SUMMARY

Although many frequency dictionaries of French words have been published over the past decades, they rarely include fixed expressions and idioms. I therefore view it as an important task to try to compile a frequency dictionary of French *locutions verbales* (fixed expressions of the predicate type). A pilot study of the frequency of about twenty expressions in the newspaper *Le Monde* 1996 and 1997 reveals that particular problems arise in creating such a dictionary. This paper discusses some of these problems, including variation (instability) and ambiguity, and concludes with a model entry for the dictionary.

Reçu le 21 janvier 1999

Linguisticæ Investigationes XXI:2. 379–400 (1997).

A COMPUTER CORPUS-BASED STUDY OF SUBJECT RAISING IN MODERN PORTUGUESE

MARK DAVIES
Illinois State University

1. Introduction

1.1. A common syntactic construction in Portuguese and many other languages is "subject to subject raising", in which the subject of the embedded clause (1a) moves into the subject position of the main clause (1b):

(1) a. __ *parece* [*que* ***a Maria*** *está doente*] non-raised
It seems that Mary is sick
b. ***A Maria*** *parece* [__ *estar doente*] raised
Mary seems to be sick

This construction has been of great interest to syntacticians working in different theoretical frameworks because of the insights that it gives into the basic nature of clauses and movement between clauses, such as the movement of the subject from the embedded clause into the subject position of the main clause.

While the construction has been studied a great deal for languages like English and certain other Romance languages such as Spanish (cf. Bolinger 1972, García Pinto and Luján 1974, Fernández Leboráns and Díaz Bautista 1992, Davies 1997a, and Davies 1997b), it has received scant attention for Portuguese. In a work devoted to a general theory of clausal structure, Quicoli (1982) discusses different models for analyzing the construction, but provides very little actual data. Besides this, there are occasional isolated comments in passing in the standard descriptive grammars of the language, but no in-depth discussion (e.g. Dunn 1928: 431, Pereira 1946: 345, Bueno 1951: 361–2, Almeida 1963: 465; Said Ali 1964: 178, Torres 1967: 119, Barros 1985: 250).

In order to extract the 4500+ cases of the *parecer* construction, we first needed to have all of the different corpora in electronic form. To acquire the texts, the Modern Portuguese newspapers (including the interviews) were downloaded from the World Wide Web, and three of the four *Linguagem Falada* corpora were scanned into the computer (the fourth was received from another researcher in electronic form). We then used *WordCruncher* to create an every-word index of the corpus, which permitted us to perform complex proximity and Boolean searches on the data. Proximity searching meant that we could find every case of *parecer* immediately followed by either an infinitive or *que*. Boolean searching meant that we could combine and limit searches, such as all cases of *parecer* followed by an infinitive where the infinitive was not *ser*. Needless to say, without the computer corpus and *WordCruncher*, extracting the 4500+ tokens from the 26,500,000+ word corpus would have been virtually impossible.

2. Basic contrast in raised / non-raised constructions

2.1. Perhaps the most basic question has to do with the frequency of the raised construction (6a) and the non-raised construction (6b) in the different dialects and registers of Modern Portuguese.

(6) a. [–raising] *Ela ... **parece que sabe** das coisas*
(Comércio 97.07.11)

b. [+raising] *E Gal **parece saber** disso* (Globo 97.09.29)

This basic contrast is noted in the standard prescriptive grammars (e.g. Dunn 1928: 431, Pereira 1946: 345, Bueno 1951: 361–2, Almeida 1963: 465; Said Ali 1964: 178, Torres 1967: 119, Barros 1985: 250), but is never quantified. Our study does examine the difference between the two constructions, and the results for both the written and spoken registers of both European and Brazilian Portuguese are summarized in Table 2.

There are two different figures for the percentage of subject raising. Column C simply indicates the total percentage of all cases that involve raising rather than the non-raised construction (e.g. *parece que...*). A substantial number of these cases of raising, however, involve the semantically simple verb *ser* (7a), in which the raising construction is rather similar to the more basic attributive use of *parecer* involving an adjective (7b):

Table 2. Raising / non-raising in Modern Portuguese

	A	B	C	D	E	F
	-raising	+raising	% raising	*ser*	% of raising with *ser*	% rai (w/o
Portugal-Written	596	1920	0.76	374	0.19	0.(
Portugal-Interviews	56	48	0.46	16	0.33	0.3
Brazil-Written	540	1010	0.65	212	0.21	0.5
Brazil-Interviews	99	83	0.46	20	0.24	0.3
Brazil-*Linguagem Falada*	188	4	0.02	3	0.75	0.(
TOTAL	1479	3065				

(7) a. *A liberdade total não **parece possível** ...* (JN 97.0
b. *O exercício formal da democracia **parecia ser** possível*
(Público 97.0

Column F provides the figures for the percentage of all constructions have raising, excluding those cases with *ser* as the embedded verb.

