Corpus

23 | 2022
Corpus et données en morphologie

Description and analysis of a Portuguese blend corpus

Description et analyse d'un corpus de mots-valises portugais

Alina Villalva and Rafael Dias Minussi

Electronic version
URL: https://journals.openedition.org/corpus/6436
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
Alina Villalva and Rafael Dias Minussi, "Description and analysis of a Portuguese blend corpus", *Corpus* [Online], 23 | 2022, Online since 25 January 2022, connection on 02 March 2022. URL: http://journals.openedition.org/corpus/6436

Description and analysis of a Portuguese blend corpus

Description et analyse d'un corpus de mots-valises portugais

Alina Villalva and Rafael Dias Minussi

Introduction

Multi-word formation processes in Portuguese comprise root-compounding (cf. 1a), word-compounding (cf. 1b), and blending (cf. 1c).

(1)
a. *toxicodependente* 'drug addict'
agridoce 'sour sweet'
b. *barco-casa* 'houseboat'
guarda-roupa 'wardrobe'
cantora-atriz 'singer / actress'
c. *cartomente* 'lying fortune teller'
< *cartomante* 'fortune teller' + *mente* 'he/she lies'
tristemunho 'sad testimony'
< *triste* 'sad' + *testemunho* 'testimony'
cantautor 'singer and composer'
< *cantor* 'singer' + *autor* 'composer'

Portuguese compound structures are thoroughly described by several authors (e.g., Villalva & Gonçalves 2015), whereas blending has yet only garnered some controversial and even contradictory analyses: some authors claim that blends and compounds have similar structures, while others consider that they have completely different structures (cf., Gries 2004b, Minussi & Nóbrega 2014, Beliaeva 2019 and Renner 2022). Therefore, blending is still a challenge for morphological analysis.

In this paper, we aim to bring a contribution to this discussion, based on the analysis of a considerable number of Portuguese blends (cf. section 1). The analysis of the data led to the compilation of a corpus, the Portuguese Blend Corpus (henceforth PBC), that will be shown in section 2. The following section is devoted to the presentation of an analysis of blending that the PBC renders possible.

The final section brings the discussion of two experimental tests related to the interpretation and processing of blends. These tests were performed for two main reasons. The first one is related to the wish to check theoretical findings against some external evidence. The second one is rooted in the belief that analyzing the processing of blends may help to deepen the knowledge on word processing in general.
Eventually, we will claim that blends have unique features that place them as a singular category within the field of complex words, and that the analysis of blend processing reveals that the presence of a full word within a blend has a negative impact for its decoding.

1. Portuguese blends

This section is devoted to the identification of features that need to be considered for the description of a large set of blends, such as those that form the PBC (presented in section 2). The features that we will discuss are related to the lexical status of the blend constituents (1.1), to the grammatical role of the constituents within the blend (1.2), and to the prosodic relation that holds between blend constituents and their base words (1.3). These features will allow us to identify different categories of blend structures[1].

1.1. Constituent status and linear relationships

The nature of the units that are used for the formation of multi-word words is quite diverse. Root compounds, word compounds and blends present specific types of constituents.
The first constituent of root compounds is systematically a root that can either be a neoclassical (e.g., *hidr* 'water') or a vernacular form (e.g., *rat* 'mouse')[2]. This root is followed by another root (e.g., *cid* 'kill') or by a word (e.g., *solúvel* 'soluble'). Furthermore, the constituents of root compounds are linked by a specifier that can be /i/, if it precedes one of a small set of Latinate forms (e.g., *raticida*), or a round vowel[3] elsewhere (e.g., *hidrossolúvel*). As for word compounds, all constituents are words. *Barco-casa* 'houseboat', for instance, is formed by the word *barco* 'boat' and the word *casa* 'house'.
The nature of the constituents of blends is not as clear-cut as it is the case of compounds, since they may assume a variety of forms that range from ad hoc truncated chunks to words.
Most blends are formed by two constituents[4]. Usually, the leftmost constituent corresponds to the initial sequence of a word (cf. 2a), and the rightmost constituent matches the final sequence of another word (cf. 2b). Other kinds of constituents, quite rarely found and generally occurring on the right-hand side of the blend, correspond to the initial sequence of the second base word (cf. 2c).

