Language Variation and Change, **4** (1992), 41–75. Printed in the U.S.A.
© 1992 Cambridge University Press 0954-3945/92 $5.00 + .00

Dimensions of register variation in Somali

DOUGLAS BIBER
MOHAMED HARED
Northern Arizona University

ABSTRACT

The present study uses a multidimensional approach to analyze the linguistic characteristics of Somali spoken and written registers. Somali is unusual in that it has a very short history of literacy (only since 1973), but at present it has a wide range of written and spoken registers, including governmental, educational, and public information uses. It thus represents a very different language type from previously described languages. We analyze the distribution of 65 linguistic features across 279 texts from 26 spoken and written registers, using factor analysis to identify five major dimensions of variation. None of these dimensions defines an absolute dichotomy between spoken and written registers, although three of the dimensions can be considered "oral/literate" parameters. As in the multidimensional analyses of other languages, the present study shows that no single dimension adequately describes the relations among spoken and written registers; rather, each dimension reflects a different set of communicative functions relating to the purpose, general topic, degree of interactiveness, personal involvement, production circumstances, and other physical mode characteristics. In the conclusion, we briefly discuss our findings relative to previous multidimensional analyses of English, Tuvaluan, and Korean, laying the foundation for cross-linguistic analyses of universal tendencies of register variation.

Over the last two decades, linguists have paid considerable attention to comparisons of spoken and written language. Earlier scholars focused on one or the other mode as "true" language, either disregarding spoken language because it was considered "corrupt" or disregarding written language because it was considered "derivative." In contrast, recent analysts have come to regard the two modes as equally valid, although clearly different, and they have thus set out to describe the linguistic characteristics of each.

Surveys, such as Akinnaso (1982) and Chafe and Tannen (1987), showed that this enterprise is not straightforward. In particular, there are two major methodological difficulties: (a) how to represent "speech" and "writing" from a textual point of view (e.g., what kinds of texts and how many texts), and (b) how to represent the linguistic characteristics of speech and writing

Support for this research project has been provided by the National Science Foundation, grant #BNS-8811720. We thank Ann Beck, Niko Besnier, Susan Foster-Cohen, and two anonymous reviewers for their helpful comments on earlier drafts of this article.

(i.e., which linguistic features should be analyzed). Studies such as Chafe (1982) addressed the second issue by identifying several functionally important linguistic features that distinguish between spoken and written varieties; other studies, such as Tannen (1982a) and Beaman (1984), additionally addressed the first issue by controlling the communicative task (narratives in these studies) compared across the two modes.

In a series of multidimensional analyses of English (e.g., Biber 1986, 1988, 1989), these methodological issues were addressed by comparing a large number of spoken and written text varieties along several linguistic dimensions, where each dimension comprises a set of co-occurring linguistic features. Interpretive labels are proposed for the dimensions based on the shared functions underlying the co-occurring linguistic features: for example, "involved versus informational production," which comprises features such as first and second person pronouns, contractions, and emphatics versus nouns, prepositions, and attributive adjectives; "narrative versus nonnarrative concerns," which comprises features such as past tense verbs, third person pronouns, and perfect aspect; and "explicit versus situation-dependent reference," which comprises features such as *wh-* relative clauses versus time and place adverbials.

Two major conclusions were drawn from these multidimensional studies. First, no single dimension of variation is adequate to account for the range of similarities and differences among registers; rather, multidimensional analyses are required. Second, there is no absolute difference between spoken and written language; rather, particular types of speech and writing are more or less similar with respect to different dimensions. However, these studies confirm the status of conversation as stereotypical "oral" language, and of academic/institutional prose as stereotypical "literate" language; they show that the spoken and written modes have different ranges of *potential* variation, with written registers having a wider range of variation than spoken registers (see esp. Biber, 1988:161–164).

One shortcoming of many previous studies is that they focused almost exclusively on English, and thus they are not representative of speech and writing in any universal sense. A few studies, though, have analyzed spoken/written differences in other languages. For example, Clancy (1982) compared spoken and written narratives in Japanese with respect to several linguistic features (e.g., verb morphology, reference, word order, and dependent clauses). Li and Thompson (1982) described some of the linguistic differences between spoken and written Chinese, considering classical Chinese, modern written Chinese, and spoken Mandarin. Other studies include Deibler (1976) on Gahuku, Duff (1973) on Amuesha, Hurd (1979) on Nasioi, and Irwin (1976) on Chimbu.

Besnier's (1988) study of speech and writing in Nukulaelae Tuvaluan was the first to analyze the overall relations among spoken and written varieties in a non-Western language. Adopting the multidimensional analytical framework developed for English, this study identified three primary dimensions of variation in Nukulaelae Tuvaluan, characterized as: "attitudinal versus authoritative discourse," "focus on information versus interaction," and "rhetorical manipulation versus structural complexity." Besnier found no overall difference between speech and writing in Nukulaelae Tuvaluan and concluded that "spoken and written registers were found to be stylistically interrelated in a complex manner" (1988:731), replicating the overall conclusions for English. However, there are some interesting differences between Nukulaelae Tuvaluan and English with respect to the linguistic and functional correlates of particular dimensions, and with respect to the situational and linguistic characteristics of particular registers.

Kim (1990), using the same analytical framework, also found a complex set of relations among spoken and written registers in Korean. This study identified five major dimensions of variation: "informal interaction versus explicit elaboration," "discourse chaining versus discourse fragmentation," "stance," "narrative concern," and "honorification." In addition, Kim compared the patterns of variation for English, Tuvaluan, and Korean, finding strong cross-linguistic correspondences for several of the dimensions but concluding that the discourse chaining and honorification dimensions are unique to Korean.

The present study focuses on register variation in Somali, a Cushitic language spoken by five to six million people in East Africa (the country of Somalia and adjoining territories in Kenya, Ethiopia, and Djibouti). Somali has existed as a written language only since 1973, when the government of Somalia named it as the official language of the country. Somali literacy expanded rapidly, though, so that in a short time after 1973 there were many official and professional varieties of writing in Somali, including dictionaries, grammars, government documents, textbooks, newspapers, histories, biographies, storybooks, and personal letters.

An analysis of Somali speech and writing complements previous multidimensional analyses in several ways. First, the languages studied to date are from four quite different language families: Indo-European, Austronesian, Altaic, and the Cushitic subfamily of Afroasiatic. They are markedly different in their geographic locations, as well as in their cultural and religious associations. The languages also differ with respect to their status. English is a world language with a long history of literacy and standardization and a wide range of spoken and written registers; Nukulaelae Tuvaluan represents the other extreme in that it has very few speakers, a relatively short history of literacy, only two written varieties, and a generally restricted range of spoken registers.[1] Korean has a long history of literacy and a wide range of spoken and written registers. Finally, Somali has a very short history of literacy, but at present it also has a wide range of spoken and written registers.

One eventual goal of these multidimensional analyses is a cross-linguistic account of register variation, to uncover universal tendencies in the relations among spoken and written varieties. Hymes (1974:41) emphasized the importance of similar goals: "it is essential to isolate the dimensions and features underlying taxonomic categories. These features and dimensions, more than particular constellations of them, will be found to be universal and hence elementary to descriptive and comparative frames of reference." Any such generalizations, however, must be based on linguistic and situational analyses of several different language types. The present study, by addressing these issues with respect to Somali, provides an additional step toward this goal.

METHODOLOGY

Speakers and writers in Muqdisho

Fieldwork for the current project was carried out in Somalia during the calendar year 1989. Most texts used in our analysis were collected in Muqdisho, the capital city.[2] Because Muqdisho is the melting pot of Somalia, the speakers and writers of the texts in our corpus come from several different geographic regions and clans. All addressors, though, used some variety of Common Somali, and most of them originally came from the northwestern regions of Somalia. The geographic dialect variation in our corpus is thus restricted primarily to spelling/pronunciation differences, lexical differences, and a few morphological variants.

We attempted to collect texts from a demographically diverse group of speakers. Social networks in Muqdisho are structured primarily along clan groupings, although factors such as education, occupation, and gender are also important. Most of the participants in our study were well-educated Somalis in professional occupations. This skewing is due to the fact that we relied primarily on Hared's own primary social networks to collect conversational data. However, to the extent possible, we included a wide range of demographic diversity in the corpus, including speakers from: (a) several educational backgrounds (ranging from little formal education to university degrees), (b) several occupations (e.g., unemployed, drivers, policemen, teachers, administrators), (c) both men and women (conversing in both same-sex and cross-sex interactions), (d) religious (Islamic) and more "modern/secular" world views, and (e) a wide range of ages (from c. 20 years to c. 60 years). All texts were recorded in naturalistic settings.

Most written texts in Somali are produced in Muqdisho by a demographically restricted range of writers. Preceding the introduction of Somali literacy in 1973, written materials were produced in Italian, English, and Arabic. The shift to Somali after 1973 was quite abrupt, so that there were very few subsequent materials produced in Italian or English. For example, after January 1973, the government required that all documents, memos, newspapers, and magazines be written in Somali.[3] In the religious domain, written Somali has not replaced Arabic for the most part, because the Quran is considered untranslatable and Arabic is preferred for related writings on Islam.

In the 1970s, most writers of Somali were administrators, teachers, and journalists who had previously been educated in Italian or English. By the early 1980s, however, writers who had been educated in Somali began to enter the workforce. There are still few actual producers of written texts, however. All published Somali texts are produced in Muqdisho by educated journalists, teachers, fiction writers, and government employees. Government employees in other cities also produce frequent memos and official letters, whereas the production of written texts by nonprofessionals is restricted primarily to personal letters, petitions, and notes.