As the table shows, raising is quite common in the written languag the level of 65% in Brazil and 76% of all cases in Portugal. If we exc the cases of raising with *ser*, then the percentage of raising is about 51 Written Brazilian Portuguese and about 61% in Written European P guese. In summary, these data show that at least in written Portug subject raising is not an uncommon phenomenon.

The data that are much harder to explain, however, are the data the spoken corpora. The 1,000,000+ word corpus of spoken Portuguese t from the newspapers from Portugal and Brazil show a much lower figu only 31% raising for Portugal and 35% for Brazil. Perhaps the most diffi data to explain are those from the *Linguagem Falada* corpus, in which 1% of all cases involve raising (just 4/192 tokens). Furthermore, notice three of these four examples of raising are with the semantically simple *ser*, and that the one single case of raising with another verb is with which is also semantically quite simple:

(8) a. ***Parece ser** um homem de muito valor* (Rio de Janeiro 93)
b. *Parecem ter duas cabeças* (Recife 216)

SUMMARY

Although many frequency dictionaries of French words have been published over the past decades, they rarely include fixed expressions and idioms. I therefore view it as an important task to try to compile a frequency dictionary of French *locutions verbales* (fixed expressions of the predicate type). A pilot study of the frequency of about twenty expressions in the newspaper *Le Monde* 1996 and 1997 reveals that particular problems arise in creating such a dictionary. This paper discusses some of these problems, including variation (instability) and ambiguity, and concludes with a model entry for the dictionary.

Reçu le 21 janvier 1999

Linguisticæ Investigationes XXI:2. 379–400 (1997).

A COMPUTER CORPUS-BASED STUDY OF SUBJECT RAISING IN MODERN PORTUGUESE

MARK DAVIES
Illinois State University

1. Introduction

1.1. A common syntactic construction in Portuguese and many othe languages is "subject to subject raising", in which the subject of the embec ded clause (1a) moves into the subject position of the main clause (1b):

(1) a. __ *parece* [*que* ***a Maria*** *está doente*] non-raise
It seems that Mary is sick
b. ***A Maria*** *parece* [__ *estar doente*] raise
Mary seems to be sick

This construction has been of great interest to syntacticians working i different theoretical frameworks because of the insights that it gives into th basic nature of clauses and movement between clauses, such as the move ment of the subject from the embedded clause into the subject position of th main clause.

While the construction has been studied a great deal for languages lik English and certain other Romance languages such as Spanish (cf. Bolinger 197 García Pinto and Luján 1974, Fernández Leboráns and Díaz Bautista 199 Davies 1997a, and Davies 1997b), it has received scant attention for Portugues In a work devoted to a general theory of clausal structure, Quicoli (198 discusses different models for analyzing the construction, but provides very litt actual data. Besides this, there are occasional isolated comments in passir in the standard descriptive grammars of the language, but no in-dept discussion (e.g. Dunn 1928: 431, Pereira 1946: 345, Bueno 1951: 361–: Almeida 1963: 465; Said Ali 1964: 178, Torres 1967: 119, Barros 1985: 250

This neglect of subject raising in Portuguese is unfortunate, because without the Portuguese data it is difficult to construct a pan-Romance description of subject raising and to see what similarities exist across all of the languages and what are the fundamental points of contrast. For example, some of the Romance languages have an indirect object "experiencer" with subject raising ((2a) for Spanish), whereas this is common only in certain dialects and registers of Portuguese (2b):

(2) a. ***Me parece ver*** *a la viejita con su vestido rojo*
(Bogotá, M44:617)

b. ***Parece-me ver*** *algum ódio* (Expresso 97.10.18)

Conversely, certain dialects and registers of Portuguese exhibit characteristics that are not present in other languages (like Spanish), such as the possibility of verbal agreement in the embedded clause, but not the main clause, as in (3):

(3) *E* ***os políticos*** *bem parece* ***saberem*** *que ...* (JN 96.10.03)
The politicians seem to know well that ...