(2)
a. ***trafi****crente* 'church-goer dealer'
< ***trafi****cante* 'dealer'+ *crente* 'believer'
b. *cão****didato*** 'candidate dog'
< *cão* 'dog' + *can****didato*** 'candidate'
c. *fut****sal*** 'indoor soccer'
< *futebol* 'football' + ***sal****ão* 'hall'

Some constituents (e.g., *homo*) are formally very similar to roots. However, the analysis of the meaning of words such as *homossensual* 'sexy homosexual' (< *homossexual* 'homosexual' + *sensual* 'sexy') reveals that the sequence *homo* is a clip of *homossexual*, not the neoclassical root *hom* meaning 'the same', followed by the linking vowel *-o-*. Therefore, although they look like morphological compounds, words such as *homossensual* are indeed blends.

The analysis of a considerable number of items has evidenced that many blends include sequences that correspond to words (henceforth W). These sequences can occur either on the leftmost position (3a) or on the rightmost position (3b) of the blend.

(3)

a. [***cão***] $_{word}$ *didato*

< ***cão*** 'dog' + *candidato* 'candidate'

b. *trafi* [***crente***] $_{word}$

< *traficante* 'dealer'+ ***crente*** 'believer'

The other constituents in these blends have a less stable nature. They may be splinters[5] that concatenate with the other constituent by juxtaposition (cf. 4a), or by some sort of agglutination[6] (cf. 4b).

(4)

a. [*caipi*] $_{clip}$ [*fruta*] $_{word}$ 'fruit caipirinha'

1. *caipi* $_{SPLINTER}$ < *caipir*] $_{NR}$ *inh*] $_{NR}$ *a*] $_{NS}$] $_{N}$ 'Brazilian drink'

2. *fruta*] $_{N}$ 'fruit'

b. [*boa*] $_{word}$ [*conha*] $_{clip}$ 'good cannabis'

1. *bo*] $_{NR}$ ***a***] $_{NS}$] $_{N}$ 'good'

2. *conha* $_{SPLINTER}$ < ***ma****conha*] $_{N}$ 'cannabis'

Alternatively, if the overlapping sequence is included, these constituents may be roots (cf. 5a), stems (cf. 5b), or even words (cf. 5c)[7]. The non-overlapping sequence, however, does not have a morphemic nature[8].

(5)

a. [*pregui*] $_{clip}$ [*sonso*] $_{word}$ 'lazy and cunning'

1. *pregui* $_{SPLINTER}$ < ***pregui***<u>ç</u>] $_{NR}$ *os*] $_{ADJR}$ *o*] $_{ADJS}$] $_{ADJ}$ 'lazy'

2. <u>*s*</u>*onso*] $_{N}$ 'cunning'

b. [*partici*]$_{clip}$ [*passivo*]$_{word}$ 'participative and passive'

1. *partici* $_{SPLINTER}$ < ***partici***<u>*p*</u>] $_{VR}$ <u>*a*</u>] $_{VS}$ *tiv*] $_{ADJR}$ *o*] $_{ADJS}$] $_{ADJ}$ 'participative'

2. <u>*pa*</u>*ssivo*] $_{ADJ}$ 'passive'

c. [*croi*] $_{clip}$ [*sandes*] $_{word}$ 'croissant and sandwich'

1. *croi* $_{SPLINTER}$ < ***croi***<u>*ssant*</u>] $_{NR}$] $_{NS}$] $_{N}$

2. <u>*san*</u>*des*] $_{N}$ 'sandwich'

We treat these constituents, collectively, as clips[9] (henceforth C). Clips thus include splinters, and roots, stems, or words that overlap with the adjacent constituent.

The overlapping sequences challenge the analysis of blends, since they may be assigned to any of the constituents, or even to both. In the case of *traficrente* (cf. 3b), for instance, the overlapping sequence (i.e., *c* = [k]) may be assigned (i) to the first constituent (cf. *trafic-rente*), which amounts to having a root-splinter sequence; (ii) to the second constituent (cf. *trafi-crente*), which implies having a splinter-word sequence; or (iii) to both constituents (cf. *trafic-crente*), which yields a root-word sequence that requires the postulation of a subsequent truncation operation. We have selected the option that favours the segmentation at a syllable boundary (henceforth $). Hence, in the case of *traficrente*, we have assigned the overlapping sequence to the rightmost constituent (cf. [*trafi*] $ [*crente*]). If the syllabic segmentation is not conclusive, the presence of a word must be prioritized since the other constituent is a splinter anyway. For instance, in the

case of *tristemunho* 'sad testimony' (< *tris**te*** 'sad' + *tes**te**munho* 'testimony'), the overlapping sequence (i.e., *te* = [te]) may be assigned to any constituent (cf. [*tris**te***] $ [*munho*] vs. [*tris*] $ [***te**munho*]). According to this last criterium, the overlapping sequence must be assigned to the first constituent (i.e., [*tris**te***] $ [*munho*]), because *triste* is a word and both *temunho* and *munho* are splinters.

