Nor Jnl Ling 37.2, 169–198 © Nordic Association of Linguists 2014 doi:10.1017/S0332586514000171

Blaxter, Tam T. 2014. Applying keyword analysis to gendered language in the *Íslendingasögur*. *Nordic Journal of Linguistics* 37(2), 169–198.

Applying keyword analysis to gendered language in the *Íslendingasögur*

Tam T. Blaxter

Keyword analysis has been used to investigate properties of style and genre, as a tool in discourse analysis, and as a method of identifying differences between the speech of distinct social groups. It has often been criticised as a blunt tool which can exaggerate what differences are present and fails to distinguish between quite distinct phenomena. However, it remains a very powerful tool for wide analysis of systematic differences between corpora when used with sufficient scepticism. This paper uses keyword analysis to examine differences between the speech of male and female characters in the *Íslendingasögur*, narrative prose texts composed in Iceland in the 13th and 14th centuries. This dataset is of particular interest because such representations of speech are the only window on the language of social groups who were not involved in text production in medieval societies. It aims to demonstrate a rigorous application of keyword analysis, exemplifying what it can and, crucially, what it cannot show.

Keywords corpus linguistics, gender, keywords, Old Icelandic, sociolinguistics

Pembroke College, Cambridge CB2 1RF, UK. ttb26@cam.ac.uk

1. INTRODUCTION AND THEORETICAL BACKGROUND

1.1 Introduction

Gender seems always to play a significant role in conditioning sociolinguistic variation. However, uncritical use of powerful methodological tools can often result in an exaggeration of gendered differences in language use. One method which has often been subject to such use is keyword analysis. This paper will present a detailed study of keywords in the speech of male and female characters in the Old Norse saga *Íslendingarsögur*. In particular, it aims to show that what initially appear to be sociolinguistic differences between the speech of these two groups are in fact a consequence of properties of narratives and the social construction of gender in the society depicted.

1.2 Keyword analysis

A lexical item which occurs significantly more frequently in one (sub)corpus than other is termed a KEYWORD (Baker 2004:346–347; 2010:134). By examining the

frequencies of a large number of lexical items across multiple corpora, it is possible to identify the keywords for each of those corpora: the lexical items which particularly distinguish them from one another. Keyword analysis has often been used to examine differences of genre and style (Culpeper 2009:32–34) or in discourse analysis (Baker 2004:347). However, if the corpora contrasted differ with regard to social features of the speakers, then keywords may indicate differences in patterns of lexical choice exhibited by those speakers. It is only a small step to claim that keywords can be used to identify differences between sociolinguistic varieties.

In principle, keyword analysis can clearly be used for this purpose; for example, if corpora of British and American English were compared, *got* and *forwards* might be key in the former while *gotten* and *forward* might emerge as keywords in the latter. In particular, keyword analysis has frequently been used to identify the differences between speech produced by female and male speakers; examples include Rayson, Leech & Hodges (1997) and Schmid (2003). In a related vein, Baker (2004) has used keyword analysis to examine the construction of gender in fiction texts.

The appeal of this method for examining the differences between sociolinguistic varieties is clear: it is rapid and undemanding (Xiao & McEnery 2005:77), can identify significant variables from a huge pool of initial possibilities and, in concentrating on patterns of lexical choice, highlights exactly those variables which are most likely to throw light on the social construction of the groups in question. However, the use of keyword analysis for this purpose raises a number of practical and theoretical problems and its use has often been criticised as unsophisticated or insensitive to these issues (Culpeper 2009:30). In addition, the investigation of gendered language raises its own specific problems.

1.3 Issues with keyword analysis

1.3.1 Statistical issues

Using keywords to investigate sociolinguistic varieties presents several practical issues. Identifying keywords involves the use of statistical significance tests, but the value of such tests is undermined by repeated use. Accepting the most commonly assumed threshold for significance, $p < .05$, implies a 5% chance of obtaining a false positive result (a type 1 error) in a random sample from normally distributed data. For a single test this is an acceptable rate of error, but if thousands of tests are undertaken, as is necessary to search for keywords among the large range of lexical items present in a typical corpus, it presents a more serious issue. Obtaining false negative results (type 2 error) can also occur, but, unlike type 1 error, these are hard to quantify without knowing in advance what differences were present between the speech of the groups in question.

1.3.2 Issues of interpretation

Putting aside the issue of statistical anomalies, keywords can reflect a wide range of different facts about corpus data and language use. One possibility is that they do indeed reflect linguistic differences between two groups of speakers: a difference in the rates at which sociolinguistic variants are selected by two groups of speakers may result in those variants being key. For example, in large corpora of British and American English, *dived* and *dove* would be likely to be key. Thus examining lists of keywords associated with two corpora can reveal such linguistic differences between the groups who produced them. Such keywords will be termed DIRECT KEYWORDS and in this study would represent linguistic differences between female and male represented speech. The variants they reflect might even be those used by speakers to EXPRESS gender.

However, keywords can also arise indirectly through differences in genres or contexts of language use within the corpora. For example, if data for spoken corpora were gathered in systematically different contexts for younger and older speakers – perhaps in a school context for younger speakers and a workplace context for older speakers – this might result in age keywords reflecting the semantic areas associated with those contexts. However, these INDIRECT KEYWORDS would not represent any linguistic difference between the groups: were data gathered from them in the same contexts and speaking about the same topics, there might be no difference in their language use. In the case of gender, indirect keywords might reflect situations in which a given variant is used in a particular speech context or to express a particular social role occupied disproportionately frequently by male or female speakers. A frequent criticism of work on language and gender has been that such indirect associations between linguistic variables and gender have been misidentified as directly gendered features (Grob, Meyers & Schuh 1997:195; Cameron 2007:48–49, 125–130, 133–139; Baker 2008:33–36).

Another complicating factor is a phenomenon known as poor dispersion. This refers to cases in which a large proportion of instances of a term are produced by a very small subset of speakers or in a very small subset of texts. In such a case, the keyword distribution may be purely down to these speakers or texts (Baker 2004:350; 2008:63; Harrington 2008:97, 101). For example, in speech in the British National Corpus the lexical item *fucking* is a strong keyword for male speakers (Rayson et al. 1997:135–136; Baker 2008:46). However, Baker (2012) has shown that just two male speakers were responsible for 50% of male instances and two female speakers for 50% of female instances and there is little difference between the proportions of male and female speakers who used the word at least once (3.63% of male speakers; 3.23% of female speakers). Thus any features of its overall distribution should be attributed to this tiny group and cannot be considered characteristic of male or female speakers generally. Similarly, Kilgarriff (2005) has demonstrated the

potential problem of poor dispersion by text, showing that even arbitrarily selected subcorpora may be distinguished by many keywords arising from differences between texts.[1]

Further possible interpretations of keywords are raised when working with data from narrative fiction. Features of the narrative may directly interact with gendered linguistic variables, creating different patterns according to the gender of primary or secondary characters or in different sections of a narrative. For example, research into gendered language in Japanese fiction has demonstrated that the speech of the heroine may be distinguished from other female characters by more markedly gendered features (Shibamoto Smith 2004, Shibamoto Smith & Occhi 2009). Although such interactions do not raise the possibility of gendered keywords which do not result from gendered differences in language use, they do create an additional factor which might unpredictably complicate such results. More worryingly, the linguistic expression of properties of the narrative in combination with the disproportionate representation of female or male characters in certain roles might result in gendered keywords. For example, if protagonists were typified by certain linguistic features regardless of gender, and protagonists were disproportionately male, this might result in male keywords that would not represent any difference in lexical choice in the speech of male and female characters.

