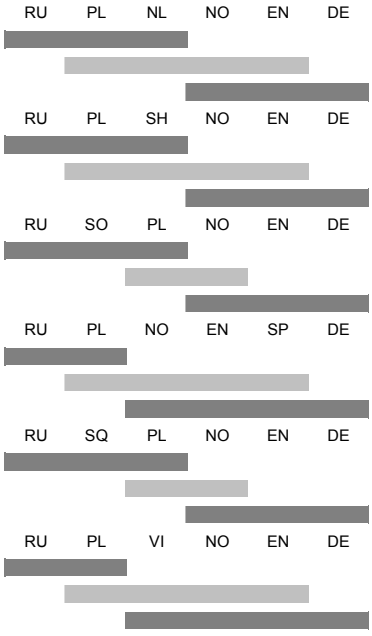
$kan



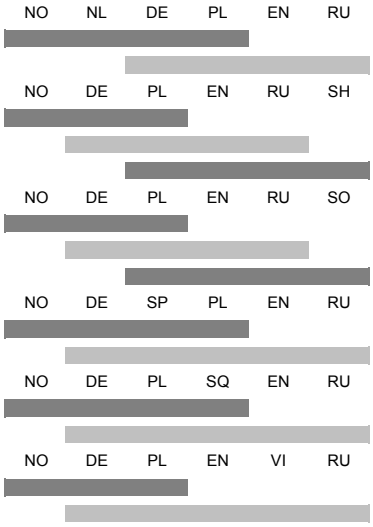
$liker



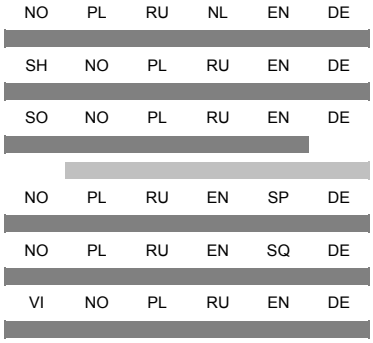
$man



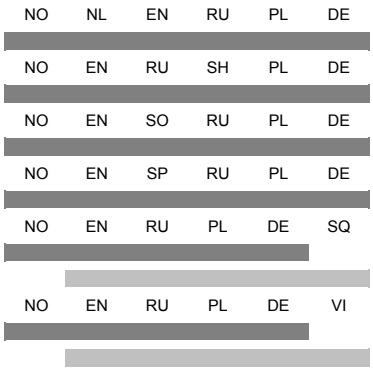
$mange



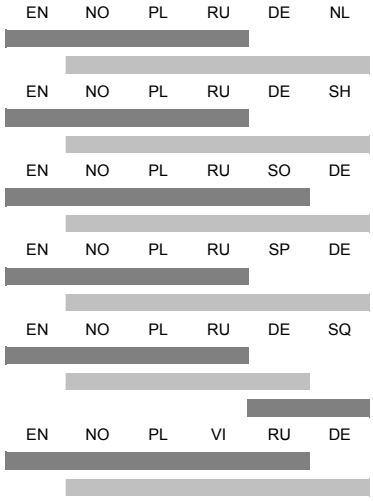
$med



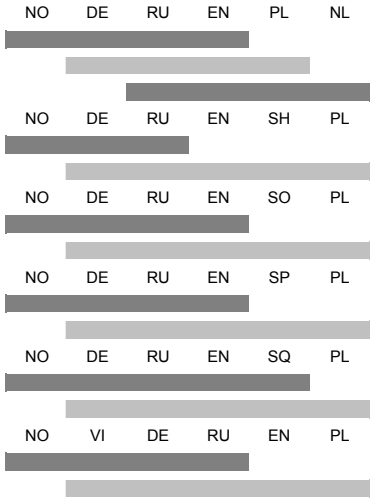
$meg



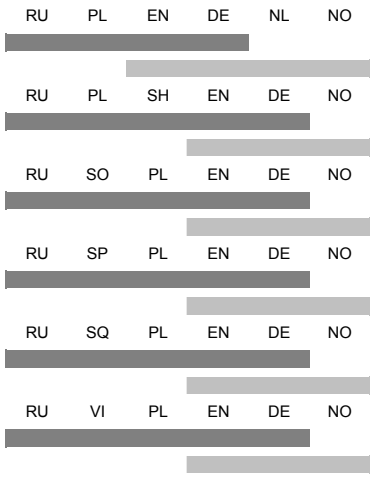
$men



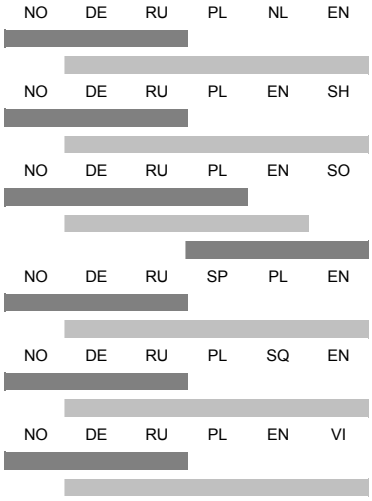
$mennesker



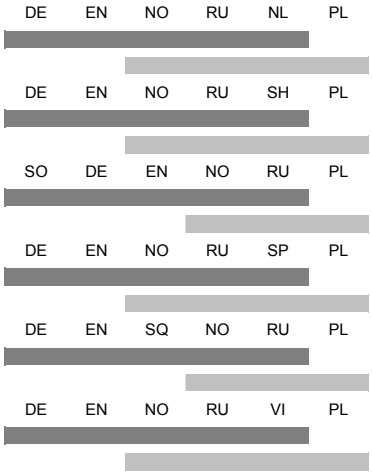
$mer



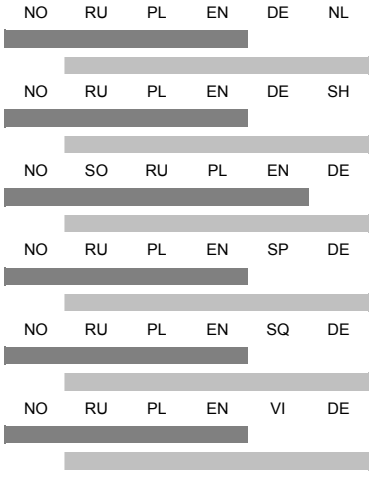