In addition to the differences between Portuguese and other Romance languages, more complete data from Portuguese would allow us to determine whether certain unexpected phenomena in the subject raising constructions of other languages are particular to those languages, or whether more general syntactic processes are involved. For example, Davies (1997b: 37–43) shows that subject raising is much more common in written Spanish than it is in spoken Spanish. This study also shows that while raising is common with third person subjects (4a), it is still quite uncommon and awkward with first and second person (4b) (Davies 1997b: 43–48):

(4) a. *La vida* ***parecía brindar*** *al poeta todas sus oportunidades*
(Havana M37: 677)

b. *Encogiendo las piernas,* ***parezco defecar*** *en el piso*
(Barro 90: 2)

Finally, the Modern Spanish data shows the existence of a curious "partial raising" construction in which both the main and embedded clause verbs are finite (Davies 1997b: 48–51). Interestingly though, in spite of examples of this construction in the corpus, nearly all speakers regard it as either very awkward or unacceptable:

(5) *¿Qué problemas te* ***parecen*** *que* ***son*** *los más importantes?*
(Sevilla, PM7:17

In summary, comprehensive data from Portuguese will enable us to examine important differences in the subject raising construction betwe Portuguese and some of the other Romance languages 2) consider syntac divergences between different dialects and registers of Portuguese, and indicate whether the phenomena in a related language such as Spanish anomalous, or whether they represent common tendencies across a rela group of languages.

1.2. In order to consider the types of questions just posed, however, obviously need a large corpus of tokens of the subject raising constructi from Modern Portuguese. In this study, we will rely on 4500+ tokens fr a corpus involving more than 26,500,000 words, which are summarized Table 1, and which are described more fully in Appendix 1.

Table 1. Composition of the corpus

	#texts/conversations	# words	# tokens (±raising
Portugal-Written	16,775	10,000,000	2,516
Portugal-Interviews	292	389,700	104
Brazil-Written	46,392	15,000,000	1,550
Brazil-Interviews	628	625,300	182
Brazil-*Linguagem Falada*	85	570,800	192
Total	64,200	26,585,800	4,544

As the table indicates, the 26,500,000+ words of text are from more t 63,000 articles in twelve newspapers from Brazil and eight newspapers fr Portugal newspaper articles that were downloaded via the Internet. addition, there are more than 1,000,000 words of text of spoken Portugu taken from 900+ interviews found in these same newspapers. Finally, th are another 570,000 words of spoken Brazilian Portuguese from the *Ling gem Falada* project in the cities of São Paulo, Rio de Janeiro, Bahia, Recife. In summary, these represent some of the largest and most div corpora available for Modern Portuguese, and provide a good samplin the different registers and dialects.

Why is it that raising is so much more common in written Portuguese than in spoken Portuguese, especially the very informal language of the *Linguagem Falada* corpora?

Our suggestion is that the difference has to do with general factors relating to the production of spoken and written language. These factors have been the focus of corpus-based research in a number of other languages. Miller (1994: 4301), for example, reminds us that one obvious difference between the two registers is that spoken language is produced in real time, whereas written language has more careful editing, which permits the writer to develop more complex syntax. In more specific terms, Biber (1988: 47) notes that infinitives are used to achieve a higher degree of clausal integration (see also Beaman 1984 and Chafe 1985). In summary, we see a cross-linguistic preference of written texts for a higher degree of clausal integration, such as with the infintival clauses found in the raising sentence *a Maria está doente.*

Shorter, more defined clauses are characteristic of spoken language. These clauses are generally introduced with conjunctions that separate the subordinate from the main clause. In studies carried out on extensive corpora of spoken English, Biber (1988: 159, 195) notes that 'the primary use of *that* complements ... seems to be for informational elaboration under real time production constraints' and 'in discourse that cannot be carefully planned and integrated'. This fits in with a more general 'one-clause-at-a-time' constraint that has been hypothesized for spoken language (Sacks, Schegloff, and Jefferson 1974, Pawley and Snyder 1983), and corresponds in turn to the non-raising *(parece) que a Maria está doente* type sentences. In summary, the difference in the degree of subject raising in the written and spoken registers, which might seem a bit unexpected at first sight, can be explained quite nicely by applying more general principles of language production for these two registers.

3. Raising with non-3SG subjects: *eles parecem estar doentes*

Let us now consider a certain subset of raising, involving constructions that involve subjects other than third person singular subjects (non-3SG). As Bolinger points out (1972: 73–74), Modern Spanish and English differ in the

sense that in English, raising is common irrespective of the person
number of the subject (9a), whereas at least non-third person subjects
rather awkward in Spanish (9b). The obvious question, then, is whet
Portuguese freely allows raising, regardless of the nature of the subject (
English), or whether it is more restricted (as in Spanish):

(9) a. *John / you / I seem(s) to have offended Mary*
b. *Juan parece / (?) tú pareces / (??) yo parezco haber ofendia María*

This issue of what type of subjects most easily allow subject raising is
that is not discussed directly in any of the prescriptive grammar (e.g. D
1928: 431, Pereira 1946: 345, Bueno 1951: 361–2, Almeida 1963: 465; S
Ali 1964: 178, Torres 1967: 119, Barros 1985: 250). The data from
corpus, on the other hand, shows that in Portuguese raising is sensitive to
nature of the subject, at least for first and second person subjects.