Uncritical interpretation of keyword differences, without consideration of this range of underlying possible patterns, is likely to result in an exaggeration of the level of difference between the varieties represented by the corpora examined. Indeed, the simple fact that identifying lists of keywords can only reveal differences and not similarities between corpora is likely to lead to such a bias (Baker 2004:349), and this criticism has been levelled at much of the earlier research into linguistic variation associated with gender.

1.4 Proposed solutions

In order to assess the extent to which these issues undermine the use of keywords to identify differences between the speech of female and male characters in the *Íslendingasögur*, it is necessary to identify ways by which keywords resulting from different underlying patterns can be distinguished.

Direct keywords, resulting from differences in lexical choice in the speech of male and female characters, would be variants of Labovian linguistic variables and must thus be able to occur in the same contexts and in the same function as other variants. Comparing the two subcorpora, these variants should then be in complementary distribution: if one variant is disproportionately favoured by one group of speakers, the other speakers must use the other variants proportionately more frequently. In the case of content keywords, such variants might be expected to be

synonymous lexical items. Thus where keywords are in complementary distribution with their synonyms, this can be taken as evidence that they are direct keywords, resulting from differences in lexical choice between represented female and male speech.

Indirect keywords, resulting from association between a social feature or role or a narrative context and gender, are predicted to show different distributional properties. Social roles and particularly narrative contexts can be expected to be characterised by not one but a cluster of semantically related terms. Thus where keywords occur in identifiable semantic groups, this can be taken as an indication that they are indirect keywords. Likewise, they can be expected to MATCH the distribution of their synonyms.

These semantic groups might be confirmed by the matching collocates of their members. In addition, cases might arise where a lexical item had other uses in addition to the use which characterised a particular context and resulted in its being a gendered keyword. In these cases, systematically different collocates could be expected for the keyword in female and in male speech.

Poor dispersion by 'speaker' is highly unlikely to prove a problem in these data because the amount of data associated with each character is relatively small. However, poor dispersion by text might occur: if a term were particularly frequent in one text and that text had an especially skewed ratio of female to male speech relative to the corpus as a whole, this might result in that term being a gendered keyword. This issue can be identified by recalculating the significance of each keyword using a weighted ratio of male to female speech according to the distribution of that keyword into texts and the ratio of male to female speech in those texts.

One way of dealing with the issue of type 1 error is simply to decrease the threshold p value. Oakes & Farrow (2007:90) advise the use of the Bonferroni correction, which aims to obtain a familywise error rate $\leq \alpha\%$ by testing n hypotheses at $p < \alpha \div n$.[2] This may solve the problem, but it is at the expense of much increased type 2 error. Additionally, Bestgen (2014) demonstrates unequivocally that even using this method and correcting for issues of dispersion by text, an unacceptable rate of type 1 errors may still arise in keyword analysis (his test suggested that 16% of Oakes & Farrow's (2007:4) results might in fact be statistical anomalies). Another possible approach is thus to reduce the total number of tests carried out by looking only at a subset of lexis (such as only examining the most frequent lexis).

Finally, some attempt can be made to avoid exaggerating overall levels of difference by comparing the number and significance of keywords identifiable between male and female speech with those which can be identified between other subcorpora.

2. METHODOLOGY

2.1 Data

The data in this study are the direct speech from the *Íslendingasögur*, a series of narrative prose texts in Old Norse produced in Iceland in the thirteenth and fourteenth centuries concerned with the settlement of Iceland and its early history. They have been taken from the *Fornrit* section of the *Mörkuð íslensk málheild* corpus of Modern Icelandic (Helgadóttir et al. 2012–), and accordingly have been respelled into Modern Icelandic orthography (Rögnvaldsson & Helgadóttir 2011:65). For this research the corpus was reencoded into the TEI XML encoding (TEI Consortium 2007) and all speech was annotated with <q> tags and tagged for the gender of the speaker. Plural speakers were tagged separately, as were occasional instances of supernatural beings. A total of 13,561 utterances were tagged as male, 1596 as female and 259 as plural; the remaining 60 were tagged either as indeterminate or as non-human.

Keywords were calculated using lemmatised data tagged for part of speech. Using part-of-speech tagging when undertaking keyword analysis is argued for by Rayson (2004, 2008) on the basis that it can allow the disambiguation of disparate uses of identical forms. The data were not semantically tagged (as argued for in keyword analysis by Rayson (2004, 2008) and Culpeper (2009:54–55)) as the relevant resources for semantic tagging are not available for Old Norse. In order to limit the number of results produced and minimise type 1 error, only lexical items among the 200 most frequent in at least one subcorpus were considered. Significance was tested using a chi-squared test and $p < .05$ was required for a distribution to be considered significant. Bestgen (2014) criticises the use of the chi-squared test in keyword analysis on the basis of the rate of type 1 errors it produces and potential problems with dispersion, as discussed above. However, he concludes that, barring the use of a vastly more processing intensive permutation test, chi-squared tests can be used as an exploratory tool to direct deeper analysis (Bestgen 2014:168–169); this is closely in line with the methodology of this paper. The reason that the set of lexical items examined was limited to the 200 most frequent in male and female speech instead of examining all lexical items and using the Bonferroni correction was that the latter approach gave only 24 results, of which eight were proper nouns and the rest were highly infrequent words. This suggested an unacceptable rate of type 2 error, presumably a result of the small total size of the corpus.

2.2 Overall degrees of difference

Although all of the *Íslendingasögur* are anonymous, it is reasonable to assume that they were written by different authors. Accordingly, the four longest individual texts have been compared to give an approximation of the degree of inter-individual

Subcorpora	Number of lexical items considered	Number of keywords	Mean deviation from expected distribution
Female and male represented speech	232	62 (26.72% of items considered)	2.97%
Njáls and *Egils*	256	182 (71.09%)	17.29%
Njáls and *Laxdæla*	251	178 (70.92%)	15.29%
Njáls and *Grettis*	245	170 (69.39%)	14.18%
Egils and *Laxdæla*	244	156 (63.93%)	15.41%
Egils and *Grettis*	244	154 (63.11%)	14.82%
Laxdæla and *Grettis*	241	163 (67.63%)	15.11%

Table 1. Keywords across different subcorpora.

variation. This then provides a context for the degree of difference between the subcorpora of female and male represented speech.

2.3 Classes of keywords

The keywords identified were grouped and investigated with a view to the properties predicted in Section 1.4. The list of keywords are given in Section 4.10 and these investigations are discussed in Section 4.2. Only a sample of the keywords are discussed here for reasons of space.

3. DEGREES OF DIFFERENCE

3.1 Results

Table 1 shows the number of keywords identified when comparing the 200 most frequent words in various different pairs of subcorpora.

3.2 Discussion

As can be seen from the numbers of in Table 1, much greater differences are found comparing the individual texts than comparing male and female dialogue: a higher proportion of frequent words are key (between 63.11% and 71.09% of lexical items considered when comparing texts, but only 26.72% when comparing male and female dialogue) and the mean deviations from expected distributions are much higher when comparing texts. An instructive comparison can be made with a study by Baker (2012), who finds that there are SIMILAR overall levels of difference between female and male subcorpora and arbitrarily selected corpora. Thus it is expected that

gender differences be comparatively limited on the scale of inter-individual variation generally. The fact that in this case lexical difference is much GREATER between texts than between female and male speech might be attributed to the systematic differences in subject matter which can be expected between texts.

4. KEYWORDS

4.1 Results

Tables 2 and 3 below show the keywords identified when comparing the subcorpora of female and male represented speech.