$min

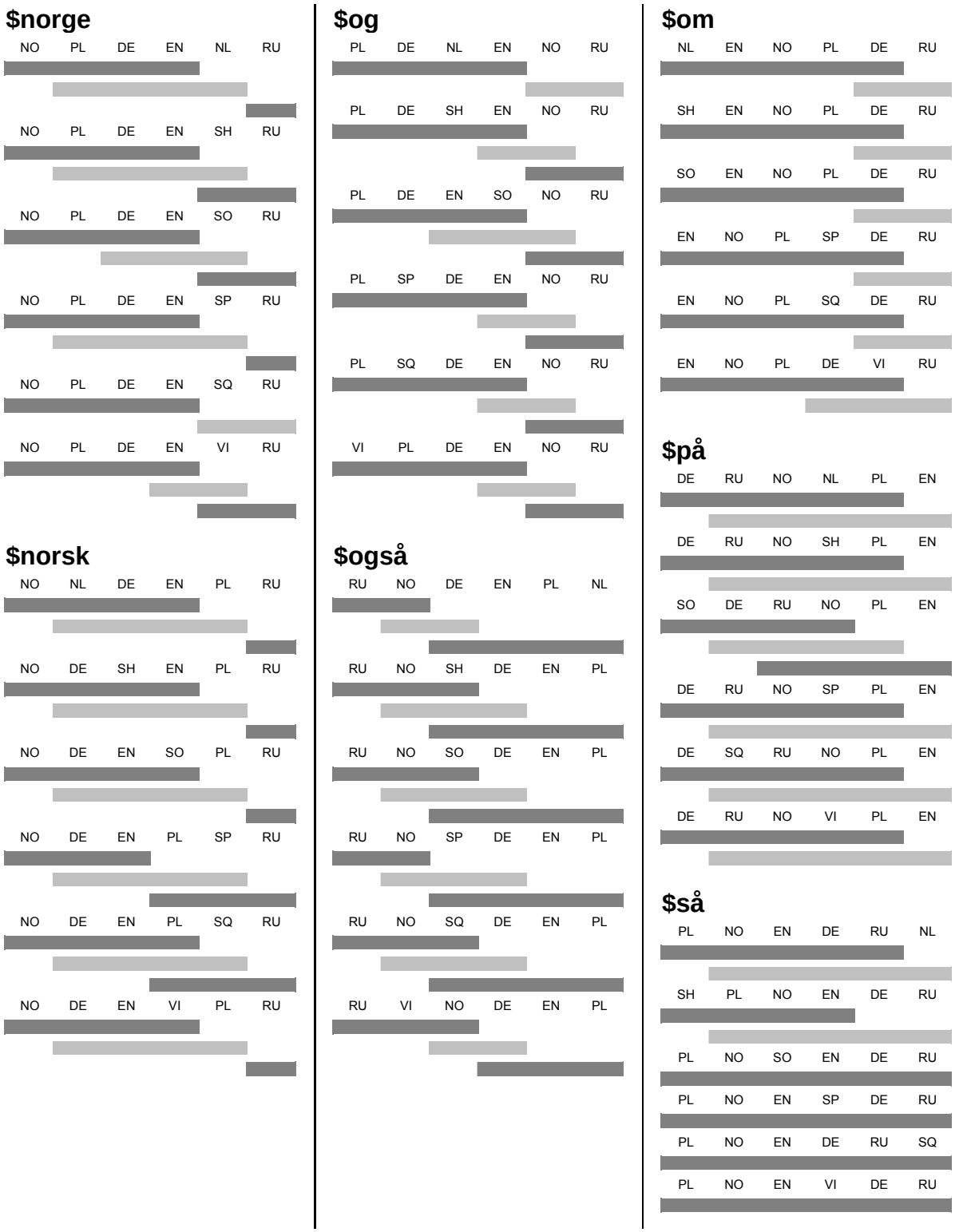


$mye

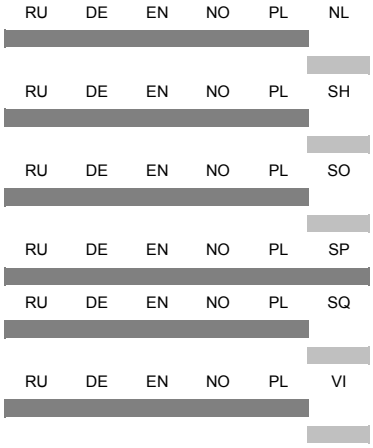


$når

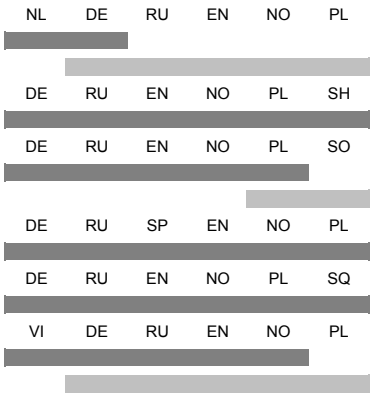



$norge
NO PL DE EN NL RU
NO PL DE EN SH RU
NO PL DE EN SO RU
NO PL DE EN SP RU
NO PL DE EN SQ RU
NO PL DE EN VI RU
$norsk
NO NL DE EN PL RU
NO DE SH EN PL RU
NO DE EN SO PL RU
NO DE EN PL SP RU
NO DE EN PL SQ RU
NO DE EN VI PL RU
$og
PL DE NL EN NO RU
PL DE SH EN NO RU
PL DE EN SO NO RU
PL SP DE EN NO RU
PL SQ DE EN NO RU
VI PL DE EN NO RU
$også
RU NO DE EN PL NL
RU NO SH DE EN PL
RU NO SO DE EN PL
RU NO SP DE EN PL
RU NO SQ DE EN PL
RU VI NO DE EN PL
$om
NL EN NO PL DE RU
SH EN NO PL DE RU
SO EN NO PL DE RU
EN NO PL SP DE RU
EN NO PL SQ DE RU
EN NO PL DE VI RU
$på
DE RU NO NL PL EN
DE RU NO SH PL EN
SO DE RU NO PL EN
DE RU NO SP PL EN
DE SQ RU NO PL EN
DE RU NO VI PL EN
$så
PL NO EN DE RU NL
SH PL NO EN DE RU
PL NO SO EN DE RU
PL NO EN SP DE RU
PL NO EN DE RU SQ
PL NO EN VI DE RU