Blends that do not include a word are formed by two clips. The rightmost constituent is always a splinter since it takes the right-hand periphery of the word. The first constituent may, however, be a splinter (cf. 6a), a root (cf. 6b), or a word (cf. 6c). The non-overlapping sequence, as in the above case, has not a morphemic nature.

(6)

a. [*democra*] $_{C}$ [*dura*] $_{C}$ 'authoritarian democracy'
< *democr**a**ci*] $_{NR}$ *a*] $_{NS}$] $_{N}$ 'democracy'
+ *dit**a**dur*]$_{NR}$ *a*] $_{NS}$] $_{N}$ 'dictatorship'

b. [*cren*] $_{C}$ [*talha*] $_{C}$ 'church-goers riffraff'
< *cr**ent***] $_{ADJR}$ *e*] $_{ADJS}$] $_{ADJ}$ 'believer'
+ *g**ent**alh*] $_{NR}$ *a*] $_{NS}$] $_{N}$ 'riffraff'

c. [*arru*] $_{C}$ [*mário*] $_{C}$ 'storing closet'
< *arru**m***] $_{VR}$ ***a***] $_{VS}$ ***r***] $_{V}$ 'to store'
+ *ar**már**i*] $_{NR}$ o] $_{NS}$] $_{N}$ 'closet'

The above analysis supports the identification of the following linear structures for Portuguese blends[10]:

(7)

CW *caipifruta* < [*caipi*] $_{C}$ *rinha* + [*fruta*] $_{P}$
WC *boaconha* < [*boa*] $_{P}$ + *ma* [*conha*] $_{C}$
CC *democradura* < [*democra*] $_{C}$ *cia* + *dita* [*dura*] $_{C}$

A final note is due to a particular kind of blends that are formed by the incorporation of a clip or a short word into a larger form (cf. 8).

(8)

*a**cão**xonado* 'in love with dogs'
< *a**pai**xonado* 'in love' *cão* 'dog'
*re**pulg**nante* 'repulsive like a flea'
< *re**pug**nante* 'repulsive' *pulga* 'flea'

These blends raise a classification issue since it is uncertain which of the constituents should be considered the first one. They deserve further research, which is out of the scope of this paper.

1.2. Grammatical relationships

Grammatical relationships that hold for compounds offer important clues about blends. There are two kinds of root compounds: modification structures (e.g., *toxicodependente* 'drug addict' < *toxic* 'drug' + *dependente* 'addict') and coordinated structures (e.g., *lusobrasileiro* 'Portuguese and Brazilian' < *lus* 'Portuguese' + *brasileiro* 'Brazilian'). Word compounds comprise three kinds of structures: modification structures (e.g., *barco-casa* 'houseboat' < *barco* 'boat' + *casa* 'house'); coordinated structures (e.g., *bar-restaurante* 'bar-restaurant' < *bar* 'bar' + *restaurante* 'restaurant'); and conversion structures (e.g., *saca-rolhas* 'corkscrew' < *saca* 'pulls' + *rolhas* 'corks').

In the case of modification structures, headedness is structurally established: root compounds are always head-final (cf. 9a), whereas word compounds are always head-initial (cf. 9b). Non-head constituents are modifiers in both cases. The position of the head indicates that root compounds (cf. 9a) are morphological structures. They are

compound roots that require further morphological and morphosyntactic specification to become words. It also indicates that word compounds (cf. 9b) are syntagmatic structures that inherit the morphosyntactic specifications (i.e., gender and number) of the head constituent. Semantics corroborates this analysis since root compounds (cf. 9a) are hyponyms of the rightmost constituent, whereas word compounds (cf. 9b) are hyponyms of the leftmost constituent.