4.2 Discussion

4.2.1 Semantic groupings

It was predicted above that indirect keywords, resulting from differences in the contexts in which characters were presented, should occur in definable semantic groupings. Several such groups were identified. These are shown in Tables 4 and 5.

Certain other groups were considered but discarded. One was religious keywords, comprised of *biðja* 'say prayers (middle voice only)' and *taka* in uses such as *taka trú* 'convert to Christianity'. However, closer examination of the data demonstrated that only six of 323 instances of *biðja* in speech in the corpus were in the middle voice, and of these just one required a religious interpretation. Accordingly, there was no good evidence for a grouping of religious keywords.

As other evidence is considered, these semantic groupings and the explanations they suggest will be put to the test. In some cases explanations suggested by different avenues of investigation will prove to be stronger than those suggested by the semantic grouping approach.

Examination of verbs in the keyword lists suggests that female verb keywords tend to be relatively patientive, taking experiencer subjects (*þykja* 'think', *njóta* 'enjoy', *þora* 'be courageous', *hljóta* 'suffer; need', *hefna* 'suffer for') or theme subjects (*hljóta* 'undergo, result from', *þykja* 'seem'). By contrast, male verb keywords tend to take agent subjects (*taka* 'take', *biðja* 'request, instruct', *bjóða* 'command, offer', *sækja* 'seek, attack').

Of the male verb keywords which do not fit this generalisation (*þiggja* 'receive', *vilja* 'want, will', *skulu* 'should, shall'), the latter two are most commonly used as auxiliary verbs: in this context, the theta role of their subject is determined by the main verb. Three female verb keywords do not fit the generalisation: *gifta* 'give in marriage', *fara* 'go' and *senda* 'send'. Eight of the 18 instances of *gifta* 'give in marriage' in female speech are in the middle voice: in this form it can mean 'be given

Lexical item	Word class	Gloss	Male rank/uses	Female rank/uses	Deviation from expected distribution	p
herra	noun	lord, sir	105th/294	437th/6	8.95%/26.84	6.8645×10^{-7}
þrír	numeral	three	185th/142	564th/5	7.54%/11.09	.0034
þiggja	verb	receive, receive hospitality, accept, get	176th/152	474th/6	7.15%/11.29	.0040
lag	noun	shape, position; fellowship; market price; regular time, tune; law (pl), law-community; communion	174th/154	395th/7	6.60%/10.62	.0073
lið	noun	people, group, troops	139th/212	292nd/10	6.44%/14.30	.0021
nærri	adverb	nearby	191st/138	397th/7	6.12%/8.87	.0183
sækja	verb	seek, come to, go, attack	127th/232	260th/12	6.03%/14.71	.0026
hvor	pronoun	who, which	199th/131	392nd/7	5.87%/8.11	.0271
konungur[a]	noun	king	69th/512	121st/30	5.41%/29.33	.0001
aftur	adverb, preposition	back	167th/161	288th/10	5.10%/8.72	.0327
bjóða	verb	command, offer, invite, forebode	94th/320	170th/20	5.06%/17.22	.0028
sjálfur	pronoun	self	145th/206	244th/13	5.01%/10.97	.0175
land	noun	land, country	101st/304	153rd/23	3.91%/12.79	.0234
mál	noun	speech, language, tale, proposition; lawsuit, charge, procedure, transactions, case; time, season, a measure; ornaments	54th/791	73rd/60	3.90%/33.15	.0003

Table 2. Male speech keywords.

Lexical item	Word class	Gloss	Male rank/uses	Female rank/uses	Deviation from expected distribution	*p*
mót(*i*)	adverb, preposition	contrary, against, opposite, towards _ in order to meet, in return	125th/236	187th/18	3.86%/9.80	.0488
biðja	verb	beg, request; court, propose	104th/299	144th/24	3.52%/11.35	.0429
annar	pronoun	other, another	46th/1090	50th/104	2.24%/26.69	.0132
taka	verb	take, choose, accept, begin, (many other uses)	53th/807	60th/78	2.13%/18.87	.0419
að	adverb, preposition	to, that	22nd/2100	29th/213	1.74%/40.18	.0072
vilja	verb	will, wish, intend, want	14th/2974	15th/311	1.48%/48.57	.0064
skulu	verb	shall, must, should	16th/2677	19th/280	1.48%/43.67	.0098
ég	pronoun	I	1st/16693	1st/1800	1.21%/224.21	5.3065×10^{-8}
til	adverb, preposition	to, concerning	11th/3697	12th/405	1.07%/44.00	.0268

[a] The phonological variant *kóngur* is used 12 times in the corpus of dialogue, all by male speakers; if these instances are included then its ranking in the corpus of dialogue by male speakers increases to 68th and its ratio deviation increases to 5.41%.

Table 2. Continued.

Lexical item	Word class	Gloss	Male rank/uses	Female rank/uses	Deviation from expected distribution	p
Þormóður	noun	(personal name)	581st/33	176th/20	26.79%/14.20	4.1745×10^{-10}
Gísli	noun	hostage, watchman; (personal name)	455th/43	198th/16	16.17%/9.54	.0001
gifta	verb	give in marriage	420th/50	189th/18	15.52%/10.56	4.1161×10^{-5}
skömm	noun	shame, disgrace	431st/48	194th/17	15.21%/9.89	.0001
vopn	noun	weapon	360th/58	175th/20	14.70%/11.46	3.2184×10^{-5}
Þorgrímur	noun	(personal name)	368th/57	196th/17	12.03%/8.90	.0009
njóta	verb	enjoy; use, benefit from; (impersonal) avail	269th/91	140th/26	11.28%/13.19	.0001
þora	verb	dare, be courageous	322nd/67	182nd/19	11.15%/9.59	.0009
bóndi	noun	farmer, householder, man, husband	193rd/137	99th/38	10.77%/18.84	5.0198×10^{-6}
Þórður	noun	(personal name)	240th/105	129th/28	10.11%/13.44	.0002
verk	noun	business, work, deed	285th/84	161st/22	9.81%/10.40	.0012
hljóta	verb	get by lot; get; suffer, undergo; result from; must, need	310th/73	178th/19	9.71%/8.93	.0029
faðir	noun	father	98th/309	57th/80	9.62%/37.42	1.1967×10^{-9}
hús	noun	building, house; household, family	305th/74	179th/19	9.48%/8.82	.0034
hefna	verb	take revenge; (impersonal) suffer for	197th/132	115th/33	9.05%/14.94	.0002
bæði	conjunction	both	227th/109	130th/27	8.91%/12.11	.0009
sonur	noun	son	95th/317	59th/78	8.80%/34.76	2.0710×10^{-8}
illa	adverb	badly, ill	132nd/223	85th/48	6.77%/18.34	.0004

Table 3. Female speech keywords.