It appears, however, that there is little difference between third per
singular and plural subjects — the 3PL subjects raise as easily as the 3SG
the corpus there are more than 500 tokens of 3PL in the written articles
the Portuguese and Brazilian newspapers (10a), and 27 tokens in the in
views from these newspapers (10b):

(10) a. *Os ladrões* ***pareciam saber*** *do funcionamento do local*
(Nordeste 97.07.
b. *Os franceses* ***parecem preferir*** *o romance psicológico*
(ESP 97.10.

Although there seems to be little difference between 3SG and 3PL subje
the same is not true for first and second person subjects, which are quite r
in Modern Portuguese. There are no examples in the spoken corpora, and th
are just five tokens in the two written corpora of more than 25,000,000 wor

(11) a. *Acho que não* ***pareço ter*** *93 anos* (Pernambuco 97.10.
b. *que* ***parecemos*** *nunca* ***saber*** *quantas coisas existem*
(Gazeta do Povo 1997.03.
c. ... *acreditando no que* ***parecemos ser*** ... (Tarde 97.03.
d. *A partir daí, bases* ***parecíamos ter*** (JN 97.06.
e. ***Parecemos estar*** *em Myst, com Myst, por Myst forever*
(Expresso 97.12.

In order to extract the 4500+ cases of the *parecer* construction, we first needed to have all of the different corpora in electronic form. To acquire the texts, the Modern Portuguese newspapers (including the interviews) were downloaded from the World Wide Web, and three of the four *Linguagem Falada* corpora were scanned into the computer (the fourth was received from another researcher in electronic form). We then used *WordCruncher* to create an every-word index of the corpus, which permitted us to perform complex proximity and Boolean searches on the data. Proximity searching meant that we could find every case of *parecer* immediately followed by either an infinitive or *que*. Boolean searching meant that we could combine and limit searches, such as all cases of *parecer* followed by an infinitive where the infinitive was not *ser*. Needless to say, without the computer corpus and *WordCruncher*, extracting the 4500+ tokens from the 26,500,000+ word corpus would have been virtually impossible.

2. Basic contrast in raised / non-raised constructions

2.1. Perhaps the most basic question has to do with the frequency of the raised construction (6a) and the non-raised construction (6b) in the different dialects and registers of Modern Portuguese.

(6) a. [–raising] *Ela ...* ***parece que sabe*** *das coisas* (Comércio 97.07.11)

b. [+raising] *E Gal* ***parece saber*** *disso* (Globo 97.09.29)

This basic contrast is noted in the standard prescriptive grammars (e.g. Dunn 1928: 431, Pereira 1946: 345, Bueno 1951: 361–2, Almeida 1963: 465; Said Ali 1964: 178, Torres 1967: 119, Barros 1985: 250), but is never quantified. Our study does examine the difference between the two constructions, and the results for both the written and spoken registers of both European and Brazilian Portuguese are summarized in Table 2.

There are two different figures for the percentage of subject raising. Column C simply indicates the total percentage of all cases that involve raising rather than the non-raised construction (e.g. *parece que...*). A substantial number of these cases of raising, however, involve the semantically simple verb *ser* (7a), in which the raising construction is rather similar to the more basic attributive use of *parecer* involving an adjective (7b):

Table 2. Raising / non-raising in Modern Portuguese

	A	B	C	D	E	F
	-raising	+raising	% raising	*ser*	% of raising with *ser*	% raisi (w/o *s*
Portugal-Written	596	1920	0.76	374	0.19	0.61
Portugal-Interviews	56	48	0.46	16	0.33	0.31
Brazil-Written	540	1010	0.65	212	0.21	0.51
Brazil-Interviews	99	83	0.46	20	0.24	0.35
Brazil-*Linguagem Falada*	188	4	0.02	3	0.75	0.01
TOTAL	1479	3065				

(7) a. *A liberdade total não* ***parece possível*** *...* (JN 97.09.

b. *O exercício formal da democracia* ***parecia ser*** *possível* (Público 97.09.

Column F provides the figures for the percentage of all constructions t have raising, excluding those cases with *ser* as the embedded verb.