TABLE 1. *Composition of the synchronic Somali corpus of written texts*

Register	No. of Texts
A. *Wargeysyada* Press	
war news reportage	14
faallo institutional editorials	10
ra'yiga iyo aqoonta invited editorials	5
maqaal guud general interest articles	6
maqaal gaar analytical articles	11
iidheh iyo ogeysiis announcements & notices	10
faaqidaadda ciyaaraha sports reviews	8
B. *Qoraalo dawladeed* Government documents	
wareegto memos	10
qoraal barabagaandha political propaganda pamphlets	3
khudbad siyaasadeed published political speeches	5
C. *Qoraalo dadweyne* Personal adult writing	
warqad letters	10
arji applications or petitions	8
D. *Qoraalo waxbarasho* Educational/Academic texts	
buugta dugsiyada sare high school textbooks	10
qoraal cilmiyeed academic essays and theses	20
arar book introductions	5
E. *Suugaan* Literature	
sheeko mala-awaal fiction	19
sheeko-xariir folklore stories	4
Total written texts	158
Total written registers 17	

A fuller analysis of literacy practices in Somalia is required, providing details of who actually produces and consumes written texts, and for what purposes. The focus of our study, however, is on the linguistic and situational correlates of Somali registers, rather than on the social correlates. For this purpose, we collected and analyzed texts from all available spoken and written registers (even though relatively few Somalis actually produce some of those registers).

Texts and registers

Tables 1 and 2 summarize the texts analyzed for the present study. Our entire corpus comprises 557 texts, representing over 700,000 words of text.

TABLE 2. *Composition of the Somali corpus of spoken texts*

Register	No. of Texts
A. *Sheekayn* Conversation and story telling	
sheeko spontaneous/conversational narratives	20
hadal caadi nonnarrative conversation	21
B. *Hadal jeedin* Public monologues	
cashar jaamcadeed university lectures	10
lakjar academic conference lectures	10
wacdi and *tafsiir* sermons	20
C. *Hadal raadiyo* Spontaneous radio broadcasts	
ciyaar-tebin live sports broadcasts	10
D. *Hadal shir* Formal conversation and meetings	
shir guddi formal committee meetings	11
shir qoys family meetings	9
falanqo lakjar discussions following lectures	10
Total spoken texts	121
Total spoken registers	9

Many of these are written texts from earlier time periods (1973–1974 and 1977–1979), which we have used to trace the diachronic development of written registers (Biber & Hared, 1992, in press). However, the present study is based entirely on texts from the contemporary period (1988–1989): 158 written texts and 121 spoken texts. We grammatically tagged and analyzed the first 1,000 words from each text, so the present study is based on approximately 279,000 words of text.[4]

We undertook a fairly exhaustive sampling of written registers, including all available written varieties and all available texts for some categories. The press registers are probably the best developed written varieties in Somali, because newspapers have been in continuous existence since January 1973. In 1989 there was both a daily paper (*Xiddigta Oktoobar*) and a weekly (*Ogaal*). We included several registers from newspapers. News reportage articles (*war*) were taken from the front page of the papers; institutional editorials (*faallo*) are commentaries with no acknowledged author taken from a titled editorial page. Invited editorials are also taken from a titled page of the newspaper (*ra'yiga iyo aqoonta* 'opinion and knowledge'); these are articles written by "experts" on specialized topics such as the economy, international relations, and Somali culture and history.

Announcements and notices (*iidheh iyo ogeysiis*) are presented in a special section of the newspaper. These include announcements about meetings and public events (e.g., new plays, new books, the opening of a new shop or hotel), as well as notices to individuals (e.g., a summons to appear in court or to pay an electricity bill). Sports reviews (*faaqidaadda ciyaaraha*) present relatively in-depth discussions of current sports news. Finally, general interest articles (*maqaal guud*) and analytical articles (*maqaal gaar*) occur on the inside pages of newspapers, but do not have their own titled sections. General interest articles are on topics such as the value of friendship, the danger of using drugs, and problems with hair dyes, whereas analytical articles are longer and deal with more specialized issues (e.g., international relations, economics, the environment).

Government documents include official memos and published pamphlets. The memos (*wareegto*) are official policy statements written by some governmental office. Political propaganda pamphlets (*qoraal barabagaandha*) and booklets of published political speeches (*khudbad siyaasadeed*) are distributed by the government on topics such as the advantages of socialism and the contributions of the Revolution to the development of the country.

Personal letters (*warqad*) were collected from friends and relatives, of various educational backgrounds, ages, and both sexes. They are truly personal, being addressed to close relatives or friends. Petitions (*arji*) are personal, but official in purpose. They are addressed to an individual in an official capacity, typically to request some action (such as a leave or transfer from a job, a passport, or a loan).

High school textbooks (*buugta dugsiyada sare*) are samples taken from the only textbook series in the country. Academic prose (*qoraal cilmiyeed*) actually combines two subregisters: published academic essays and theses. Book introductions, which are labeled either as *arar* or *hordhac*, present the background and writing history of a book.

The category of fiction or imaginative stories (*sheeko mala-awaal*) also combines two subregisters: short novels and serial stories published in the newspaper. Folklore stories (*sheeko-xariir*, lit. 'silk [i.e., entertaining] stories') are booklets of traditional stories.

Our collection of spoken texts includes a full range of urban spoken registers. We include two conversational registers: plain conversation (*hadal caadi*) and conversational narratives (*sheeko*, lit. 'stories'). These both represent face-to-face interactions among friends or relatives. Conversational narratives are produced primarily by a single speaker, describing either the personal experiences of the narrator or stories about other participants. Conversations include discussions of the day's events, jokes, and other casual forms of interaction.

Two types of lectures were collected: from the university and from an academic conference. *Cashar jaamcadeed* are class lectures given by teachers at Somali National University. We use the borrowed term *lakjar* for academic conference lectures given at the Fourth International Congress of Somali Studies in Muqdisho (June 1989). *Falanqo lakjar* are taken from the discussions after lectures at this same conference.

Sermons (*wacdi* and *tafsiir*) are formal religious presentations, either at the weekly Islamic service or in a more private setting. They include didactic explanations of the Quran and Hadith, as well as social and spiritual exhortations.

Public sporting events were rarely covered by radio broadcasts during the year that we were in Somalia, but one soccer game combined with a marathon race was broadcast. Approximately five reporters took turns covering the event, describing the action on the playing field plus occasionally providing additional background information. The 10 sports broadcasts in our sample are from this single event.

Finally, we collected texts from two kinds of committee meetings. The first, *shir guddi*, are formal meetings, such as the regular staff meetings of the newspaper *Ogaal*. Family meetings (*shir qoys*) are more intimate planning sessions to discuss various financial and (inter)personal issues. Our text samples for this register are taken from a single meeting of Hared with his family.

Linguistic features used in the analysis

We analyzed over 70 linguistic characteristics of Somali texts using computer programs written in Pascal. The first program tags each word in a text for its grammatical category. This program works in a cyclical fashion to build an on-line dictionary containing an entry for every word in the Somali corpus: new words are checked interactively and entered into the dictionary; subsequent occurrences of a word are tagged automatically by the program. Ambiguous forms, including many verbs, are tagged interactively. For example, there is no relative pronoun in Somali, so typically there are no overt surface contextual differences between relative clauses and main clauses; these forms thus must be tagged interactively to ensure accurate identification. After tagging was completed, a second program was run to count the frequency of each feature in each text; this program also computed additional features, such as t-unit length and type/token ratio. The statistical analyses in the following section are based on these frequency counts. (All frequency counts are normalized to their frequency per 1,000 words of text.)

We set out to include all linguistic features that might have functional associations in Somali. We used Saeed (1984, 1987) as our main grammatical reference, although we added several features based on our own analyses of the text corpus. There are marked linguistic differences in the grammatical systems of English and Somali. For example, every sentence in Somali must have some kind of overt focus marker, typically a focus particle (*baa*, *ayaa*) that marks a noun phrase as new information or a clefting construction (*waxaa*) that can put an entire clause into focus. Another focus particle (*waa*) has been analyzed as the unmarked case, but our corpus shows that this construction is relatively rare; one use of this particle is for verbless clauses (*waa* + NP). As Saeed (1984) showed, relative clause constructions are central to Somali syntax, including the syntax of focus constructions. In fact, in addition to postnominal modification and clefting constructions, relative clauses are used for many types of adverbial subordination in Somali. (For example, temporal clauses, such as *markii aan tagay* 'when I went . . .', are literally relative clauses with 'the time' as the head noun, i.e., 'the time [that] I went'.)

Although it is not possible to include grammatical descriptions of these linguistic features here, many of the features are illustrated in the text samples included in the Appendix. Further information about the grammatical characteristics of Somali is available in the Saeed references. Several features were excluded from the factor analysis, either because they overlapped completely with other forms that were included (e.g., definite nouns, indefinite nouns, and demonstrative nouns overlap with common nouns, derived nouns, and compound nouns), or because they had low communalities in preliminary factor analyses (e.g., manner clauses, reason clauses, and *must* clauses were dropped for this reason). Other features were combined to reduce the total number of features in the analysis (e.g., adjectival and verbal dependent clauses were combined into single categories). The final factor analysis was thus based on the 65 linguistic features listed in Table 3, representing 11 grammatical and functional categories: dependent clauses, main clause and verb features, nominal features, pronouns, adjectival features, special lexical classes, features reflecting lexical choice, preverbal particles, reduced and interactive features, coördination, and focus constructions.[5]

Linguistic co-occurrence and the Multidimensional approach

The Multidimensional (MD) approach to genre or register variation (earlier referred to as the Multifeature/Multidimension approach) is outlined in Biber (1986) and developed more fully in Biber (1988). The approach is based on the centrality of linguistic co-occurrence in analyses of text variation. Theoretical antecedents to this approach are provided by Ervin-Tripp (1972), Hymes (1974), and Brown and Fraser (1979). For example, Brown and Fraser (1979:38–39) observed that it can be "misleading to concentrate on specific, isolated [linguistic] markers without taking into account systematic variations which involve the cooccurrence of sets of markers."

In the MD approach, linguistic co-occurrence is analyzed in terms of underlying dimensions of variation, with the explicit assumption that multiple dimensions will typically be required to account adequately for the range of linguistic variation among registers in a language. Dimensions are continuous scales of variation (rather than dichotomous distinctions), identified quantitatively by a factor analysis (rather than on an a priori functional basis).