Lexical item	Word class	Gloss	Male rank/uses	Female rank/uses	Deviation from expected distribution	p
senda	verb	send, send for; throw	203rd/127	135th/27	6.59%/10.14	.0088
yfir	adverb, preposition	above, over; through, across	153rd/192	95th/40	6.30%/14.61	.0021
kona	noun	woman, wife	126th/232	86th/48	6.20%/17.35	.0009
niður	adverb	down	226th/111	155th/22	5.60%/7.44	.0387
síðan	adverb	since; since that	154th/187	109th/36	5.20%/11.59	.0129
Þorsteinn	noun	(personal name)	205th/126	149th/24	5.05%/7.58	.0474
illur	adjective	bad, evil; bad, ineffective; difficult	129th/225	91st/42	4.78%/12.77	.0122
lengi	adverb	long, for a long time	135th/220	93rd/41	4.76%/12.43	.0137
hugur	noun	mind; thought; mood, temper; desire, wish; courage	143rd/209	104th/37	4.09%/10.07	.0396
hlutur	noun	lot, share; part; case, thing, deed	114th/262	88th/45	3.71%/11.40	.0371
bróðir	noun	brother	81st/374	70th/64	3.67%/16.06	.0139
ekki	particle	not	45th/1151	32nd/186	2.97%/39.65	.0005
minn	pronoun	my	24th/1945	16th/310	2.80%/63.17	1.9025×10^{-5}
þinn	pronoun	your	29th/1646	24th/245	2.01%/38.01	.0050
fara	verb	go, come; travel, go through; become, be, fare; do, begin; suit	25th/1765	22nd/260	1.89%/38.35	.0062
þykja	verb	seem, be thought; (impersonal) think	32nd/1593	28th/228	1.57%/28.68	.0309
hann	pronoun	he	7th/5031	7th/719	1.56%/89.62	.0001
þú	pronoun	you	3rd/11488	3rd/1624	1.44%/188.79	6.9453×10^{-8}

Table 3. Continued.

Keywords		
bóndi ‘husband’, *kona* ‘wife’, *faðir* ‘father’, *bróðir* ‘brother’, *sonur* ‘son’	terms of address and kinship	family and household
gifta ‘give in marriage’, *hús* ‘household’		
hugur ‘thought, mood desire’	mental experience	
þykja ‘seem, think’, *njóta* ‘enjoy, benefit from’		
þora ‘be courageous’, *hljóta* ‘suffer, undergo; need’, *hefna* ‘suffer for’		verbs which take experiencer subjects
skömm ‘shame, disgrace’, *þora* ‘dare’, *hefna* ‘get revenge’	feuds and (dis)honour	
þú, ‘you’, *þinn* ‘your’, *minn* ‘my’, *hann* ‘she, he, it’	personal pronouns	
Þormóður, *Gísli*, *Þorgrímur*, *Þórður*, *Þorsteinn*	personal names	

Table 4. Proposed semantic groupings of female keywords.

Keywords		
lið ‘group, troops’, *sækja* ‘seek, attack’, *bjóða* ‘command’, *taka* ‘steal, capture, (uses in various legal terms)’	military terms	
mál ‘proposotion, lawsuit, charge, procedure, transactions, case’, *lag* ‘law, law-community, market price’		legal terms
til ‘to’, *nærri* ‘nearby’, *mót*(*i*) ‘to meet, against, opposite’, *aftur* ‘back’	physical location and movement	
vilja ‘will, want’, *skulu* ‘shall, should’	futurity, intention	
biðja ‘request’, *þiggja* ‘accept’	exchange and interchange	
herra ‘lord’, *konungur* ‘king’	terms of address, aristocratic ranks	
taka ‘take’, *biðja* ‘request, instruct’, *bjóða* ‘command, offer’, *sækja* ‘seek, attack’	verbs that take agent subjects	

Table 5. Proposed semantic groupings of male keywords.

in marriage’. *Fara* has a great variety of meanings in different contexts and phrases; uses such as *fara eigi einnsaman* ‘not be alone, be pregnant’, *fara flatt* ‘tumble’, *fara til svefns* ‘fall asleep’ and *fara í vöxt* ‘increase’ demonstrate that the verb is not universally agentive. Thus it can be seen that this generalisation about male and female verb keywords fits the data relatively well.

It seems likely that this tendency results from power imbalances between male and female characters. A greater use of experiencer, theme or patientive verbs suggests

that female speakers tend to represent the subject as a less powerful entity. As all speakers frequently produce predicates with first person subjects (although the first person pronoun *ég* is a male keyword, its skew is less than 2%), this suggests that female speakers tend to present themselves as less powerful.

4.2.1.1 Terms of address, names and pronouns

The male keywords *herra* 'lord' and *konungur* 'king' are both terms of address in addition to their uses as common nouns. The female keyword *þú* 'you' is also a term of address, as are the five personal name keywords; *þinn* 'your' might also be seen in this connection. Together these keywords suggest a difference in the way in which male and female characters address other interlocutors.

The distribution of personal names will be determined by narrative and textual factors to a much greater degree than for other lexical items, so any apparent correlations with external factors must be suspect; thus uses of personal names were investigated further. When the top 100 personal name lemmas are taken together, there is a highly significant skew in distribution towards use by female speakers,[3] confirming that female speakers used personal names more frequently.

Personal name NPs in Old Norse are structurally diverse. Names may:

- have an adjective adjunct or appositional noun nickname
- appear with a juxtaposed kin term noun
- appear with a juxtaposed patronymic
- appear with a juxtaposed title
- appear with a demonstrative, article or personal pronoun

All of these elements are optional and all can co-occur freely:[4] the possibility was investigated that speaker or referent gender might correlate with these elements in addition to correlating with the occurrence of names overall. The grammatical elements were of particular interest as they would be less likely to be sensitive to narrative contextual factors.

Each instance of 26 personal names[5] in direct speech was identified. These were classified according to which elements occurred in the NP: third person pronouns, non-third person pronouns, determiners and demonstratives, titles, patronymics, other kin terms and other non-pronominal elements. A total of 1174 instances were examined in total; of these, 1111 were singular. The details of these are found in Table 6 (only rows with one or more instances have been included).

The breakdown according to the gender of the referent is given, but it is clear that this is likely to be highly structured by narrative factors: a single major character in a particular text could greatly skew the overall figures for male or female referents. It would be extremely difficult to distinguish these effects from any other effect of referent gender.[6]

±title[a]	±kin[b]	±adj	±patr	±det/dem	±3rd pro[c]	±2nd pro[d]	Speaker		Referent	
							Female	Male	Female	Male
–	–	+	–	–	+	–	0	1	0	1
–	–	–	+	–	+	–	0	1	0	1
–	–	–	–	–	+	–	1	6	1	6
+	–	–	–	–	–	+	0	1	0	1
–	–	–	+	–	–	+	0	1	0	1
–	–	–	–	–	–	+	4	47	4	47
+	+	+	–	+	–	–	0	2	0	2
+	+	–	–	–	–	–	0	4	0	4
+	–	+	+	–	–	–	0	1	0	1
+	–	+	–	+	–	–	0	2	0	2
+	–	+	–	–	–	–	0	1	0	1
+	–	–	+	–	–	–	0	6	0	6
+	–	–	–	+	–	–	0	1	0	1
+	–	–	–	–	–	–	5	91	5	91
–	+	+	–	+	–	–	0	1	0	1
–	+	+	–	–	–	–	0	2	1	1

Table 6. Breakdown of combinations of elements accompanying personal names.

±title[a]	±kin[b]	±adj	±patr	±det/dem	±3rd pro[c]	±2nd pro[d]	Speaker		Referent	
							Female	Male	Female	Male
–	+	–	+	–	–	–	1	4	1	4
–	+	–	–	–	–	–	24	78	35	67
–	–	+	+	–	–	–	0	1	0	1
–	–	+	–	+	–	–	0	9	1	8
–	–	+	–	–	–	–	8	38	4	42
–	–	–	+	–	–	–	3	28	7	24
–	–	–	–	+	–	–	1	3	0	4
–	–	–	–	–	–	–	133	603	133	603
					Totals:		180	932	192	920

[a] The titles which occurred were: *bóndi* 'householder', *búmaður* 'householder', *dróttning* 'queen', *goði* 'priest, chief', *húsfreyja* 'housewife', *jarl* 'earl', *konungur* 'king', *lögsögumaður* 'law speaker', and *spákona* 'prophetess'.

