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

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):

\[
\begin{align*}
\text{(1) a. } & \text{ _ parece [que a Maria está doente] } & \text{ non-raised} \\
& \text{ It seems that Mary is sick} \\
\text{ b. } & \text{ A Maria parece [_ estar doente] } & \text{ raised} \\
& \text{ Mary seems to be sick}
\end{align*}
\]

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, Quiroga (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>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portugal-Written</td>
<td>596</td>
<td>1920</td>
<td>0.76</td>
<td>374</td>
<td>0.19</td>
<td>0.6</td>
</tr>
<tr>
<td>Portugal-Interviews</td>
<td>56</td>
<td>48</td>
<td>0.46</td>
<td>16</td>
<td>0.33</td>
<td>0.2</td>
</tr>
<tr>
<td>Brazil-Written</td>
<td>540</td>
<td>1010</td>
<td>0.65</td>
<td>212</td>
<td>0.21</td>
<td>0.5</td>
</tr>
<tr>
<td>Brazil-Interviews</td>
<td>99</td>
<td>83</td>
<td>0.46</td>
<td>20</td>
<td>0.24</td>
<td>0.5</td>
</tr>
<tr>
<td>Brazil-Linguagem Falada</td>
<td>188</td>
<td>4</td>
<td>0.02</td>
<td>3</td>
<td>0.75</td>
<td>0.0</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1479</td>
<td>3065</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(7) a. *A liberdade total não parece possível ...* (IN 97.08)  
     b. *O exercício formal da democracia parece ser possível*  
                (Público 97.09)

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

The data that are much harder to explain, however, are the data from the spoken corpora. The 1,000,000+ word corpus of spoken Portuguese from the newspapers from Portugal and Brazil show a much lower figure only 31% raising for Portugal and 35% for Brazil. Perhaps the most difficult data to explain are those from the *Linguagem Falada* corpus, in which 1% of all cases involve raising (just 4/192 tokens). Furthermore, notice that 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

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):

\[ \begin{align*}
(1) & \quad \text{a.} \quad \underline{\text{parece}} \quad [\text{que} \quad \underline{\text{a Maria está doente}}] & \text{non-raise} \\
& \quad \text{It seems that Mary is sick} \\
& \quad \text{b.} \quad \underline{\text{Maria parece}} \quad [\underline{\text{estar doente}}] & \text{raise} \\
& \quad \text{Mary seems to be sick}
\end{align*} \]

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 1977; García Pinto and Luján 1974, Fernández Leboráns and Díaz Bautista 1993; Davies 1997a, and Davies 1997b), it has received scant attention for Portuguese. In a work devoted to a general theory of clausal structure, Quicoli (1987) discusses different models for analyzing the construction, but provides very little actual data. Besides this, there are occasional isolated comments in passim in the standard descriptive grammars of the language, but no in-depth discussion (e.g. Dunn 1928: 431, Percival 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. Encogiéndose las piernas, parece debo calzar 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 between Portuguese and some of the other Romance languages 2) consider syntactic divergences between different dialects and registers of Portuguese, and indicate whether the phenomena in a related language such as Spanish is anomalous, or whether they represent common tendencies across a related 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 from a corpus involving more than 26,500,000 words, which are summarized in Table 1, and which are described more fully in Appendix 1.

<table>
<thead>
<tr>
<th></th>
<th>#texts/conversations</th>
<th># words</th>
<th># tokens (raising)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portugal-Written</td>
<td>16,775</td>
<td>10,000,000</td>
<td>2,516</td>
</tr>
<tr>
<td>Portugal-Interviews</td>
<td>292</td>
<td>389,700</td>
<td>104</td>
</tr>
<tr>
<td>Brazil-Written</td>
<td>46,392</td>
<td>15,000,000</td>
<td>1,550</td>
</tr>
<tr>
<td>Brazil-Interviews</td>
<td>628</td>
<td>625,300</td>
<td>182</td>
</tr>
<tr>
<td>Brazil-Linguaagem Falada</td>
<td>85</td>
<td>570,800</td>
<td>192</td>
</tr>
<tr>
<td>Total</td>
<td>64,200</td>
<td>26,585,800</td>
<td>4,544</td>
</tr>
</tbody>
</table>