Each dimension (or factor) comprises a group of linguistic features (e.g., nominalizations, adjectives, relative clauses) that co-occur with a markedly high frequency in texts. Factor analysis is used to identify the groups of linguistic features associated with each dimension. The interpretation of the factors as functional dimensions is based on the assumption that co-occurrence reflects shared function; that is, features co-occur frequently in texts because

TABLE 3. *List of linguistic features used in the analysis*

Dependent clauses
1. total dependent clauses
2. conditional clauses
3. purpose clauses
4. concessive clauses
5. temporal clauses
6. framing clauses (similar to nonrestrictive relative clauses)
7. *ah* relative clauses (with a reduced copula and no object)
8. full relative clauses
9. verb complements (Verb + *in-*)
10. demonstrative relative clauses (with demonstrative pronoun as head; concludes a series of relative clauses)
11. *ahaan* adverbials (e.g., *guud ahaan* 'being general' = 'generally')

Main clause and verbal features
12. total main clauses
13. average t-unit length (main clause plus associated dependent clauses)
14. verbless clauses (with copula deleted)
15. independent verbs
16. imperatives
17. optative clauses (*ha* + Verb- 'let X do Y')
18. compound verbs
19. present tense (verbs and adjectives)
20. past tense (verbs and adjectives)
21. possibility modals (*kar-*)
22. future modals (*doon-*)
23. habitual modals (*lah-*)

Nominal features
24. common nouns
25. proper nouns
26. possessive nouns
27. nominalizations (e.g., *-nimo*, *-tooyo*, *-aan*)
28. gerunds (e.g., *-id*, *-in*, *-is*)
29. agentive nouns (*-e*, *-te*, *-to*, *-so*)
30. compound nouns
31. *-eed* genitives

Pronouns
32. 1st person pronouns (e.g., *ani-*, *-aan*, *i* + Verb)
33. 2nd person pronouns (e.g., *adi-*, *-aad*, *ku* + Verb)
34. 3rd person pronouns (e.g., *isa-*, *iya-*, *-uu*, *-ay*)

Adjectival features
35. derived adjectives (*-(s)an*)
36. attributive adjectives
37. predicative adjectives

Lexical classes
38. stance adjectives (e.g., *jecel* 'like', *neceb* 'hate', *hilmaansan* 'forget'; these function as verbs)
39. stance verbs (e.g., *garo* 'understand', *hilmaan* 'forget', *baq* 'become afraid')
40. speech act verbs (e.g., *sheeg* 'say', *sharax* 'explain')
41. time deictics (e.g., *maanta* 'today', *marar* 'sometimes')
42. place deictics (e.g., *hoos* 'under', *dib* + Verb 'behind')
43. downtoners (e.g., *malaha* 'perhaps', *yara* 'just', 'a little')
44. amplifiers (e.g., *aad* 'really', 'very', *shaki la'aan* 'without doubt')
45. concessive conjuncts (*hase yeeshee* 'however', *laakiin* 'however')
46. reason conjuncts (*waayo* 'the reason [is]', *sidaas darteed* 'as a result')

continued

TABLE 3 (*continued*)

Lexical choice
47. word length
48. hapax legomena (number of once-occurring words in first 500 words)
49. type–token ratio (number of different words in first 500 words)

Preverbal particles
50. single case particles (*u* 'to', 'for', *ku* 'in', 'on', 'at', 'by means of', *ka* '(away) from', *la* 'together with')
51. case particle sequences (e.g., *uga kaga*)
52. impersonal particles (*la*)
53. locative/directional particles (*sii*, *soo*)

Reduced and interactive features
54. contractions
55. *yes/no* questions (*ma* + Verb)
56. *what if* questions (*soo*, *sow*)
57. *wh-* questions (*maxaa-* 'what')
58. simple responses (e.g., *haa* 'yes', *haye* 'ok', *nacam* 'true')

Coordination
59. clause/phrase coordination (*oo*)
60. phrase coordination (*iyo*)
61. contrastive clause coordination (*eh*)
62. clitic topic (clause) coordination (*-na*, *-se*)

Focus constructions
63. *waa* focus markers
64. *baa* focus markers
65. *waxaa* clefts

they serve shared, underlying communicative functions associated with the situational contexts of the texts. The functional interpretations are based on prior analyses of individual linguistic features and on the distribution of the co-occurring features across registers. In the next section, we present the co-occurring features associated with six Somali dimensions of variation, and in the section following that, we present the distribution of registers along five of the dimensions and offer functional interpretations.

PATTERNS OF REGISTER VARIATION IN SOMALI

Six dimensions of variation

In the factor analysis for our investigation of Somali registers, we extracted six factors for consideration.[6] Table 4 summarizes the co-occurring features associated with each of these factors. The decimal numbers on this table represent the factor loadings for each linguistic feature. Loadings can run from −1.0 to +1.0; the further from 0.0 a loading is, the more one can generalize from the factor in question to the particular linguistic feature. Features with larger loadings are thus better representatives of the dimension under-

TABLE 4. *Summary of the six dimensions derived from factor analysis of 279 contemporary texts representing 26 spoken and written registers (see note 7)*

Dimension 1	
Positive features	
simple responses	.97
yes/no questions	.91
contrast clause coordination (*eh*)	.88
stance adjectives	.81
contractions	.74
independent verbs	.73
what if questions (*soo*)	.70
time deictics	.68
waa focus markers	.67
main clauses	.63
baa focus markers	.60
downtoners	.59
imperatives	.58
wh- questions	.47
conditional clauses	.43
2nd person pronouns	.41
1st person pronouns	.30
verbless clauses	.30
Negative features	
dependent clauses	−.63
full relative clauses	−.63
waxaa clefts	−.60
ah relative clauses	−.58
clause coordination (*oo*)	−.54
word length	−.53
common nouns	−.52
derived adjectives	−.52
phrase coordination (*iyo*)	−.52
-eed genitives	−.46
verb complements	−.45
case particle sequences	−.44
single case particle	−.34
t-unit length	−.34
agentive nouns	−.33
compound nouns	−.33
attributive adjectives	−.32
purpose clauses	−.30
ahaan adverbials	−.28
Dimension 2	
Positive features	
hapax legomena	.92
type/token ratio	.88
nominalizations	.54
compound verbs	.48
single case particle	.47
demonstrative relatives	.40
clitic topic coordination	.35
gerunds	.35
purpose clauses	.31
word length	.30
No negative features	
Dimension 3	
Positive features	
present tense	.89
predicative adjectives	.55
possibility modals	.50
concessive conjuncts	.46
verbless clauses	.38
attributive adjectives	.38
derived adjectives	.38
impersonal particles	.37
conditional clauses	.33
dependent clauses	.31
Negative features	
past tense	−.58
proper nouns	−.54
agentive nouns	−.45
framing clauses	−.36
future modals	−.32
speech act verbs	−.29
Dimension 4	
Positive features	
3rd person pronouns	.87
past tense	.69
temporal clauses	.61
waxaa clefts	.43
habitual modals	.40
stance verbs	.31
possessive nouns	.31
concessive clauses	.29
Negative features	
compound nouns	−.37
gerunds	−.30
agentive nouns	−.30
t-unit length	−.31
phrase coordination (*iyo*)	−.31
Dimension 5	
Positive features	
optative clauses	.60
2nd person pronouns	.59
1st person pronouns	.55
directional preverbal particles	.47
single case particles	.43
clitic topic coordination	.38
imperatives	.36
independent verbs	.32
possessive nouns	.32
case particle sequences	.32
No negative features	
Dimension 6 (Personal persuasion)	
Positive features	
amplifiers	.60
1st person pronouns	.52
reason conjuncts	.46
verb complements	.40
framing clauses	.39
future modals	.31
2nd person pronouns	.30
Negative features	
place deictics	−.39
phrase coordination	−.38
-eed genitives	−.35

lying a factor. In Table 4, only features with loadings larger than 0.30 (plus or minus) are included.[7]

Most of the dimensions consist of two groupings of features, having positive and negative loadings. Positive or negative sign does not indicate the strength of the relationship; rather, these two groups represent sets of features that occur in a complementary pattern. That is, when the features in one group occur together frequently in a text, the features in the other group are markedly infrequent in that text, and vice versa. To interpret the dimensions, it is important to consider likely reasons for the complementary distribution of these two groups of features, as well as the reasons for the co-occurrence pattern within each group.

Consider the features grouped on Dimension 1 in Table 4. The positive features include: (a) nondeclarative, interactive, sentence types: *yes/no* questions, *what if* questions, imperatives, *wh-* questions; (b) interactive or involved lexical classes: responses (e.g., *haye* 'ok'), stance adjectives (adjectives such as *neceb* 'hate' and *jecel* 'like' functioning in predicative positions as personal expressions of feeling, e.g., *waan jeclahay* . . . 'I like . . .'), time deictics (e.g., *shaley* 'yesterday'), and downtoners (e.g., *waa laga yaabaa* 'maybe'); (c) main clause features: contrastive main clause coordination (*eh*), independent verbs, main clause focus markers (*waa*, *baa*), total main clauses, and verbless clauses; (d) other "involved" features: contractions, conditional clauses, first and second person pronouns. The negative features include: (a) dependent clause features: dependent clauses, full relative clauses, *waxaa* clefts, *ah* relative clauses, *oo* coordination (which connects dependent clauses, independent clauses, or verb phrases), and verb complements; (b) nominal elaboration: word length, common nouns, derived adjectives, phrase coordination (*iyo*), *-eed* genitives, and attributive adjectives; (c) elaborating phrases in clauses: case particle sequences (marking the inclusion of multiple indirect object case roles in a clause) and single case particles.