[b] The kinship terms which occurred were: *bóndi* 'husband', *bróðir* 'brother', *bróðurdóttir* 'niece', *dóttir* 'daughter', *faðir* 'father', *fóstra* 'foster daughter', *fóstri* 'foster son', *frændi* 'relative', *frændkona* 'female relative', *húsbóndi* 'husband', *kona* 'wife', *mágur* 'kinsman in law', *móðir* 'mother', *móðurfaðir* 'grandfather', *næsturbræðra* 'female second cousin', *sonur* 'son', and *systursonur* 'nephew'.

[c] In modern Scandinavian varieties, there exists a demonstrative identical in form to the third person personal pronoun. It is possible that the third person pronouns adjacent to personal names here are examples of that demonstrative and thus should be treated together with other determiners. However, as it is argued that this demonstrative is a recent innovation in modern varieties (Johannessen 2008:163), this has not been assumed here.

[d] Second person pronouns were treated separately from third person pronouns because, although they clearly cannot co-occur, they appear to pattern differently as regards the elements with which they do co-occur.

Table 6. Continued.

None of the grammatical elements was distributed differently for female and male speakers. Of the lexical elements, male speakers used titles significantly more frequently than female speakers,[7] and female speakers used kinship terms more frequently than male speakers, although this only approaches significance.[8] This is exactly in line with the groups of keywords noted above. In addition, it is interesting to note that statistically significant skews in the same direction are found when the gender of the referent is considered instead of the gender of the speaker.[9,10] Although when taken alone it is impossible to disentangle these figures from effects of narrative factors, it is clear that more systematic, social factors are also relevant: importantly, five of the nine titles considered are entirely restricted to male reference.

Taking these observations together, it could be suggested that these two systems of categorisation are in a partial complementary distribution for male and female individuals. When talking of men, speakers have access to both titles and kinship terms in order to distinguish between like-named individuals; by contrast, if speakers wish to attain greater specificity when talking about women, they have less access to titles and use kinship terms correspondingly more frequently. This is then reflected in the usage by male and female speakers: female speakers make less use of the system of classification to which they have less access themselves (titles), making correspondingly greater use of the system to which they have full access (kinship terms). This latter fact could alternatively be seen as a consequence of the fact that female speakers are more likely to talk about female referents and less likely to talk about male referents than male speakers.

This account relies on the proposition that kinship terms and titles serve the same communicative purpose. This is clearly possible because both can be used to disambiguate like-named individuals. Furthermore, it is conceivable that both are used to communicate respect and acknowledge a referent's social prestige or power: this area would benefit from further research to identify in detail what pragmatic effect might be intended by these terms.

4.2.1.1.1 Pronouns

It appears from the keyword lists (Tables 2 and 3 above) that female speakers use personal and possessive pronouns more frequently than male speakers; this is contradicted only by the male keyword *ég* 'I'. One possible account for these observations, which would also be an account for the distribution of personal names, is that male speech in the corpus is more self-centric while female speech is more other-centric: that is, that first person arguments feature as core participants more frequently in male utterances whereas in female utterances second and third person arguments are more frequent and where reference is made to the first person it is as a nominal adjunct (*minn* 'my').

A variation on this explanation is that principal characters are more likely to be the topics of stretches of speech and that male characters are more likely to be principal characters, resulting in an overall impression that male characters are more likely to be talking about themselves. This could be tested by tagging speech for the identity of individual characters and classifying characters by the total amount of speech they produced. These keywords could then be examined in the subcorpora of male and female principal and subsidiary characters separately.

4.2.1.1.2 Cross-linguistic comparisons

Cross-linguistic comparisons can be made regarding the distributions of both personal names and pronouns. In a corpus study of Modern English, Rayson et al. (1997:135–137) found that female speakers used personal pronouns and personal names more frequently than male speakers. The authors suggest that this may indicate that female speakers are 'more concerned with persons as individuals' than male speakers (page 137). This seems a somewhat crude explanation and it should be noted that without tagging corpus data for details of speech act context, as neither the present study nor that of Rayson et al. has done, it is difficult to exclude the possibility that the difference is due to the situations in which speakers are recorded or portrayed. If this explanation is accepted, the male keyword *ég* 'I' sits alone as an anomaly that must be explained differently.

In a another study, of gendered usage in recorded small-group conversations in English, Lynette Hirschman found that female speakers used more of all personal pronouns, including the first person pronoun, and suggested that '[t]his correlated with the subjective impression ... that the females tend to talk more about their own experiences and feelings, while the males tend to generalize and talk rather abstractly' (Hirschman 1994:434).

4.2.2 Collocates

It was predicted that keywords which resulted from differences in the contexts in which characters were presented would show collocates in related semantic areas and might have clearly distinct collocates for male and female speakers. Accordingly, collocates were examined for a series of different keywords in order to further investigate the best explanation for their distributions; this was especially useful in determining the best explanation for highly polysemous words. Such examinations might help to confirm the accounts suggested by the semantic groupings, disambiguate between such accounts, or provide alternative explanations.

In each case, the Antconc collocation analysis tool (Anthony 2012) was used to examine a keyword's collocates, usually examining lemmas only. Antconc outputs collocates alongside a probability value assessing the significance of the frequency of co-occurrence of two lexical items. This measure, the Mutual Information measure

(MI),[11] takes into account the fact that frequent co-occurrence with an otherwise frequent word is a less significant finding than co-occurrence with an infrequent word.

4.2.2.1 Taka *'take'*

The collocates of the male keyword *taka* 'take' within a distance of three tokens in both directions were considered. This distance was selected because many of the different meanings of *taka* are distinguished by co-occurrence with particles, prepositions and nominals which usually occur within a range of three words (Cleasby, Vigfússon & Dasent 1894:622–624).

Firstly, only collocates with a co-occurrence frequency greater than 20 and which were among the top ten according to female and/or male MI score were considered, see Table 7. As there was relatively little difference between the collocates for male and female speakers, it was not possible to account for the skewed distribution by these usages. In addition, these do not help to determine whether *taka* should be classified with one of the semantic groups above to explain its frequency: *taka af* 'put to death', *taka upp mál* 'take up a legal case' and *eftir taka* 'receive a reward' might be classified with legal terminology, but *taka af* could be associated with military terms.

However, when lower frequency collocates of *taka* were considered as well, a different pattern emerged. Among the ten collocates with the highest MI measures for male speakers were seven legal terms and two religious terms, see Table 8. None of these terms were collocates for female speakers. Although all relatively low frequency terms with low frequencies of co-occurrence with *taka*, taken together the frequency of co-occurrence of *taka* with these legal terms alone was 31, easily sufficient to account for its distribution (which deviated from expected frequencies by 18.87).

Thus this keyword fits the predictions made for contextual/narrative keywords: its collocates demonstrate that it is frequently used by one group of speakers in a semantic context not typical of the other group.