(9)
a. [*toxicodependent*] *e* $_{\text{sing}}$ 'drug addict'
[*toxicodependent*] *es* $_{\text{pl}}$ 'drug addicts'
kind of *dependente* 'addict'
[*apicultor*] *ø* $_{\text{masc}}$ 'beekeeper (masc)'
[*apicultor*] *a* $_{\text{fem}}$ 'beekeeper (fem)'
kind of *cultor* 'keeper'
b. [*bomba* $_{\text{sing}}$ *relógio*] $_{\text{sing}}$ 'time bomb'
[*bombas* $_{\text{pl}}$ *relógio*] $_{\text{pl}}$ 'time bombs'
kind of *bomba* 'bomb'
[*águia* $_{\text{fem}}$ *macho* $_{\text{masc}}$] $_{\text{fem}}$ 'male eagle'
kind of *águia* 'eagle'
[*elefante* $_{\text{masc}}$ *fêmea* $_{\text{fem}}$] $_{\text{masc}}$ 'female elephant'
kind of *elefante* 'elephant'

In the case of coordination structures, all the constituents are heads because they are evenly involved in the structure, which means that these compounds may be analyzed as either multiheaded or as headless structures. The semantics of both kinds of coordinated compounds (adjectives, and nouns alike) is quite similar. They can either refer to a (property of an) entity that accumulates the properties of all the compound constituents (cf. 10a), or to a set of (properties of) entities formed by the compound constituents (cf. 10b).

(10)
a. *cidadão* [*luso-brasileiro*] $_{\text{ADJ}}$ 'Luso-Brazilian citizen'
< *cidadão* 'citizen' + *lus* 'Portuguese' + *brasileiro* 'Brazilian'
[*saia-calça*] $_{\text{N}}$ 'culotte'
< *saia* 'skirt' + *calça* 'pants'
b. *acordo luso-brasileiro*] $_{\text{ADJ}}$ 'Luso-Brazilian agreement'
< acordo 'agreement' + *lus* 'Portuguese' + *brasileiro* 'Brazilian'
[*saia-casaco*] $_{\text{N}}$ 'skirt suit'
< *saia* 'skirt' + *casaco* 'coat'

The examples in (10a) are cumulative: a Luso-Brazilian citizen is a citizen that has two nationalities (Portuguese and Brazilian); a *saia-calça* is a single garment. The examples in (10b) refer to a set: a Luso-Brazilian agreement is an agreement between the members of a set (i.e., Portugal and Brazil); a *saia-casaco* is a garment formed by two pieces (i.e, a skirt and a coat).

The distinction between root and word compounds that have a coordinated structure is due to how their morphosyntactic features are computed. Root compounds (cf. 11) have no internal morphosyntactic specifiers. Therefore, morphosyntactic specification has scope over the whole compound.

(11)
a. [*lusobrasileir*] *o* $_{\text{ADJmasc, sg}}$ 'Luso-Brazilian (masc)'
[*lusobrasileir*] *a* $_{\text{ADJfem, sg}}$ 'Luso-Brazilian (fem)'
b. [*socioeconómic*] *o* $_{\text{ADJmasc, sg}}$ 'socioeconomic (sg)'
[*socioeconómic*] *os* $_{\text{ADJmasc, pl}}$ 'socioeconomic (pl)'

Word compounds (cf. 12) behave differently. Animate word compounds (cf. 12a) require gender and number agreement: these values are jointly assigned to the compound. In the case of inanimate word compounds, only number agreement is required (cf. 12b, 12c). Gender is inherited from the value of both constituents, when they are identical (cf. 12b), or it is set as masculine, if the value of the constituents differs, probably because masculine is the generic gender value in Portuguese (cf. 12c).

(12)
a. [*cantor* $_{\text{Nmasc; sg}}$ *compositor* $_{\text{Nmasc; sg}}$] $_{\text{Nmasc; sg}}$
'singer-songwriter (masc, sg)'
[*cantora* $_{\text{Nfem; sg}}$ *compositora* $_{\text{Nfem; sg}}$] $_{\text{Nfem; sg}}$
'singer-songwriter (fem, sg)'
[*cantores* $_{\text{Nmasc; pl}}$ *compositores* $_{\text{Nmasc; pl}}$] $_{\text{Nmasc; pl}}$
'singer-songwriter (masc, pl)'
[*cantoras* $_{\text{Nfem; pl}}$ *compositoras* $_{\text{Nfem; pl}}$] $_{\text{Nfem; pl}}$
'singer-songwriter (fem, pl)'
b. [*bar* $_{\text{Nmasc; sg}}$ *restaurante* $_{\text{Nmasc; sg}}$] $_{\text{Nmasc; sg}}$
'bar-restaurant (masc, sg)'
[*bares* $_{\text{Nmasc; pl}}$ *restaurantes* $_{\text{Nmasc; pl}}$] $_{\text{Nmasc; pl}}$
'bar-restaurant (masc, pl)'
[*saia* $_{\text{Nfem; sg}}$ *calça* $_{\text{Nfem; sg}}$] $_{\text{Nfem; sg}}$ 'culotte (fem, sg)'
[*saias* $_{\text{Nfem; pl}}$ *calças* $_{\text{Nfem; pl}}$] $_{\text{Nfem; pl}}$ 'culotte (fem, pl)'
c. [*saia* $_{\text{Nfem: sg}}$ *casaco* $_{\text{Nmasc; sg}}$] $_{\text{Nmasc; sg}}$ 'skirt suit (masc, sg)'
[*sofá* $_{\text{Nmasc; sg}}$ *cama* $_{\text{Nfem; sg}}$] $_{\text{Nmasc; sg}}$ 'sofa bed (masc, sg)'