The features with positive loadings tend to co-occur in texts. That is, when there are frequent simple responses, *yes/no* questions and contrast clause coordinators in a text, there are also frequent stance adjectives, contractions, and so on. Similarly, the group of features with negative loadings represent a set of co-occurring features; for example, when there are frequent total dependent clauses, relative clauses, and *waxaa* clefts in a text, there will also tend to be high frequencies of common nouns, derived adjectives, and so on. The positive and negative groupings of features belong to a single dimension because they have a strong complementary relation to one another – when the positive features are markedly frequent in a text, the negative features are relatively absent from that text, and vice versa. Thus the interactive, involved, main clause features (with positive loadings) have a complementary distribution to the dependent clause and structural elaboration features (with negative loadings). Based on the distribution of similar linguistic features in English, it is possible to propose a preliminary functional interpretation of this dimension: a cline distinguishing between highly interactive, involved texts (characterized by frequent occurrences of the positive features) and informational, noninvolved texts (characterized by the negative features). We examine the actual distribution of texts and propose a fuller interpretation of this dimension in the following section.

Dimension 2 has only positive features. The stronger loadings on this dimension are lexical characteristics: hapax legomena (once-occurring words), type-token ratio (the number of different words), nominalizations, and compound verbs.[8] Gerunds and word length also load on this factor. Thus this dimension distinguishes primarily between texts having careful and elaborated lexical choice (lexical diversity, rare words, and derivationally complex words) and those using frequent repeated lexical forms that are short and derivationally simple.[9]

Dimension 3 shows a basic opposition between present tense and past tense. However, as indicated by the co-occurrence of past tense verbs and future tense modals (both with negative loadings), this dimension does not represent a simple dichotomy between present and past events. In addition to present tense, the positive features in Dimension 3 are: adjectival forms (predicative adjectives, attributive adjectives, and derived adjectives), qualified statements (possibility modals, concession conjuncts, and conditional clauses), impersonal constructions (clauses with impersonal agents), and other clausal features (verbless clauses and total dependent clauses). The features with negative loadings in Dimension 3, in addition to past tense, are: animate/human references (proper nouns and agentive derived nouns, which are similar to nouns derived by *-er* in English), future modals, framing clauses, and speech act verbs. The negative features represent projected time (past or future) with a focus on specific human referents; the positive features represent present time, with frequent elaborating details and qualifying conditions and concessions.

The grouping of features in Dimension 4 corresponds closely to a dimension identified in the analyses of English and Korean. The major features are third person pronouns, past tense verbs, temporal clauses, *waxaa* clefts, and habitual modals; based on cross-linguistic expectations, this co-occurrence pattern marks narrative discourse versus other discourse types.[10]

The strongest features in Dimension 5 are optative clauses, which function as polite commands or wishes (translated as 'let X do Y'), second person pronouns, first person pronouns, and directional preverbal particles (*soo*, *sii*, which mark action toward or away from the speaker/writer). Single case particles, *-na* coordination (which coordinates clauses while topicalizing the preceding noun phrase), and imperatives (direct commands) have lower loadings. These features seem to relate to a type of direct interaction between addressor and addressee.

Finally, the major positive features in Dimension 6 are amplifiers, first person pronouns, reason conjuncts, verb complement clauses, and framing clauses. The major features with negative loadings are place deictics, phrase coordination, and *-eed* genitives. Our preliminary interpretation of this dimension is that it reflects a kind of personal persuasion. However, because it is less well represented linguistically, and less transparent functionally, we do not discuss the interpretation of this dimension further in the present article.

TABLE 5. F *scores and correlations for mean dimension score differences among 26 spoken and written registers* (df = *25,253)*

Dimension	*F* Value	Probability	r^2
1	101.0	$p < .0001$	90.9%
2	13.7	$p < .0001$	57.5%
3	11.8	$p < .0001$	53.8%
4	19.4	$p < .0001$	65.7%
5	22.9	$p < .0001$	69.3%

Relations among Somali spoken and written registers

Although the identification of these dimensions is important in itself – in that it isolates several of the basic parameters of variation in Somali – the primary use of the dimensions is to analyze the linguistic characteristics of texts and registers. This can be accomplished by computing dimension scores for each text: a summation, for each dimension, of the frequencies of all features having salient loadings on the dimension. We used only features with loadings over 0.45 in the computation of dimension scores, and we transformed the dimension scores to aid in comparability across dimensions.[11] Dimension scores should not be interpreted in absolute terms; they are useful only for relative comparisons among texts and registers. The transformations do not alter the relative relations among registers or the strength of each dimension; their purpose is simply to facilitate comparisons across dimensions.

For example, the Dimension 1 score for each text is computed by adding together the frequencies of simple responses, *yes/no* questions, *eh*-coordination, stance adjectives, and so forth (the features with positive loadings) and then subtracting the frequencies of total dependent clauses, relative clauses, *waxaa* clefts, and so forth (the features with negative loadings) (see Table 4). The resulting score provides an overall characterization of each text with respect to Dimension 1. Then, the mean of these Dimension 1 scores for each register is computed. Consideration of these dimension scores enables linguistic characterization of any given text or register, comparison of the relations between any two registers, and a fuller functional interpretation of the underlying dimension.

Table 5 shows that all five dimensions are associated with important, systematic differences among the registers. The *F* values (and probabilities) show that the registers are significant discriminators for each dimension; the r^2

values show that they are important (r^2 is a direct measure of the percentage of variation in the dimension score that can be predicted on the basis of the register distinctions). All five of the dimensions have r^2 values over 50%, and three of them have r^2 values over 60%. Dimension 1 has an extremely large r^2 value of 90.9%. Thus, all five dimensions are important predictors of register variation. However, each dimension defines a different set of relations among registers, described in the following sections.

Relations along Dimension 1. Consider Figure 1, which plots the mean dimension scores of the 26 spoken and written Somali registers with respect to Dimension 1. The registers with large positive values, such as conversations and family meetings, have high frequencies of *yes/no* questions, stance adjectives, contractions, main clauses, and so on (the features with positive loadings in Dimension 1), together with markedly low frequencies of total dependent clauses, relative clauses, nouns, derived adjectives, and so on (the features with negative loadings in Dimension 1). Text Sample 1 in the Appendix, from a personal conversation, illustrates these linguistic characteristics. Registers with large negative values, such as political pamphlets and editorials, have the opposite linguistic characteristics: very high frequencies of dependent clauses, nouns, and so forth, plus low frequencies of *yes/no* questions, contractions, and so forth. Text Sample 2 in the Appendix, from a press editorial, illustrates many of the features with negative loadings in Dimension 1.

The characterizations of registers shown in Figure 1, together with consideration of the linguistic features grouped in Dimension 1 (Table 4), enable a fuller functional interpretation of this dimension. (Spoken registers are capitalized and written registers are underlined in Figures 1–5.) The two extremes of Dimension 1 characterize personal involvement versus informational exposition. The positive extreme characterizes three markedly involved registers: conversations, family meetings, and conversational narratives, whereas at the negative extreme there is a very tight cluster of informational, expository registers (e.g., editorials, political pamphlets, press reportage). In between these two extremes, there are a number of spoken registers and three written registers. These intermediate registers are also distributed according to their focus on personal involvement versus informational exposition. Among these intermediate spoken registers, sermons and conference discussions are relatively involved, whereas lectures and sports broadcasting are relatively informational. Among the intermediate written registers, folk stories and personal letters are relatively involved, whereas general fiction is more informational.

Although the poles clearly separate spoken and written registers, this dimension does not define a spoken/written dichotomy; rather, folk stories and personal letters are written, but have relatively involved characterizations, whereas lectures, formal meetings, and sports broadcasts are spoken with relatively expository and elaborated characterizations. Similarly, Dimension 1 does not clearly distinguish between interactive and noninteractive registers. For example, among the spoken registers, conversational narratives are somewhat less interactive than conversations, but just as marked on this dimension. Among the written registers, personal letters are more interactive than folk stories, but they have nearly the same characterization (with folk stories actually being slightly higher).

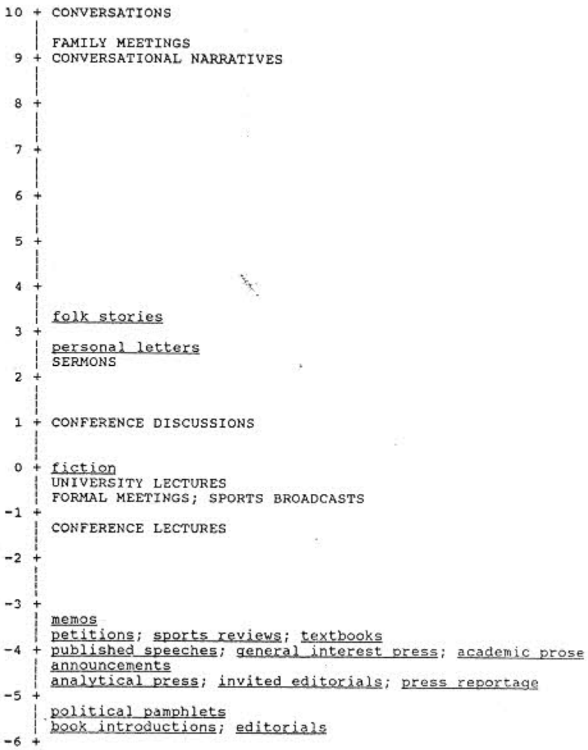

FIGURE 1. Distribution of registers along Dimension 1 – Structural elaboration: Involvement versus exposition ($F = 101.0$, $p < .0001$, $df = (25,253)$, $r^2 = 90.9\%$).

Although this dimension is functionally related to mode and interactiveness distinctions, the primary underlying parameter here seems to reflect different author/speaker purposes: a cline from personal involved expression to informational exposition. Linguistically, many of the features in Dimension 1 relate to structural elaboration (the negative features) or the lack of it (the positive features). This dimension can thus be labeled "Structural elaboration: Involvement versus exposition." (We refer to this as the "structural elaboration" dimension for short.)