As the table shows, raising is quite common in the written language the level of 65% in Brazil and 76% of all cases in Portugal. If we excl the cases of raising with *ser*, then the percentage of raising is about 51% Written Brazilian Portuguese and about 61% in Written European Po guese. In summary, these data show that at least in written Portugu subject raising is not an uncommon phenomenon.

The data that are much harder to explain, however, are the data fr the spoken corpora. The 1,000,000+ word corpus of spoken Portuguese ta from the newspapers from Portugal and Brazil show a much lower figur only 31% raising for Portugal and 35% for Brazil. Perhaps the most diffi data to explain are those from the *Linguagem Falada* corpus, in which c 1% of all cases involve raising (just 4/192 tokens). Furthermore, notice three of these four examples of raising are with the semantically simple v *ser*, and that the one single case of raising with another verb is with which is also semantically quite simple:

(8) a. ***Parece ser*** *um homem de muito valor* (Rio de Janeiro 93)

b. *Parecem ter duas cabeças* (Recife 216)

This neglect of subject raising in Portuguese is unfortunate, because without the Portuguese data it is difficult to construct a pan-Romance description of subject raising and to see what similarities exist across all of the languages and what are the fundamental points of contrast. For example, some of the Romance languages have an indirect object "experiencer" with subject raising ((2a) for Spanish), whereas this is common only in certain dialects and registers of Portuguese (2b):

(2) a. ***Me parece ver*** *a la viejita con su vestido rojo*
(Bogotá, M44:617)

b. ***Parece-me ver*** *algum ódio* (Expresso 97.10.18)

Conversely, certain dialects and registers of Portuguese exhibit characteristics that are not present in other languages (like Spanish), such as the possibility of verbal agreement in the embedded clause, but not the main clause, as in (3):

(3) *E* ***os políticos*** *bem parece* ***saberem*** *que ...* (JN 96.10.03)
The politicians seem to know well that ...

In addition to the differences between Portuguese and other Romance languages, more complete data from Portuguese would allow us to determine whether certain unexpected phenomena in the subject raising constructions of other languages are particular to those languages, or whether more general syntactic processes are involved. For example, Davies (1997b: 37–43) shows that subject raising is much more common in written Spanish than it is in spoken Spanish. This study also shows that while raising is common with third person subjects (4a), it is still quite uncommon and awkward with first and second person (4b) (Davies 1997b: 43–48):

(4) a. *La vida* ***parecía brindar*** *al poeta todas sus oportunidades*
(Havana M37: 677)

b. *Encogiendo las piernas,* ***parezco defecar*** *en el piso*
(Barro 90: 2)

Finally, the Modern Spanish data shows the existence of a curious "partial raising" construction in which both the main and embedded clause verbs are finite (Davies 1997b: 48–51). Interestingly though, in spite of examples of this construction in the corpus, nearly all speakers regard it as either very awkward or unacceptable:

(5) *¿Qué problemas te* ***parecen*** *que* ***son*** *los más importantes?*
(Sevilla, PM7:1

In summary, comprehensive data from Portuguese will enable us to examine important differences in the subject raising construction betw Portuguese and some of the other Romance languages 2) consider syntac divergences between different dialects and registers of Portuguese, and indicate whether the phenomena in a related language such as Spanish anomalous, or whether they represent common tendencies across a rela group of languages.

1.2. In order to consider the types of questions just posed, however, obviously need a large corpus of tokens of the subject raising construct from Modern Portuguese. In this study, we will rely on 4500+ tokens fr a corpus involving more than 26,500,000 words, which are summarizec Table 1, and which are described more fully in Appendix 1.

Table 1. Composition of the corpus

	#texts/conversations	# words	# tokens (±raisin
Portugal-Written	16,775	10,000,000	2,516
Portugal-Interviews	292	389,700	104
Brazil-Written	46,392	15,000,000	1,550
Brazil-Interviews	628	625,300	182
Brazil-*Linguagem Falada*	85	570,800	192
Total	64,200	26,585,800	4,544

As the table indicates, the 26,500,000+ words of text are from more t 63,000 articles in twelve newspapers from Brazil and eight newspapers f Portugal newspaper articles that were downloaded via the Internet addition, there are more than 1,000,000 words of text of spoken Portugu taken from 900+ interviews found in these same newspapers. Finally, tl are another 570,000 words of spoken Brazilian Portuguese from the *Lin gem Falada* project in the cities of São Paulo, Rio de Janeiro, Bahia, Recife. In summary, these represent some of the largest and most div corpora available for Modern Portuguese, and provide a good samplin the different registers and dialects.