As the table indicates, the 26,500,000+ words of text are from more than 63,000 articles in twelve newspapers from Brazil and eight newspapers from Portugal newspaper articles that were downloaded via the Internet. In addition, there are more than 1,000,000 words of text of spoken Portuguese taken from 900+ interviews found in these same newspapers. Finally, there are another 570,000 words of spoken Brazilian Portuguese from the Linguaagem Falada project in the cities of Sào Paulo, Rio de Janeiro, Bahia, and Recife. In summary, these represent some of the largest and most diverse corpora available for Modern Portuguese, and provide a good sampling of 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 infinitive 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 (párecer) *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 whether Portuguese freely allows raising, regardless of the nature of the subject (1 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 ofendido Maria*

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. Du 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 person 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 interviews 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 subject 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 there are just five tokens in the two written corpora of more than 25,000,000 w ords.

(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* (IN 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)  
ad(raising) *Ela ... parece que sabe das coisas*  
(Comércio 97.07.11)  
b. [raising] *Ele disse que parecia*  
(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>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>-raising</td>
<td>+raising</td>
<td>% raising</td>
<td>% with ser</td>
<td>% raising</td>
<td>% raising</td>
</tr>
<tr>
<td>Portugal-Written</td>
<td>596</td>
<td>1920</td>
<td>0.76</td>
<td>0.19</td>
<td>0.61</td>
</tr>
<tr>
<td>Portugal-Interviews</td>
<td>56</td>
<td>48</td>
<td>0.46</td>
<td>0.33</td>
<td>0.31</td>
</tr>
<tr>
<td>Brazil-Written</td>
<td>540</td>
<td>1010</td>
<td>0.65</td>
<td>0.21</td>
<td>0.31</td>
</tr>
<tr>
<td>Brazil-Interviews</td>
<td>99</td>
<td>83</td>
<td>0.46</td>
<td>0.24</td>
<td>0.25</td>
</tr>
<tr>
<td>Brazil-Linguagem Falada</td>
<td>188</td>
<td>4</td>
<td>0.02</td>
<td>0.75</td>
<td>0.01</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1479</td>
<td>3065</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Column F provides the figures for the percentage of all constructions that 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 Portuguese subject raising is not an uncommon phenomenon.

The data that are much harder to explain, however, are the data for the spoken corpora. The 1,000,000+ word corpus of spoken Portuguese text from the newspapers from Portugal and Brazil show a much lower figure only 31% raising for Portugal and 35% for Brazil. Perhaps the most diff 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 * 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. Encogiéndole las piernas, parezca defecer 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:

In summary, comprehensive data from Portuguese will enable us to examine important differences in the subject raising construction between Portuguese and some of the other Romance languages 2) consider syntactic 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 from a corpus involving more than 26,500,000 words, which are summarized in Table 1, and which are described more fully in Appendix 1.

<table>
<thead>
<tr>
<th></th>
<th>#texts/conversations</th>
<th># words</th>
<th># tokens (3 raising)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portugal-Written</td>
<td>16,775</td>
<td>10,000,000</td>
<td>2,516</td>
</tr>
<tr>
<td>Portugal-Interviews</td>
<td>292</td>
<td>389,700</td>
<td>104</td>
</tr>
<tr>
<td>Brazil-Written</td>
<td>46,392</td>
<td>15,000,000</td>
<td>1,550</td>
</tr>
<tr>
<td>Brazil-Interviews</td>
<td>628</td>
<td>625,300</td>
<td>182</td>
</tr>
<tr>
<td>Brazil-LinguaM Falada</td>
<td>85</td>
<td>570,800</td>
<td>192</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>64,200</strong></td>
<td><strong>26,585,800</strong></td>
<td><strong>4,544</strong></td>
</tr>
</tbody>
</table>

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