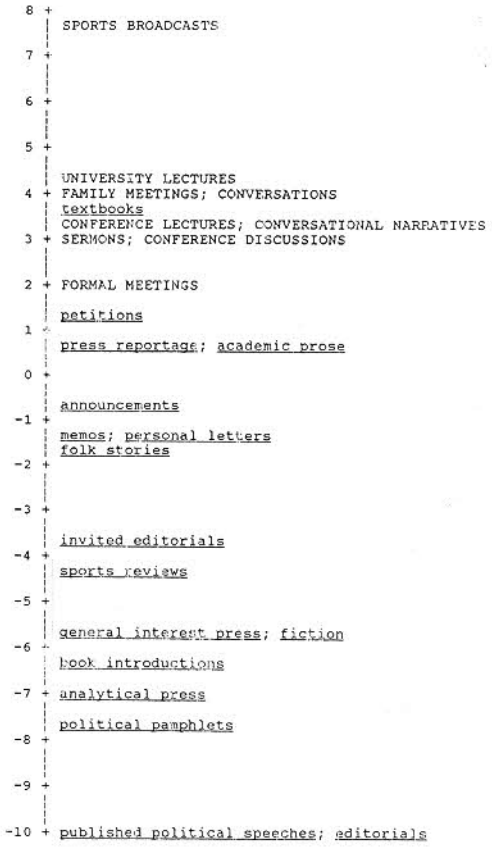

FIGURE 2. Distribution of registers along Dimension 2—Lexical elaboration: On-line versus planned/integrated production (polarity reversed) ($F = 13.7$, $p = .0001$, $df = (25,253)$, $r^2 = 57.5\%$).

Relations along Dimension 2. Figure 2 plots the mean dimension scores of Somali registers with respect to Dimension 2. (The poles of this dimension have been reversed to facilitate comparisons across dimensions; see note 11.) Sports broadcasts have the highest positive score, and registers such as lectures, conversations, and family meetings have relatively high positive values; these scores reflect markedly low frequencies of once-occurring words, nominalizations, compound verbs, and gerunds, together with markedly short words and little lexical diversity (low type/token ratio). One of the differences between Dimension 2 and Dimension 1 is that all spoken registers are similar to one another in this dimension (vs. the relatively wide spread of spoken registers seen in the structural elaboration dimension). That is, regardless of purpose (e.g., informational vs. interpersonal), topic (e.g., scientific vs. everyday), and interactiveness (monologue and dialogue), all spoken registers are marked as having little lexical variety and an absence of elaborated lexical items in this dimension.

In contrast, written registers show a wide range of variation in Dimension 2 (vs. the more restricted range of variation seen in Dimension 1, where most written registers were markedly elaborated in structure). Some written registers, such as editorials, published political speeches, political pamphlets, and analytical press, are markedly elaborated in their lexical choice, showing extreme lexical diversity and very frequent use of derived words and longer words (see Text Sample 2 in the Appendix). Other written registers, such as folk stories, memos, and personal letters, are less informational in purpose and thus have more moderate scores here, although they still show greater lexical diversity and elaboration than the spoken registers. Surprisingly, some informational registers (such as press reportage and academic prose) have intermediate scores on this dimension, whereas high school textbooks have a quite high score. These scores reflect the frequent repetition of technical terms, which are often borrowed (from English or Italian) rather than derived from native Somali words. In contrast, the institutional lexical items common in editorials and political registers have generally been created through nominalizing and compounding processes; these forms tend to be longer than borrowed forms, and they are also derivationally complex.[12]

Considering the grouping of linguistic features in Dimension 2, which primarily represent lexical diversity and lexical elaboration, together with the distribution of registers seen in Figure 2, we propose the label "Lexical elaboration: On-line versus planned/integrated production." (We refer to this as the "lexical elaboration" dimension for short.) This dimension seems to represent a basic difference between the production possibilities of speech and writing. All spoken registers, regardless of purpose or interactiveness, are produced on-line and thus show little lexical diversity or elaboration. This restriction is most pronounced in sports broadcasts, where broadcasters must describe events in progress. In contrast, writers have extensive opportunity for careful word choice, and thus written registers can show extreme lexical diversity and elaboration. However, writers of registers such as personal letters, folk stories, and announcements can choose not to exploit the production possibilities of the written mode (because deliberate production is not required by their purposes and topics), resulting in relatively little lexical diversity and elaboration. And writers of registers such as academic prose and textbooks can deliberately restrict the range of lexical diversity, due to the need for precise, technical vocabulary.

Dimensions 1 and 2 show reverse patterns with respect to the range of variation *within* speech and writing. Both dimensions polarize interactive, interpersonal speech at one extreme and informational exposition at the other extreme. Dimension 1, though, shows a quite restricted range of variation among written registers versus a wide range of variation among spoken registers. Apart from folk stories, personal letters, and fiction, all written registers are markedly elaborated in structure, whereas spoken registers range

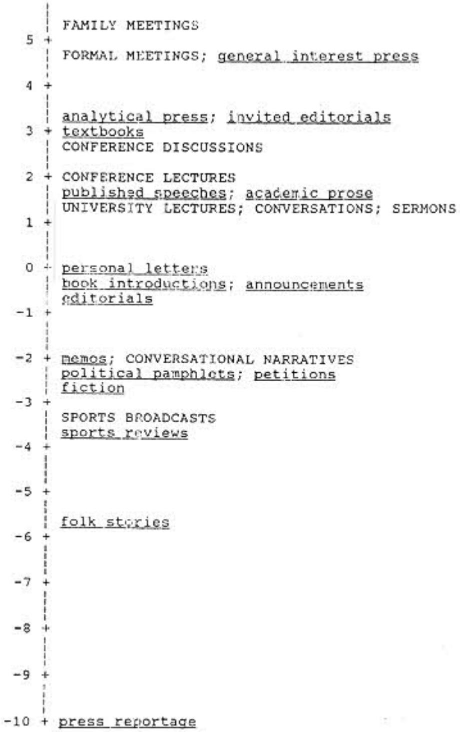

FIGURE 3. Distribution of registers along Dimension 3—Argumentative versus reported presentation of information ($F = 11.8$, $p < .0001$, $df = (25,253)$, $r^2 = 53.8\%$).

from the involved, nonelaborated characteristics of conversations and family meetings to the relatively informational and elaborated characteristics of lectures, formal meetings, and sports broadcasts. In contrast, Dimension 2 defines an extremely restricted range of variation among spoken registers versus a quite wide range of variation among written registers. Both dimensions reflect differences among registers relating to purpose and topic, but Dimension 2 further reflects differences in the production possibilities of speech versus writing—spoken registers are restricted in lexical elaboration due to on-line production constraints, regardless of purpose.

Relations along Dimension 3. Figure 3 plots the distribution of Somali registers with respect to Dimension 3. Family meetings have the highest score on this dimension, similar to Dimension 1. Conversations, however, differ from family meetings in this dimension in having an intermediate score, whereas formal meetings are quite similar to family meetings here. There are also several written registers with quite high scores in Dimension 3, such as general interest press, analytical press, textbooks, and invited editorials. All

registers with high positive scores are characterized by a heavy reliance on present tense forms, plus frequent adjectives, possibility modals, and concessive conjuncts (the positive features in Dimension 3), combined with a marked absence of past tense forms, proper nouns, and agentive nouns (the negative features in Dimension 3).

In contrast, press reportage has by far the largest negative score in Dimension 3, followed by folk stories with a relatively large negative score. These registers have the opposite linguistic characteristics: a heavy reliance on past tense forms plus frequent proper nouns and agentive nouns, combined with markedly few present tense forms, adjectives, possibility modals, and so forth.

Several other registers, both spoken and written, have moderately high scores in Dimension 3. These include the more informational spoken registers, such as conference discussions, conference lectures, university lectures, and sermons, as well as face-to-face conversations. Published political speeches and academic prose also have relatively high scores here.

Considering this distribution of registers, together with the linguistic features grouped on this dimension, we propose the label "Argumentative versus reported presentation of information." We use the term "argumentative" here to refer to a qualified presentation of information, considering a number of different possibilities, whereas "reported" styles simply present the facts, with little consideration of alternative possibilities. This reported presentation of information, as in press reportage, focuses on past events and specific individuals, resulting in the high frequencies of past tense forms, proper nouns, and agentive nouns. In contrast, the argumentative registers, such as family and formal meetings (spoken), or general interest and analytical press (written), focus on the relative merits of present possibilities, resulting in frequent present tense forms, possibility modals, concessive conjuncts, conditional clauses, and so on.

Relations along Dimension 4. The distribution of registers shown in Figure 4 confirms the interpretation of Dimension 4 as distinguishing between "narrative versus nonnarrative discourse organization." Folk stories and general fiction have the highest scores in this dimension, reflecting very frequent use of third person pronouns, past tense forms, temporal clauses, *waxaa* clefts, and habitual modals. These features are associated with the discourse development of narrative story lines, consisting of a temporal sequence of past events in relation to several third persons (the characters). This dimension should thus be contrasted with Dimension 3, which focuses on the description of past (vs. present) events, but often does not include sequencing of a series of events. It should also be noted that conversational narratives do not share these stereotypical narrative characteristics (and thus they have an intermediate score in Dimension 4). Conversational narratives are composed of a sequence of past events, but they often describe events as if they were in the present (or even the future), and they also include several evaluative comments on the described events.

```
 10 + folk stories
    |
  9 +
    |
  8 +
    |
  7 +
    |
  6 +
    | fiction
  5 +
    |
  4 +
    |
  3 +
    |
    | SERMONS
  2 +
    |
    | UNIVERSITY LECTURES; CONVERSATIONAL NARRATIVES
  1 +
    | sports reviews; CONFERENCE LECTURES; CONFERENCE DISCUSSIONS
    |
  0 + invited editorials
    | textbooks; SPORTS BROADCASTS
    | general interest press; CONVERSATIONS
 -1 + academic prose; published speeches
    | press reportage; editorials; FORMAL MEETINGS
    | analytical press; book introductions; FAMILY MEETINGS
 -2 + political pamphlets; personal letters
    |
    |
 -3 + announcements; memos
    |
    |
 -4 +
    | petitions
    |
 -5 +
```

FIGURE 4. Distribution of registers along Dimension 4—Narrative versus nonnarrative discourse organization ($F = 19.4$, $p < .0001$, $df = (25,253)$, $r^2 = 65.7\%$).

```
 10 + personal letters
    |
  9 +
    |
  8 +
    |
  7 +
    |
  6 +
    |
  5 +
    |
  4 +
    | FAMILY MEETINGS
    |
  3 +
    |
    |
  2 + CONVERSATIONAL NARRATIVES; CONVERSATIONS
    | SERMONS
    |
  1 + fiction; petitions; FORMAL MEETINGS
    | SPORTS BROADCASTS; UNIVERSITY LECTURES
    |
  0 +
    | folk stories
    | published speeches
 -1 +
    | analytical press; memos; CONFERENCE DISCUSSIONS
    | sports reviews; CONFERENCE LECTURES
 -2 + political pamphlets; general interest press
    | book introductions; announcements;
    | textbooks; academic prose;
 -3 + invited editorials; editorials
    | press reportage
    |
 -4 +
```

FIGURE 5. Distribution of registers along Dimension 5—Distanced, directive interaction ($F = 22.9$, $p < .0001$, $df = (25,253)$, $r^2 = 69.3\%$).