4.2.2.2 Biðja *'request'*

Biðja 'request' presented a slightly different problem as it was used too infrequently by female speakers for it to be possible to identify collocates for female speakers alone. When collocates in the combined corpus of female AND male speech with a collocation frequency of greater than 20 were considered, only one content word could be identified in the top ten for MI: *dóttir* 'daughter'; this was the most significant collocate by far (MI = 6.8368) with a collocation frequency of 22. When the threshold for consideration was lowered to ten occurrences, a different pattern emerged, seen in Table 9. Note the female-specified terms *kona* 'woman', *dóttir* 'daughter' and *hún*

Lexical item	Gloss	Frequency of co-occurrence	Female MI	Male MI	Set phrases
af	off	92	4.2567	4.2829	taka *af* 'put to death'; *taka e-n af e-m* 'deprive of'; *taka mikinn/lítinn af e-u* 'make a lot/little of something'[a]
ef	if, whether, lest	46	3.6001	2.8553	
eftir[b]	after, along	24	n/a	3.6349	*eftir taka* 'receive a reward'; *taka eftir* 'notice, observe'
fé	money	40	5.1886	4.7425	
fyrir	before, in front of, against, around	40	3.3562	2.7148	*taka fyrir e-t* 'stop, interrupt, refuse'
mál[c]	speech, matter	38	n/a	3.9299	*taka upp mál* 'take up a legal case'; *taka til máls* 'begin to speak'
nú	now	98	2.922	3.1727	
ráð	advice	60	4.8516	4.7806	*taka ráð* 'begin to speak'; *taka e-t* (*til*) *ráðs* 'resort to'; *takast ráð* 'begin a marriage'
skulu	shall, must, need	80	3.8922	3.0908	
til	to	139	3.4973	3.4958	*taka e-t til ráðs* 'resort to'; *taka til* 'begin', *taka til máls/orðs/orða* 'begin to speak'
upp	up	72	5.0557	5.7246	*taka upp* 'pick up; assume'
vel	well	22	3.6228	2.5263	*taka vel við* 'receive well'
við	against, with	192	4.4575	4.7031	*taka við* 'receive (in return), take up in turn'

[a] Note that *mikinn/mikið* were low frequency but significant collocates of *taka* for both male and female speakers.
[b] Male speakers only.
[c] Male speakers only.

Table 7. Collocates of *taka* 'take' with frequency > 20.

Lexical item	Gloss	Frequency of co-occurrence	Male MI
Legal terms			
sektarfé	the property of an outlaw to be legally confiscated	10	8.38262
fjárheimtur	sheep returning from mountain pastures	3	6.64565
fébót	offer of money, bribe	3	6.64565
goðorð	office of goði	4	5.73876
arfur[a]	bull	5	6.57526
gjald	payment, compensation	3	5.96758
bót	cure; compensation, redress	3	4.88012
Religious terms			
trú[b]	faith	12	6.76113
goðorð	office of goði	4	5.73876

[a] *Taka arf* meant 'receive a legal inheritance'.
[b] *Taka trú* meant 'convert to Christianity'.

Table 8. Low frequency male collocates of *taka* 'take'.

Lexical item	Gloss	Frequency of co-occurrence	MI
dóttir	daughter	22	6.83676
kona	woman, wife	10	5.02940
út	out, from abroad	10	4.92987
vilja	will, want	93	4.69424
hún	she	13	4.20404
þú	you	205	3.83764
hann	he	69	3.45594
ég	I	207	3.35555
að	to, that	119	3.11067
þinn	your	17	3.03929

Table 9. Collocates of *biðja* 'request', frequency ≥ 10.

'she'. If no threshold for frequency were applied, this pattern, seen in Table 10, is clearer still.

Note, in Table 10, the occurrence of three female personal names. Examining the actual instances in question showed that in nearly all of these cases of *biðja* + *dóttir*/*kona*/*hún*/female personal name, the female referent was the object of *biðja*. This seems to suggest that *biðja* may be a word particularly selected by male speakers when describing a situation of making a request of a female referent; this provides

Lexical item	Gloss	Frequency of co-occurrence	MI
hvers	of whom, of what, of each	7	9.32219
Hallgerðar	(personal name)	3	8.25180
griði	horseman, servant	3	8.25180
Kolfinna	(personal name)	3	7.72128
ásjá	help, protection	6	7.42172
ljá	lend, allow	3	6.96229
dóttir	daughter	22	6.83676
Helga	(personal name)	6	6.61437
liðveisla	granting of help, support	3	5.86713
hitta	meet, find, hit, visit	4	5.19290

Table 10. Low frequency collocates of *biðja* 'request'.

a different possible explanation for the distribution of this word than the association with *þiggja* 'accept' or the classification as a more agentive verb, both noted above.

4.2.2.3 Að *'to'*

The word *að* has three functions: (i) it is a preposition meaning 'to'; (ii) it is a marker of the infinitive; and (iii) it is a complementiser. In the MÍM, the first of these functions is distinguished from the other two by part-of-speech tag, and it is this prepositional usage which is a male keyword; however, the part-of-speech tagging in the MÍM does not achieve 100% accuracy and has not been manually corrected. It seemed striking that a function word such as *að* should be a gendered keyword, raising the possibility that this was a direct keyword perhaps associated with some sort of ongoing syntactic change.

It should be possible to distinguish the uses of most instances of *að* on the basis of the following token: prepositional *að* is followed by a dative or accusative noun, determiner or pronoun, infinitive marker *að* is followed by an infinitive, all other following forms imply complementiser *að*. Thus the distribution of word classes for the tokens following *að* were examined. Looking at all instances of *að*, including those tagged as a conjunction (the tag which covered its complementiser and infinitive marker uses), there were no significant differences in the distribution of different word classes in following tokens, although a non-significantly greater proportion of instances in male speech were followed by a dative noun, pronoun or determiner. This implied greater use of prepositional *að*, confirming the keyword identified using the part-of-speech-tagged data.

Looking then at SIGNIFICANT collocates of *að* one token to the right with frequency greater than 10 in male speech, several items were identified, shown in Table 11. Among these collocates are several legal terms (*lögbergi* 'law-rock', *lögum* 'law', *dómi* 'court, judgement', *málum* 'case, matter, speech') as well as military terms (*bana* 'death, doom', *liði* 'host, people, forces'), none of which occur with *að* in female speech in the corpus. The frequency of these collocations is sufficient to account for the appearance of *að* as a male keyword. Thus *að* should be considered an indirect keyword grouped with other legal terms or military terms.

4.2.3 Synonyms

It was predicted that while the synonyms of indirect keywords resulting from contextual biases would show matching distributions, the synonyms of direct keywords representing linguistic differences would show complementary distributions. Investigating the distribution of synonyms thus provided a further method for testing the predictions of the semantic groupings of keywords and for the interpretation of problematic keywords.

4.2.3.1 Intention and desire keywords

It was noted above that the female keywords *hugur* 'mind' and *þykja* 'seem' could be associated with *njóta* 'suffer, experience' in the semantic area of mental experience. However *hugur* and *þykja* can also be placed in a more specific grouping. The primary uses of both of these lexical items are in indirect expressions of opinion and thought: *þykja mér* 'it seems to me, I think'; *segir mér hugur um* 'my mind says to me, I forebode, I suspect'; *er í hug mér* 'it is in my mind, I think'; *kemur mér í hug* 'it comes into my mind, it occurs to me'; *ég leggur á hug* 'I am interested in'; etc. These might be compared with more direct expressions such as *ég hygg* 'I think'. Indirect expressions of opinion such as these are cross-linguistically typical of negative politeness[12] and therefore likely to be distributed differently according to speech act context and social status.