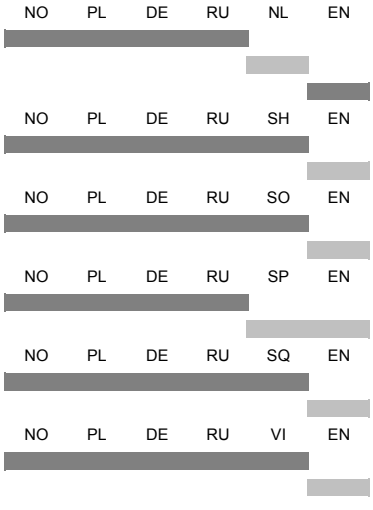
$skal



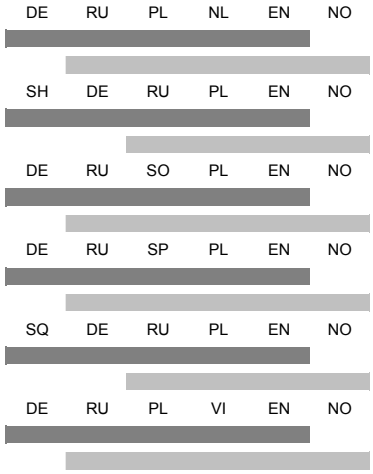
$som



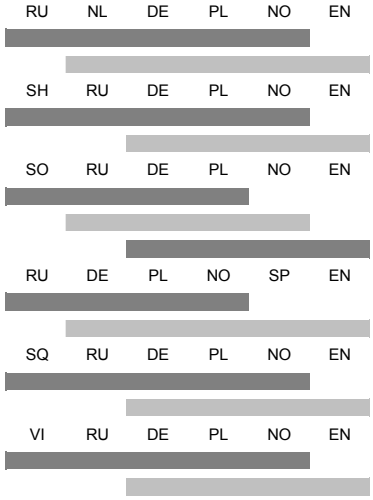
$sted



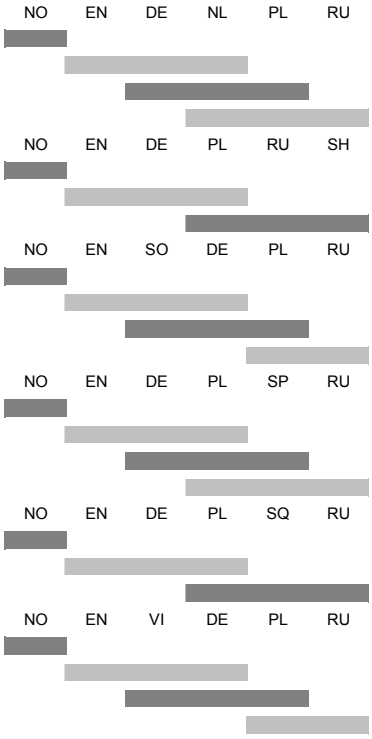
$til



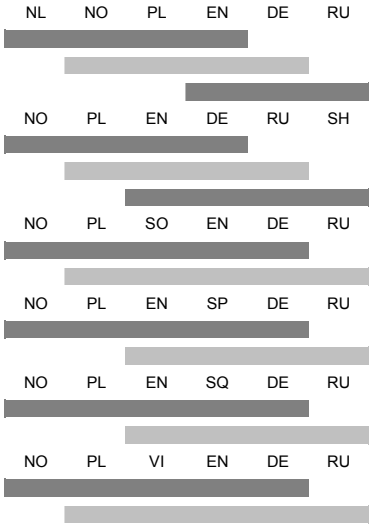
$være



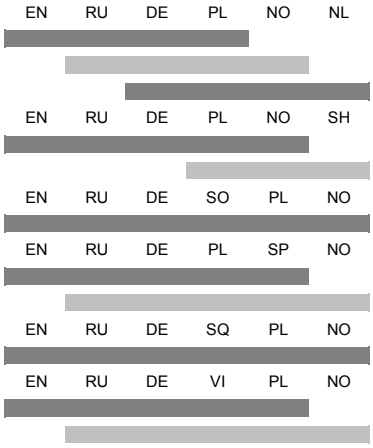
$veldig



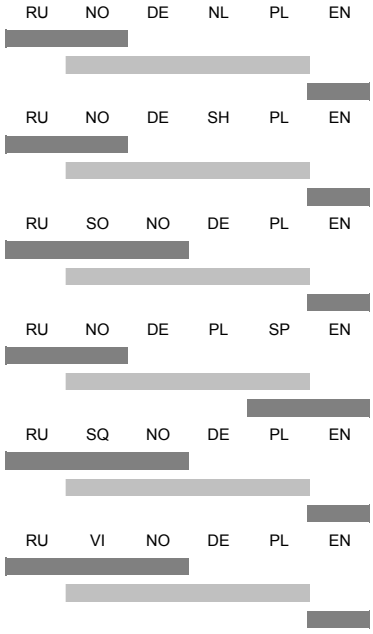
$venner



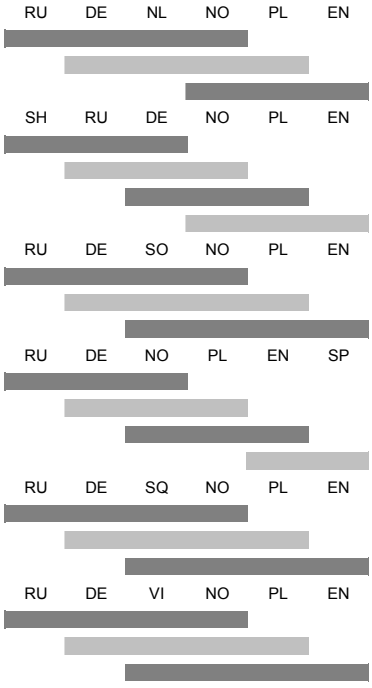
$vi


EN RU DE PL NO NL
EN RU DE PL NO SH
EN RU DE SO PL NO
EN RU DE PL SP NO
EN RU DE SQ PL NO
EN RU DE VI PL NO

$viktig


RU NO DE NL PL EN
RU NO DE SH PL EN
RU SO NO DE PL EN
RU NO DE PL SP EN
RU SQ NO DE PL EN
RU VI NO DE PL EN

$å


RU DE NL NO PL EN
SH RU DE NO PL EN
RU DE SO NO PL EN
RU DE NO PL EN SP
RU DE SQ NO PL EN
RU DE VI NO PL EN

This study investigates cross-linguistic influence in Norwegian interlanguage using predictive data mining technology and with a focus on lexical transfer. The following research questions are addressed:

- Can data mining techniques be used to identify the L1 background of Norwegian language learners on the basis of their use of lexical features of the target language?
- If so, what are the best predictors of L1 background?
- Can those predictors be traced to cross-linguistic influence?

Lexical transfer in Norwegian interlanguage – A detection-based approach

Steve Pepper, University of Oslo

The study is based on L2 data from ASK, the Norwegian Second Language Corpus, and draws on resources from the ASKeladden project. The source data consists of essays written by 1,736 second language learners of Norwegian from ten different L1 backgrounds, and a control corpus of 200 texts written by native speakers. Word frequencies computed from this data are analysed using the multivariate statistical method of Discriminant Analysis, and the results are subjected to Contrastive Analysis.

This combination of quantitative and qualitative analysis yields all three types of evidence called for by Scott Jarvis in his requirements for rigour in transfer research. Well-known effects, such as the tendency for Russian learners to omit the indefinite article, are confirmed, but also other, more subtle patterns of learner language are revealed, such as the marked tendency for Dutch learners to overuse the modal verb *skal*. These results provide abundant material for future research, and the statistical methods can be applied in many areas of quantitative language research.