Figure 4 shows that several other registers mix narrative and nonnarrative discourse organizations. Among the spoken registers, sermons, lectures, and conference discussions all show a relatively high use of narrative forms, whereas sports reviews, invited editorials, and textbooks show moderately high use of narrative forms among the expository written registers. Other registers, such as conversations, formal and family meetings, editorials, and personal letters, are not primarily narrative in purpose, but do make use of narrative features to support their primary goals (whether informative or interpersonal). Finally, three written registers are marked for the near total absence of narrative features: announcements, memos, and petitions. These all have extremely restricted purposes and do not typically use narratives even in supporting roles. Announcements and memos are directly informative, with little elaboration of any kind, whereas petitions are formal requests, which are not enhanced by the inclusion of narratives.

The overall pattern in this dimension represents a cline associated with the extent to which registers depend on narrative discourse organizations. Only folk stories and general fiction make extensive use of these features, but most other Somali registers show at least some use of narrative patterns.

Relations along Dimension 5. On first consideration of Dimension 1, we found it surprising that first and second person pronouns do not group more strongly with the positive involved features such as questions, contractions, and so on. The reason for that pattern is that those pronominal features are two of the strongest loadings in Dimension 5 (see Figure 5). As discussed earlier, the major communicative functions associated with Dimension 1 relate to speaker/writer purpose rather than to interactiveness; Dimension 5, on the other hand, seems to reflect the communicative requirements of certain types of interaction.

Personal letters have by far the largest positive score in Dimension 5. In addition to frequent first and second person pronouns, this score reflects a frequent use of commands (represented by optative clauses, translated as 'let X do Y', and imperatives) and directional preverbal particles (marking action toward or away from the addressor). Family meetings, conversational narratives, conversations, and sermons all have relatively high scores on this dimension. At the other extreme are the expository written registers (e.g., academic prose and political pamphlets), with press reportage and editorials having the lowest scores; these registers are characterized by the absence of Dimension 5 features.

We interpret this clustering of features and distribution of registers to reflect "distanced, directive interaction." The communicative functions associated with these linguistic features are most pronounced in personal letters, where there are frequent references to *I* (the writer) and *you* (the reader[s]) plus frequent directives of various types, reflecting the need to be interactive and directive over great physical distances where there is no possibility of direct, face-to-face interaction.

Other registers, such as conversations and family meetings, are face-to-face, but still directive in many respects, and thus they have relatively high scores here. Fiction and folk stories include dialogue sections that have many of these characteristics. Conversational narratives are in some respects interactive and directive, and they also include reported dialogue with these characteristics. Sermons can be considered a type of distanced interaction: although the addressee cannot ask for clarification, the addressor (the sheikh) makes frequent reference to *I* and *you*, and he is notably directive in exhorting listeners regarding prescribed and proscribed actions. This dimension thus distinguishes among types of interactions that are distanced and directive to varying degrees.

Multidimensional characterizations of registers

Table 6 summarizes the underlying functions, major linguistic features, and distribution of registers for each dimension. Consideration of this table, and the more detailed presentations in Figures 1–5, shows that the dimensions provide fairly complex linguistic characterizations of each spoken or written register. For example, folk stories make the most extensive use of narrative features (Dimension 4), but they are also relatively involved with little structural elaboration (Dimension 1), have relatively little lexical diversity and elaboration (Dimension 2), have a generally reported style (Dimension 3), and have an intermediate score for directive interaction (Dimension 5). This combination of characteristics reflects the fact that these texts usually combine a straightforward narration of events with extensive dialogue.

Another example of a complex multidimensional characterization is press reportage. This register is markedly expository and structurally elaborated in Dimension 1, but it has a more intermediate score in the lexical elaboration dimension (2), reflecting a moderate integration of information through careful word choice. It has by far the lowest score in the argumentative/reported dimension (3), reflecting its direct reportage of information, and also the lowest score in the distanced interaction dimension (5), reflecting a nearly complete absence of direct interaction. Interestingly, press reportage has an intermediate score in the narrative dimension (4); thus, it reports past events (Dimension 3), but makes only moderate use of narrative discourse organizations (Dimension 4).

TABLE 6. *Comparison of functions, linguistic features, and characteristic registers for five Somali dimensions*

Functions	Linguistic Features	Characteristic Registers
DIMENSION 1 *Positive*:		
Interactive (Inter)personal focus Involved Personal stance (On-line production)	main clause features questions imperatives contractions stance adjectives downtoners 1st & 2nd person pronouns	conversations family meetings conversational narratives
Negative:		
Monologue Informational Faceless (Careful production)	dependent clauses relative clauses clefts verb complements nouns adjectives	written expository registers
DIMENSION 2 *Positive*:		
On-line production (Situation dependent)	–	sports broadcast (other spoken registers)
Negative:		
Careful production Informational	once-occurring words high type/token ratio nominalizations compound verbs (see note 11)	editorials written political speeches & pamphlets analytical press
DIMENSION 3 *Positive*:		
Overt argumentation Persuasion	present tense adjectives possibility modals concessive conjuncts conditional clauses	family & formal meetings general interest & analytical press (invited editorials)
Negative:		
Reported presentation	past tense proper & agentive nouns future modals	press reportage (folk stories)
DIMENSION 4 *Positive*:		
Narrative discourse	3rd person pronouns past tense verbs temporal clauses clefts habitual modals	folk stories (serial stories) (general fiction)
Negative:		
Nonnarrative discourse	compound nouns gerunds agentive nouns	petitions announcements memos

continued

TABLE 6 (*continued*)

Functions	Linguistic Features	Characteristic Registers
DIMENSION 5		
Positive:		
Interactive	optative clauses	personal letters
Distanced & directive communication	1st & 2nd person pronouns	(family meetings)
	directional particles	(Quranic exposition)
	imperatives	
Negative:		
Noninteractive	—	press reportage & editorials
Nondirective		written expository registers

Analytical press articles are situationally similar to press reportage, but they differ in purpose and typical topics; these differences correspond to linguistic differences in Dimensions 2, 3, and 5. Analytical press shows a denser integration of information through careful word choice (Dimension 2), it is markedly argumentative in style as opposed to the direct reported style of press reportage (Dimension 3), and it permits some of the features of distanced interaction (Dimension 5), as opposed to their near total absence in press reportage.

Similar characterizations can be given for each of the registers; and a unidimensional description is inadequate for each of them. For example, it is not sufficient to simply describe conversational narratives as on-line production having little lexical elaboration (Dimension 2). Conversational narratives and conference lectures are quite similar in this respect, but they are quite different with respect to the structural elaboration dimension (1); while the lectures are relatively elaborated, the conversational narratives are markedly involved and nonelaborated. Conversational narratives are quite similar to family meetings in Dimensions 1 and 2, with both registers being involved (having little structural elaboration) and produced on-line (having little lexical elaboration). These two registers differ, however, in Dimension 3. Family meetings are extremely argumentative, whereas conversational narratives have an intermediate reported characterization. Similarly for any register, descriptions with respect to a single dimension provide an incomplete characterization, and such characterizations can easily result in inaccurate conclusions concerning the extent of similarities and differences between registers.

Comparing Somali speech and writing

Turning to the comparison of speech and writing in Somali, Table 6 and Figures 1–5 also show that there is a complex set of relations among spoken and

written registers. To aid in this comparison, spoken registers are capitalized, and written registers are underlined in the figures. As in the multidimensional analyses of English, Tuvaluan, and Korean, no dimension in Somali defines an absolute difference between spoken and written registers. Dimensions 1, 2, and 5, though, are closely associated with speech versus writing. Linguistically, all three of these dimensions are defined by some features associated with interactiveness and personal involvement versus structural or lexical complexity and elaboration. Along all three dimensions, conversations (stereotypical speech) are near one pole; expository prose registers (stereotypical writing) are near the other pole; the range of spoken registers tend toward the conversational pole; and the range of written registers tend toward the expository pole. Thus, all three of these dimensions can be considered oral versus literate, in the sense of Biber (1988) and Biber and Finegan (1989).

Dimension 1 shows that there is a limit to the extent to which spoken registers can be structurally elaborated (relative to written registers), even when they have markedly informational purposes (as in lectures and formal meetings). Conversely, written registers are somewhat elaborated structurally (relative to conversational spoken registers), even when they have involved purposes (as in personal letters and folk stories). There is considerable overlap among spoken and written registers in this dimension, however, with the involved written registers having more oral characterizations than the informational spoken registers.

In contrast, Dimension 2 nearly defines an absolute difference between speech and writing; it shows that production differences between the two modes have an extremely strong influence in Somali relative to linguistic features reflecting lexical diversity and elaboration. All spoken registers, regardless of purpose or interactiveness, are markedly nonelaborated in their lexical characteristics, due to their on-line production circumstances. Written registers, on the other hand, occupy a very wide range of variation along this dimension, from the extreme lexical diversity and elaboration seen in registers such as editorials, political published speeches, and political propaganda, to the relatively restricted range of vocabulary found in textbooks and petitions.