It is then interesting to compare *þykja* and *hugur* to the male keywords *vilja* 'want, will' and *skulu* 'shall, should'. Here, perhaps, is an instance of complementary distribution of near-synonyms. Whilst *vilja* and *skulu* are clearly not straightforward synonyms of *hugur* and *þykja* or the predicates formed with them, there are contexts in which they feature in communicatively equivalent constructions, e.g. 'I want X' or 'I shall do Y' vs. 'it seems to me that X is good' or 'it occurs to me to do Y'. Consider the following two examples, the first from a female character constructed with *þykir mér* 'it seems to me, I think' and the second from a male character with *ég vil* 'I will, I want':

Form	Gloss	Frequency of co-occurrence	Male MI	Set phrases
lögbergi	'law-rock' (dat.sg)	11	7.42670	
lögum	'law' (dat.pl)	27	7.25503	*að lögum* 'legal' (compare the oath formula: *sem ég veit sannast og réttast ok helzt að lögum*)
dómi	'court, judgement' (dat.sg)	10	7.05474	*ganga að dómi* 'go to court'
bana	'death, doom' (obl.sg)	19	6.96281	*verða að bana* 'kill', *kominn að bana* 'decline towards death'
þér	'you' (dat.sg)	303	5.60268	
vísu	'certain' (dat.sg.nt)	37	5.57769	*að vísu* 'certainly'
engu	'none' (dat.sg.nt)	15	5.56931	*að engu* 'for naught'
einu	'one' (dat.sg.nt)	16	5.56288	*að einu* 'only, but'
sinni	'her, his, its' (dat.sg.nt)	54	5.43676	
þessu	'this' (dat.sg.nt.)	54	5.16757	
málum	'case, matter, speech' (dat.pl)	12	5.15588	
bænum	'request, prayer' (dat.pl)	11	4.82767	
liði	'host, people, forces' (dat.sg)	13	4.78939	

Table 11. Right-adjacent collocates of *að* with frequency ≥ 10.

(1) **Gott þykir mér að fara** til vistar með þér en vita skaltu það að ég nenni lítt að gefa fyrir mig því að ég er vel verkfær.
'I would like to go to live with you, but you should know that I would feel little like paying for myself because I am well able to work.'
(Eybryggja saga, *c*. 50)

(2) **Ég vil fara** herra ef þér viljið.
'I would like to go, sir, if you want.'
(Þorsteins þáttur uxafóts)

Of course, (2) could equally be translated as expressing future tense: 'I will go' instead of 'I would like to go'. Nevertheless, it can be seen that the bolded phrases in these two examples are communicatively equivalent, expressing an intention to go somewhere followed by a caveat to do with a listener's opinion. Thus *þykja* and *hugur* on the one hand compared with *vilja* and *skulu* on the other appear to represent a difference in the communicative strategies of male and female speakers in the corpus: in similar contexts, male speakers tend to choose the more direct expressions with *vilja* and *skulu* whereas female speakers tend to choose the more indirect expressions constructed with *þykja* and *hugur*.

Politeness features of this type are associated with power imbalances. Given this, it would be overly simplistic to assume that this feature was used by speakers to express gender (Grob et al. 1997:283, 285–287): especially regarding such socially marked linguistic features, research has often found that association with gender is indirect (Grob et al. 1997:195; Cameron 2007:48–49, 125–130, 133–139; Baker 2008:33–36). Here, the association between gender and this politeness feature may in fact be an indirect reflection of a tendency for female characters to be represented in less powerful positions. This also appears to support the observations made above concerning classes of verbs; again, it might be productive to research this further by examining the subjects of these predicates in female and male speech.

4.2.3.2 Feuds and (dis)honour

Feuds, honour and dishonour form the primary subject matter of many of the *Íslendingasögur* and it is primarily male characters who undertake the revenge killings involved; accordingly, the implication of the female keywords *skömm* 'shame, disgrace', *þora* 'dare' and *hefna* 'get revenge' that female characters discuss these topics more frequently than male characters seems surprising and it might be suspected that they instead point towards differences in lexical choice.

Eighteen straightforward synonyms of *skömm* 'shame' were identified in direct speech in the corpus. In contrast to the case of the intention and desire keywords discussed above, it was found that these patterned as a group, being slightly

skewed towards use by female speakers.[13] Only three synonyms of *þora* 'dare' were identified and these were not used significantly more frequently by male or female characters;[14] no direct synonyms of *hefna* could be identified. Nouns meaning 'daring, bravery' and 'vengeance' were also identified, but exhibited no significant patterns.[15]

This reinforced the conclusion that the distribution of this group of keywords should be explained in terms of contextual or narrative patterns and not as a difference in the communicative strategies or linguistic variety used by female and male speakers.

5. SUMMARY AND CONCLUSIONS

Several methods for exploring the implications of gendered keywords have been exemplified. Identifying possible semantic groupings demonstrated that many keywords did occur in groups, suggesting that they were indirect keywords, resulting from contextual rather than linguistic biases. However, it was clear that these groupings could not be taken at face value; one, that of religious keywords, was rejected when uses were examined in detail.

The examination of the collocates of two keywords, *að* 'to, that' and *taka* 'take', further demonstrated that more investigation was required to confirm and specify keyword groupings. This examination also revealed that even in large datasets skewed distributions may be due to a handful of low frequency constructions. The examination of the collocates of another keyword, *biðja* 'request', revealed features of the usage of this word that would not have been visible merely from its distribution into male and female speech. Similarly, further investigations into personal names, pronouns and terms of address demonstrated that even where biases may be linguistic and not contextual, their exact significance may not be obvious from sets of keywords and different investigatory approaches are needed to illuminate them.

Other further investigations confirmed the original semantic groupings. Examination of the synonyms of the female keyword *skömm* 'shame' confirmed the counter-intuitive suggestion that this was an indirect keyword, substantiating the effectiveness of the semantic groupings in signposting such patterns. The discussion of synonymous constructions with the terms *þykja* 'seem', *hugur* 'mind', *vilja* 'will, want' and *skulu* 'shall, should' demonstrated that semantic groupings may be correct and yet not exclude the possibility that direct differences in communication strategies have a role to play in explaining keyword distributions. Furthermore, these data alongside the investigation into agentive and experiencer verb keywords demonstrated that semantic groupings of keywords can signpost much wider tendencies in male and female speech.

A key implication of these results is that to draw conclusions about speaker roles or speech patterns directly from a list of keywords would be highly simplistic. This approach would not only fail to make the basic distinction between direct and indirect keywords but would also miss the opportunity to identify subtler linguistic patterns which keywords signpost. It should be clear that when faced with the ambiguous evidence that keyword data provide, it is important that the researcher use every tool at their disposal to illuminate more detail before attempting to propose accounts for the observed patterns.

In the particular case of these data, the evidence has largely pointed towards narrative and contextual explanations for the keywords identified. Some evidence has been found for pragmatic differences which might result from systematic power imbalances between the male and female social roles depicted in the saga narratives and some for differences in terms of address, again likely to result from different properties of gender roles. None of the keywords examined here were shown unequivocally to be direct keywords; thus no clear evidence has been identified here for differences in the LINGUISTIC VARIETIES represented for male and female characters in these texts. The one keyword in these data which can be argued to point to such a difference (*ekki* 'not') is discussed in Blaxter (2013).

ACKNOWLEDGEMENTS

I am indebted to my advisors at Oxford and Cambridge, Aditi Lahiri, Peter Barber and David Willis, for all of their help and guidance. I would also like to thank two anonymous reviewers at the *Nordic Journal of Linguistics* for their insightful comments.