Blends display some identical and some different properties[11]. Like compounds, they split over modification and coordination structures. However, modification blends can be either head-final[12] (cf. 13a), like root compounds (cf. 9a), or head-initial[13] (cf. 13b), like word compounds (cf. 9b). Word class and gender value of the blends are always set by the head constituent:

(13)
a. MH [*cãominhada*] $_{\textbf{Nfem}}$ 'walk with dogs'
< *cão* $_{\text{masc}}$ 'dog' + *caminhada* $_{\textbf{Nfem}}$ 'walk'
MH [*tristemunho*] $_{\textbf{Nmasc}}$ 'sad testimony'
< *triste* $_{\text{Adj}}$ 'sad' + *testemunho* $_{\textbf{Nmasc}}$ 'testimony'
b. HM [*cartomente*] $_{\textbf{Nfem}}$ 'lying fortune teller'
< *cartomante* $_{\textbf{Nfem}}$ 'fortune teller' + *mente*$_{\text{V}}$ 'lies'
HM [*caligrafeia*] $_{\textbf{Nfem}}$ 'ugly calligraphy'
< *caligrafia* $_{\textbf{Nfem}}$ 'calligraphy' + *feia* $_{\text{ADJfem}}$ 'ugly'
HM [*pirilimpo*] $_{\textbf{Nmasc}}$ 'clean firefly'
< *pirilampo* $_{\textbf{Nfmasc}}$ 'firefly' + *limpo* $_{\text{ADJmasc}}$ 'clean'

From a formal point of view, coordinated blends contrast with both types of coordinated compounds. Noun gender is again an important feature. Blends that refer to animate entities require internal and external gender agreement (cf. 14a), as seen with word compounds (cf. 12a). But unlike word compounds, the gender of inanimate blends (cf. 14b) is apparently set by the rightmost constituent, which suggests its prominence and a closeness but not an identity to root compounds.

(14)
a. *namorido* $_{\text{Nmasc}}$ 'each of the boyfriends sharing a house'
< *namorado* $_{\text{Nmasc}}$ 'boyfriend' + *marido* $_{\text{Nmasc}}$ 'husband'
cantriz $_{\text{Nfem}}$
< *cantora* $_{\text{Nfem}}$ 'singer (fem)' + *atriz* $_{\text{Nfem}}$ 'actress'

b. *burkini* $_{Nmasc}$
< *burka* $_{Nfem}$ 'burka' + *biquíni* $_{Nmasc}$ 'bikini'
diciopédia $_{Nfem}$
< *dicionário* $_{Nmasc}$ 'dictionary' + *enciclopédia* $_{Nfem}$
'encyclopaedia'

The analysis of grammatical relationships within blends has therefore allowed us to identify the following structures[14]:

(15)
HM *caipifruta* < [*caipi*] $_{H}$ (*rinha*) + [*fruta*] $_{M}$
MH *boaconha* < [*boa*] $_{M}$ + (*ma*) [*conha*] $_{H}$
HH *burkini* < [*bur*] $_{H}$ (*ka*) + (*bi*) [*quíni*] $_{H}$

1.3. Phonetic / prosodic relationships

Phonetics and prosody are not as relevant for the analysis of compounds as they are for the analysis of blends[15]. Therefore, in this section, we will not bring the compounds to the discussion.

Stress position in blends coincides with the position of the stress of its rightmost constituent[16]. Therefore, they are single prosodic domains.