Finally, although we characterize Dimension 5 (distanced, directive interaction) as an oral/literate dimension, the most oral register here is personal letters, which has very frequent use of first and second person pronouns and directive forms such as optative clauses and imperatives. Family meetings, the conversational registers, and sermons also have quite oral characterizations in this dimension, although they are less marked than letters. At the other extreme is the range of written expository registers, marked by the absence of these interactive and directive features. Fiction and the informational spoken registers (e.g., lectures) have intermediate characterizations in this dimension. This dimension shows the most overlap among spoken and written registers of the three oral/literate dimensions.

The other two dimensions (3 and 4) relate primarily to discourse organization and purpose, and thus they are largely independent of mode consid-

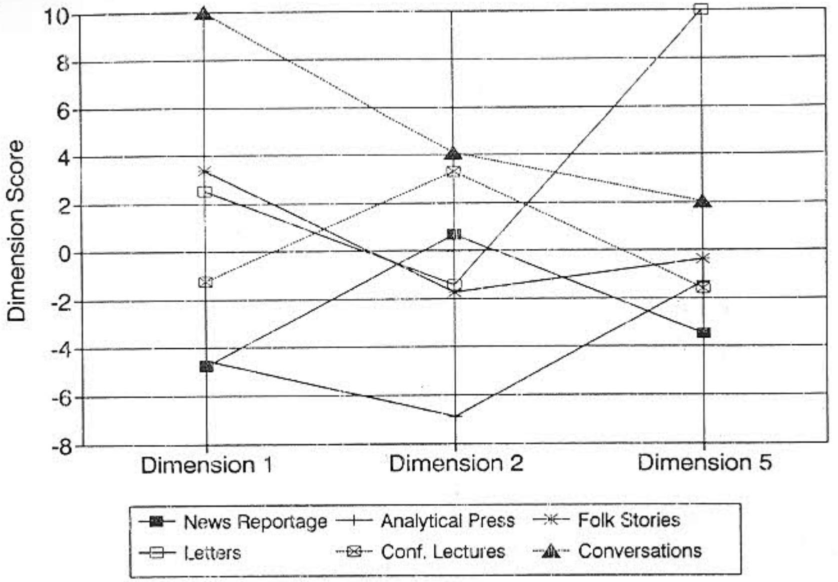

FIGURE 6. Three-dimensional plot of two spoken and three written registers.

erations. Along Dimension 3, which represents argumentative versus reported presentation of information, written press registers are near both poles: general interest press and analytical press near the argumentative pole, and press reportage at the reported pole. Family and formal meetings are the most argumentative, but the other spoken registers (conversational and informational) have intermediate characterizations in this dimension. Similarly, written registers as diverse as fiction, editorials, and press announcements have intermediate characterizations in this dimension. Along Dimension 4, which distinguishes narrative discourse organizations from other discourse types, folk stories and fiction are highly marked (as extremely narrative), whereas most other registers, whether spoken or written, have intermediate characterizations.

Overall, these patterns show that simple unidimensional comparisons between registers (or between speech and writing) are inadequate. Even if we limit the comparison to the three oral/literate dimensions, a multidimensional account is required. For example, Figure 6 plots the relations among two spoken and four written registers with respect to the three oral/literate dimensions (1, 2, and 5). The two spoken registers (conversations and conference lectures) are represented by dashed lines, whereas the four written registers (news reportage, analytical press, folk stories, and personal letters) are represented by solid lines. This figure shows that each of these registers has a distinctive profile with respect to the three oral/literate dimensions. For example, conversations are extremely oral with respect to structural elaboration (Dimension 1) and lexical diversity (Dimension 2), but they have a more intermediate score with respect to directive interaction (Dimension 5). Personal letters have intermediate characterizations in the first two dimensions (structural elaboration and lexical diversity), but are the most oral with respect to directive interaction (Dimension 5). Analytical press and press reportage are both extremely literate with respect to structural elaboration (Dimension 1) and directive interaction (Dimension 5), but with respect to lexical diversity (Dimension 2), analytical press is extremely literate and press reportage has an intermediate characterization.

Thus, even the notion of stereotypical speaking and writing is multidimensional in Somali, as it is in English. That is, depending on the particular purposes, topics, and communicative circumstances, each register will be more or less oral or literate with respect to each dimension. Adequate generalizations concerning the similarities or differences between registers, or concerning the overall orality or literacy of registers, must be based on a comparison of register profiles across dimensions.[13]

CONCLUSION

The present study of Somali further confirms the general conclusions concerning spoken and written language reached in previous multidimensional analyses of register variation in English (Biber, 1986, 1988), Nukulaelae Tuvaluan (Besnier, 1988), and Korean (Kim, 1990). All four studies showed that the linguistic relations among spoken and written registers are quite complex, and that a multidimensional analysis is required, because no single dimension by itself adequately captures the similarities and differences among registers. In addition, all four studies failed to find any absolute dichotomy between speech and writing; rather, situational factors such as purpose, topic, and interactiveness work together with the physical mode distinction to define the salient linguistic differences among registers.

There are also certain specific cross-linguistic generalizations that hold across these four languages. The most notable is that all of these languages have one or more oral/literate dimensions. These do not define absolute differences between speech and writing, but they are associated with stereotypical spoken and written registers, and they are defined linguistically by interactive/involved features versus features reflecting structural/lexical elaboration and complexity. In Nukulaelae Tuvaluan, there are two dimensions that can be considered oral/literate, even though there is an extremely restricted range of register variation, with only two written registers (personal letters and sermons) and six spoken registers (conversations, Council of Elders meetings, Island Council meetings, community and private speeches, and a one-time political discussion on the radio). In Somali, English, and Korean, this similarity is even more striking. The first dimension in all three languages is extremely strong (in terms of the number of defining features); it is defined by involved, interactive features versus features marking structural and lexical elaboration; and it distributes registers along a cline ranging from involvement/interactiveness to informational exposition.

In addition, Somali, English, and Korean all have a narrative dimension, defined by past tense and temporal features, distinguishing fiction and traditional folk stories from other registers. Nukulaelae Tuvaluan lacks a comparable dimension, possibly because there are no written narratives; oral narratives constitute an integral part of many spoken Tuvaluan registers, but they were not analyzed as a separate register (Besnier, 1988:711–714; personal communication).[14]

There are thus important specific cross-linguistic similarities in the patterns of register variation, suggesting the possibility of universal tendencies (perhaps reflecting communicative needs that occur cross-culturally). On the other hand, all four of these languages have dimensions that are unique, reflecting the fact that each culture has certain communicative requirements that are peculiar to its speech and writing situations and purposes. A comparison of similarities *and* differences is required to develop a cross-linguistic account of register variation.

Diachronic comparisons can also facilitate the analysis of cross-linguistic patterns. In Biber and Hared (1992, in press), we used the dimensions of variation described in the present article to analyze the development of Somali written registers from their inception in 1972 to the present. In ongoing research, we are comparing the developmental patterns in Somali with those found by Biber and Finegan (1989) for English written registers.

Additional study of register variation in other languages, which differ in their linguistic classifications, their range of spoken and written registers, and their literacy traditions, will further enhance the cross-linguistic study of register variation. The four languages already analyzed, however, indicate that there are certain strong cross-linguistic similarities in the underlying dimensions of variation and in the relations among spoken and written registers.

NOTES

1. Nukulaelae is a small atoll with a population of c. 300, which is part of the Tuvalu group of atolls and islands. The range of registers on Nukulaelae is much more restricted than it is within Tuvaluan generally; in particular, there are a relatively wide range of spoken and written registers in regular use on Funafuti, Tuvalu's capital.

2. A few of the conversational texts and the family meetings were recorded in Hared's hometown of Beled Weyn, located approximately 205 miles from Muqdisho.

3. Textbooks written in Italian and English are still used at the university level, and academic theses can be written in Somali, Italian, or English.

4. Because spoken texts were considerably more difficult to collect than written texts (at least for some categories), we have extracted two text samples from many spoken texts. For example, a single "conversation" or "lecture" would typically last longer than a recording tape and would thus be much longer than the required 1,000-word sample. In these cases, we transcribed large portions of the interaction and then selected two samples corresponding to topic or participant breaks. The 10 text samples from live sports broadcasts actually come from a single radio broadcast, which was the only sporting event covered during our year in Somalia.

5. Some features could be classified in more than one of these categories.

6. We used a common factor analysis with a Promax rotation. We examined a scree plot together with the solutions for four, five, and six factors and decided to present the six-factor solution as the most adequate. We combined several features (such as adjectival and verbal dependent clauses) before running the factor analysis. In addition, three features were dropped from the analysis because they had low communalities (reason clauses, manner clauses, and *must* clauses). The first factor in the analysis accounted for 27.8% of the shared variance; all six factors together account for 53.3% of the shared variance. Further details are given in Biber (1992).

7. Some features do not have loadings greater than .30 on any factor; the largest loading for these is thus listed (for example, *ahaan* adverbials with a loading of .28 on factor 1). Features with loadings greater than .45 are separated from lesser loadings in Table 4; the features with higher loadings should be given greater weight in the interpretation of each factor, and only those with loadings over .45 are used in the computation of factor scores (see the following section).

8. Hapax legomena and type-token ratio are both based on the first 500 words in a text.

9. These forms also show weaker correlations with certain structural elaboration features: single case particles, demonstrative relatives (relative clauses with a demonstrative pronoun as head, which conclude a series of relative clauses), clitic topic coordination, and purpose clauses.

10. Dimension 4 has weaker positive loadings for stance verbs (e.g., *u malee* 'think', *fil* 'hope'), possessive nouns, and concession clauses. There are also several relatively weak negative loadings: compound nouns, gerunds, agentive nouns, t-unit length, and phrasal coordination.

11. Following the practice in Biber (1988), all frequencies are standardized to a mean of 0.0 and a standard deviation of 1.0 before the dimension scores are computed. In addition, each dimension score was multiplied by a scaling coefficient so that all dimensions used a scale running from +10 to −10. The scaling coefficients are:

Dimension	Scaling Coefficient
1	.314
2	−1.693
3	1.067
4	1.310
5	1.188

Dimension 2 is inverted (reversing the positive and negative poles) to facilitate comparisons across Dimensions 1, 2, and 5; after inversion, conversational registers are at or near the positive pole of all three dimensions, whereas expository registers are at or near the negative pole.