NOTES

1. Gries (2005) tempers this warning, pointing out that Kilgarriff (2005) did not use the Bonferroni correction (see Section 1.4 below) and carried out the experiment of selecting pseudocorpora only once.
2. For example, to obtain only a 5% chance of type 1 error occurring over 100 tests, the threshold p-value would be $.05 \div 100 = .0005$.
3. The results are as follows: f frequency = 544, m frequency = 3274, ratio deviation = 3.30%, $p = 4.8517 \times 10^{-11}$.
4. A relatively complex example is *hinn helgi Ólafur konungur bróðir minn* 'my brother the holy king Ólafur' (from 'Þorsteins þáttur forvitna').
5. The female names investigated were: *Auður*, *Bergþóra*, *Guðrún*, *Gunnhildur*, *Hallgerður*, *Hrefna*, *Jórunn*, *Melkorka*, *Steingerður*, *Unnur*, *Vigdís*, *Þórdís*, *Þorgerður*, *Þórunn* and *Þuríður*; the male names investigated were: *Björn*, *Egill*, *Gísli*, *Gunnar*, *Hákon*, *Höskuldur*, *Ketill*, *Kjartan*, *Ólafur*, *Snorri*, and *Þorgrímur*.

6. One incidental observation that was clear from these data is that male speakers (m sp) were more likely to talk about male individuals (m ref) and female speakers (f sp) to talk about female individuals (f ref): f sp f ref = 47 (24.48%), m ref = 145 (75.52%); m sp f ref = 159 (16.24%), m ref = 820 (83.76%); $p = .0061$. Again, this should probably be seen as a property of the narratives.
7. The results are as follows: f sp –title = 175 (97.22%), +title = 5 (2.78%); m sp –title = 823 (88.30%), +title = 109 (11.70%); $p = .0002$.
8. The results are as follows: f sp –kin = 155 (86.11%), +kin = 25 (14.89%); m sp –kin = 841 (90.24%), +kin = 91 (9.76%); $p = .0623$.
9. The results are as follows: f ref –kin = 155 (80.73%), +kin = 37 (19.27%); m ref –kin = 841 (91.41%), +kin = 79 (8.59%); $p = 1.2645 \times 10^{-7}$.
10. The results are as follows: f ref –title = 187 (97.40%), +title = 5 (2.60%); m ref –title = 811 (88.15%), +title = 109 (11.85%); $p = 1.3916 \times 10^{-5}$.
11. See Stubbs (1995) for an overview of the MI measure.
12. Strategies of politeness, typically involving the avoidance of direct or bald statements, designed to avoid harm to the listener's negative face. See Brown & Levinson (1987).
13. The results are as follows: f frequency = 26; m frequency = 110; ratio deviation = 8.17%; $p = .0017$.
14. The results are as follows: f frequency = 1; m frequency = 18; ratio deviation = 5.68%; $p = .4275$.
15. Results for 'daring': f frequency = 43, m frequency = 261, ratio deviation = 3.20%, $p = .0739$; results for 'vengeance': f frequency = 30, m frequency = 166, ratio deviation = 4.36%, $p = .0505$.

REFERENCES

Anthony, Lawrence. 2012. Antconc: A Freeware Concordance Program for Windows, Mactintosh OS X, and Linux. http://www.antlab.sci.waseda.ac.jp/software.html (30 November 2012).

Baker, Paul. 2004. Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics* 32, 346–359.

Baker, Paul. 2008. *Sexed Texts: Language, Sexuality and Gender*. London: Equinox.

Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, Paul. 2012. Mars and Venus reappraised: Using the Manhattan Distance to explore the sex differences paradigm in the BNC. Presented at the UCREL Corpus Research Seminar Series, Lancaster.

Bestgen, Yves. 2014. Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary & Linguistic Computing* 29(2), 164–170.

Blaxter, Tam T. 2013. Sociolinguistic variation in the Old Icelandic family sagas. MPhil thesis, University of Oxford.

Brown, Penelope & Steven C. Levinson. 1987. *Politeness: Some Universals in Language Usage* (Studies in Interactional Sociolinguistics 4). Cambridge: Cambridge University Press.

Cameron, Deborah. 2007. *The Myth of Mars and Venus*. Oxford: Oxford University Press.

Cleasby, Richard, Guðbrandur Vigfússon & George Webbe Dasent. 1894. *An Icelandic–English Dictionary, Based on the Ms. Collections of the late Richard Cleasby*. Oxford: Clarendon Press.

Culpeper, Jonathan. 2009. Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics* 14(1), 29–59.

Gries, Stefan Th. 2005. Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1(2), 277–294.

Grob, Lindsey M., Renee A. Meyers & Renee Schuh. 1997. Powerful/powerless language use in group interactions: Sex differences or similarities? *Communications Quarterly* 45(3), 282–303.

Harrington, Kate. 2008. Perpetuating difference? Corpus linguistics and the gendering of reported dialogue. In Kate Harrington, Lia Litosseliti, Helen Sauntson & Jane Sunderland (eds.), *Gender and Language Research Methodologies*, 85–102. New York: Palgrave Macmillan.

Helgadóttir, Sigrún, Eyrún Valsdóttir, Auður Rögnvaldsdóttir & Hjördís Stefánsdóttir. 2012–. *Mörkuð íslensk málheild*. Reykjavík: Stofnun Árna Magnússonar í íslenskum fræðum. http://mim.hi.is/index.php (12 May 2012).

Hirschman, Lynette. 1994. Female–male differences in conversational interaction. *Language in Society* 23, 427–442.

Johannessen, Janne Bondi. 2008. The pronominal psychological demonstrative in Scandinavian: Its syntax, semantics and pragmatics. *Nordic Journal of Linguistics* 31(2), 161–192.

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2), 263–276.

Oakes, Michael & Malcom Farrow. 2007. Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries. *Literary & Linguistic Computing* 22(1), 85–99.

Rayson, Paul. 2004. Keywords are not enough. Presented at the Invited Talk for JAECS (Japan Association for English Corpus Studies), Chuo University, Tokyo. http://www.comp.lancs.ac.uk/~paul/publications/jaecs_tokyo04.pdf (22 December 2013).

Rayson, Paul. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4), 519–549.

Rayson, Paul, Geoffrey Leech & Mary Hodges. 1997. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2(1), 133–152.

Rögnvaldsson, Eiríkur & Sigrún Helgadóttir. 2011. Morphological tagging of Old Norse texts and its use in studying syntactic variation and change. In Caroline Sporleder, Antal van den Bosch & Kalliopi Zervanou (eds.), *Language Technology for Cultural Heritage*, vol. 2, 63–76. Berlin: Springer.

Schmid, Hans-Jörg. 2003. Do men and women really live in different cultures? Evidence from the BNC. In Andrew Wilson, Paul Rayson & Tony McEnery (eds.), *Corpus Linguistics by the Lune* (Łódź Studies in Language 8), 185–221. Frankfurt: Peter Lang.

Shibamoto Smith, Janet S. 2004. Language and gender in the (hetero)romance: 'Reading' the ideal hero/ine through lover's dialogue in Japanese romance fiction. In Shigeko Okamoto & Janet S. Shibamoto Smith (eds.), *Japanese Language, Gender, and Ideology*, 113–130. Oxford & New York: Oxford University Press.

Shibamoto Smith, Janet S. & Deborah J. Occhi. 2009. The green leaves of love: Japanese romantic heroines, authentic femininity, and dialect. *Journal of Sociolinguistics* 13(4), 524–546.

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of trouble with quantitative studies. *Functions of Language* 1(2), 23–55.

TEI Consortium (ed.). 2007. TEI P5: Guidelines for electronic text encoding and interchange, 2.1.0. Last updated 17/06/12. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (16 August 2012).

Xiao, Zhonghua & Anthony McEnery. 2005. Two approaches to genre analysis: Three genres in Modern American English. *Journal of English Linguistics* 33, 62–82.