12. The high score for textbooks might also reflect a conscious attempt to increase the comprehensibility of difficult subject matter for high school students by restricting word choice.

13. Although the present analysis shows that there are important linguistic differences among registers in Somali, there has been no attempt to validate the register categories in terms of their linguistic coherence. In fact, because registers are defined situationally (in terms of interactiveness, production circumstances, purpose, etc.) rather than on a linguistic basis, they are not equally coherent in their linguistic characteristics. A complete description of a register should include a linguistic characterization of typical texts (the mean scores) and analysis of the internal range of variation. Some registers, such as petitions and press announcements, have quite focused norms and therefore show little internal linguistic variation. Other registers, such as general press articles, include a wide range of purposes and therefore have an extensive range of linguistic differences among the texts within the register.

A complementary perspective, not explored in the present article, is to analyze the *linguistically* well-defined text categories, or "text types" (see Biber, 1989). Registers and text types represent two alternative approaches to linguistic variation. Registers are defined on the basis of their situations and purposes, but they can be analyzed linguistically; text types are defined on linguistic grounds, but the types can be interpreted functionally. Given a text-type perspective, linguistically distinct texts within a register would be taken to represent different types, whereas linguistically similar texts from different registers would represent a single text type. Biber (1992) identified eight basic text types in Somali and compared the range of types in Somali and English, complementing the comparison of registers given here.

14. The fourth dimension in Besnier (1988) is interesting in this regard in that it shows an opposition between past tense (with a loading of −.33) and nonpast tense (with a loading of .62). However, this dimension was not interpreted because its overall functional basis is not clear.

REFERENCES

Akinnaso, F. Niyi. (1982). On the differences between spoken and written language. *Language and Speech* 25:97–125.

Beaman, Karen. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In Deborah Tannen (ed.), *Coherence in spoken and written discourse*. Norwood, NJ: Ablex. 45–80.

Besnier, Niko. (1988). The linguistic relationships of spoken and written Nukulaelae registers. *Language* 64:707–736.

Biber, Douglas. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language* 62:384–414.

______ (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

______ (1989). A typology of English texts. *Linguistics* 27:3–43.

______ (1992). *Cross-linguistic patterns of register variation: A multi-dimensional comparison of English, Tuvaluan, Korean, and Somali*. Unpublished manuscript.

Biber, Douglas, & Finegan, Edward. (1989). Drift and the evolution of English style: A history of three genres. *Language* 65:487–517.

Biber, Douglas, & Hared, Mohamed. (1992). Literacy in Somali: Linguistic consequences. *Annual Review of Applied Linguistics* 12:260–282.

______ (in press). Linguistic correlates of the transition to literacy in Somali: Language adaptation in six press registers. In Douglas Biber & Edward Finegan (eds.), *Sociolinguistic perspectives on register*. New York: Oxford University Press.

Brown, Penelope, & Fraser, Colin. (1979). Speech as a marker of situation. In Klaus R. Scherer & Howard Giles (eds.), *Social markers in speech*. Cambridge: Cambridge University Press. 33–62.

Chafe, Wallace L. (1982). Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen (ed.), *Spoken and written language: Exploring orality and literacy*. Norwood, NJ: Ablex. 35–53.

Chafe, Wallace L., & Tannen, Deborah. (1987). The relation between written and spoken language. *Annual Review of Anthropology* 16:383–407.

Clancy, Patricia M. (1982). Written and spoken style in Japanese narratives. In Deborah Tannen (ed.), *Spoken and written language: Exploring orality and literacy*. Norwood, NJ: Ablex. 55–76.

Deibler, Ellis W., Jr. (1976). Differences between written and oral styles in languages near Goroka. *Read* 11:77–79.

Duff, Martha. (1973). Contrastive features of oral and written texts in Amuesha. *Notes on Translation* 50:2–13.

Ervin-Tripp, Susan M. (1972). On sociolinguistic rules: Alternation and co-occurrence. In John J. Gumperz & Dell Hymes (eds.), *Directions in sociolinguistics*. New York: Holt, Rinehart, and Winston. 213–250.

Hurd, Conrad. (1979). A study of oral versus written Nasioi discourse. *Read* 14:84–86.

Hymes, Dell. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Irwin, Barry. (1976). Written and oral language in Southern Chimbu. *Read* 11:74–76.

Kim, Yong-Jin. (1990). *Register variation in Korean: A corpus-based study*. Ph.D. Dissertation, University of Southern California.

Li, Charles N., & Thompson, Sandra A. (1982). The gulf between spoken and written language: A case study in Chinese. In Deborah Tannen (ed.), *Spoken and written language: Exploring orality and literacy*. Norwood, NJ: Ablex. 77–88.

Nelson, Harold D. (ed.). (1982). *Somalia: A country study*. Washington, DC: U.S. Government.

Saeed, John Ibrahim. (1984). *The syntax of focus and topic in Somali*. Hamburg: Helmut Buske.

______ (1987). *Somali reference grammar*. Wheaton, MD: Dunwoody.

Tannen, Deborah. (1982a). Oral and literate strategies in spoken and written narratives. *Language* 58:1–21.

______ (ed.). (1982b). *Spoken and written language: Exploring orality and literacy*. Norwood, NJ: Ablex.

APPENDIX

SOMALI TEXT SAMPLE 1: CONVERSATION

(Some young women are discussing whether they had meddled in a relationship between a married couple. Speaker A feels unjustly accused.)

B: *wallaahi dee way iska fiicnayd* (pause)
swear to God uh FM-she just was fine (pause)
suurahay taqaan haye?
coyness-FM-she knew, isn't it

D: *walaal meherkeedii* (unintelligible words)
oh sister legal wedding-her (. . .)

A: *waxaa iigu dambeysayba waa kaas*
What me-for last-time-was FM that [time]
waxay iigu darnayd ayaantay
what-she me-for was the worst [was] day-the-she
Aamina ku tidhi "ninkayga / ninkayga=
Amina to she-said man-my [pause] man-my=

B: *=ninkaygay igu dirayaan"*
=man-my-FM-they me-against they-set

A: *adduunka, kelmaddaasi weli waa xasuustaa, ka*
world-the word-that still FM-I remember about
warran!
report

B: *dee horta waa runoo waan ku dirnee*
uh first-the FM truth-and FM-we to we send-and
ma og tahay? Taasi ma been baa?
QM know being That QM lie FM

A: *kuma dirin*
to-NEG send

B: *Illaahay baan kugu dhaarshee ma been baa?*
God-my FM-I from-with swore-and QM lie FM
laakiin adaa adaa gardarnaa Sacaado.
But you-FM you-FM justice-without Sa'ado
markaan ku idhi ha u sheegin, ha u sheegin,
Time-the-I to said OPT to tell-not OPT to tell-not
ma tidhi saaxiibaan nahay?
QM said friends-we being
ma naag baa iyo ninkeedaa la dhexgalaa?
QM women FM and man-her-FM IMP middle-enter

Translation

B: I swear, she [i.e., the wife] was just fine. She knew how to be coy, didn't she?

D: Sister, her legal-wedding . . .

A: The last time that I saw her was that time. What was the worst thing for me was the day that she said to Amina "my man, my man=

B: =They are setting my man against me."

A: Can you imagine! I still remember that sentence, what do you think of that?

B: Uhm, first of all it is true and we set her against him, don't you know? Is that a lie?

A: We didn't set her against him.

B: I swear to you upon God, is it a lie? But you, you were at fault, Sa'ado. When I said to you, "don't tell [her], don't tell [her]," didn't you say, "we are friends."? Are a woman and her husband meddled with? [i.e., isn't it wrong to meddle with a woman and her husband?]

SOMALI TEXT SAMPLE 2: PRESS EDITORIAL

Dhinaca keenista daawada iyo qaybinteedaba,
the side (of) the-bringing (of) the medicine and even their distribution
dhibaatooyinka ka jira waxaa ka mid ah daawo
the problems (which) at exist what from one is medicine (which)
boorso lagu sido oo aan la ogeyn waxa ay ka
purse 'they'-with carry and (which) not 'they' know what it from
sameysan tahay, cidda sameysay iyo waxa ay tarto toona,
made is people-the (who) made (it) and what it does neither
oo si xaaraan ah dalka u soo galeysa, taasoo
and (that) way forbidden being country-the into enters that one (which)
aynnu ognahay khasaarooyinka wax-yeellada caafimaad leh
we know losses-the (which) damages-the health have
ee bulshadeenna ka soo gaara. Qorsheyntii
and (which) society-our from toward reaches Planning-the (of)
qaybinta daawooyinka dalkana waxa ku
distributing-the (of) medicine-the (of) country-the-and what with
guuleystay geddisley yaryar ah oo ku sugan waxa
succeeded (was) traders (which) small are and (which) in are what
loogu yeero farmashiyayaasha, guul-darradaas iyo
'they'-in call pharmacies victory-without-that and
marin-habaabintaas waxa dhaliyay ka-gaabinta
path-misleading-that what caused (was) from-becoming-short-the (of)
xil-gudashadii looga baahnaa Wakaaladda
responsibility-fulfilling-the (which) 'they'-from-to needed agency-the (of)
ASPIMA.
ASPIMA

Translation

'With respect to importing medicine and distributing it, the problems that exist include medicine carried in a purse [i.e., sold on the blackmarket], which it is not known what it is made of, the people who made it, or what it does, and which enters the country in a forbidden way, that one [i.e., the medicine] that we know the losses that include the health damages that come from it to our society. And what has succeeded in planning the distribution of the medicine of the country is the small traders who are in the so-called pharmacies; what caused that failure and that deception was the lack of fulfillment of responsibility that was required from the agency of ASPIMA.'

KEY TO ABBREVIATIONS USED IN THE ENGLISH INTERLINEAR GLOSSES

FM	focus marker
QM	question marker
OPT	optative marker
IMP	impersonal marker
NEG	negative marker