(16)
enxadachim
enxada 'hoe' + *espadachim* 'swordsman'
pistralhadora
pistola 'pistol'+ *metralhadora* 'machine gun'
dramédia
drama 'drama' + *comédia* 'comedy'

Another important feature of blends regards the way their constituents are phonetically concatenated[17]. Some are simply juxtaposed (cf. 17a), but, as previously mentioned, a large majority of cases involves a more complex operation of concatenation that overlaps the end of the leftmost constituent with the beginning of the rightmost. The array of overlapping possibilities deserves a closer look, since it ranges from a single segment (cf. 17b), or a syllable (cf. 17c) to larger and more complex sequences (cf. 17d), but this discussion is out of the scope of this paper.

(17)
a. *bara*(*lhado*) 'shuffled' + (*con*)*fundido* 'confused'> *barafundido*
b. *bur***[k]** (*a*) 'burka' + (*bi*) **[k]***íni* 'bikini' > *burkini*
c. *tris***[ti]** 'sad' + (*tes*) **[ti]***munho* 'testimony' > *tristemunho*
d. *diplo**mata*** 'diplomat' + *ma**mata*** 'gravy train' > *diplomamata*

Finally, the analysis of the Portuguese data suggests that the length[18] of the blend frequently coincides with the length of one of the base words (cf. 18a and 18b), and it may even coincide with the length of them both (cf. 18c). In a smaller number of cases, the blend is longer than any of its base words (cf. 18d).

(18)
a. ***namorido*** (4) ***namorado*** (4), *marido* (3)
b. ***tristemunho*** (4) *triste* (2), ***testemunho*** (4)
c. ***portunhol*** (3) ***português*** (3), ***espanhol*** (3)
d. *aminimigo* (5) *amigo* (3), *inimigo* (4)

This finding has allowed us to set another typology of blend structures, based on the prosodic prominence of one of the base words[19]:

(19)
prominence of the 1st base word (1BW) *namorido*
prominence of the 2nd base word (2BW) *tristemunho*

prominence of both base words (12BW) *portunhol*
no prosodic prominence (0BW) *aminimigo*

2. The Portuguese blend corpus (PBC)

Studying blends requires access to raw data, which is not easy to get. The marginal status of most of these words tends to exclude them from dictionaries or other lexicographic registers. Therefore, building a blend corpus is mandatory for research on this kind of words. For the moment, PBC is an Excel file. Although new coinages make this a never-ending work in progress, it is presently formed by (circa) 300 blends. Hopefully, the corpus will be publicly available in a near future.

Portuguese blends have received more attention from Brazilian linguists than from Portuguese linguists, and somehow, the idea that European Portuguese lacked this kind of complex words has long dominated. Building a blend corpus has proven otherwise. In fact, blends that originate in Brazilian Portuguese are easily traceable, but European Portuguese has contributed with a considerable number of items too. Some Angolan and Mozambican blends were also included.

Since Portuguese is a multi-centered language and the familiarity of the Portuguese speakers with social events that are often crucial to interpret blends is highly constrained by nationality, we have decided to discard blends that use proper nouns, since they are utterly opaque for non-resident speakers. All the remaining blends have been marked according to the language variety of the original coinage: Angolan Portuguese (AP), Brazilian Portuguese (BP), European Portuguese (EP), and Mozambican Portuguese (MP)[20].

PBC items comprise data reported in the literature [21], and original data, including some very recent forms, such as *pãodemia* ('bread + pandemic'), a blend attested in 2020, during the first COVID-19 lockdown. The corpus includes blends that have a documented literary origin and many others, namely those that are constantly coined in social and political contexts[22]. Since blending is an exercise of linguistic creativity, bound to no explicit constraints, a wide diversity of cases may stream. Therefore, it is important to annotate the corpus as thoroughly as possible, considering structural features such as those presented in section 1, but also information regarding their coinage and usage. Thus, the PBC annotation includes:

I. information regarding the blend
i. first attestation and authorship, whenever traceable, or a good example of its usage;
ii. POS, according to the registered context;
iii. frequency (corpus NOW);
iv. morphophonological representation;
v. phonetic transcription in EP and BP;
vi. number of syllables;
vii. identification of the stressed syllable;
viii. identification of the prosodically prominent base word (1BW, 2BW, 12BW, 0BW).
II. information regarding each constituent
i. identification of the base word;
ii. POS of the base word;
iii. phonetic transcription in EP and BP;
iv. number of syllables;
v. identification of the stressed syllable;

vi. status of each of constituent (clip / word);
vii. grammatical role of each constituent (head / modifier);
viii. identification of the most frequent words that share the sequence in the clip.
The following sections present some of the features of the corpus that were not previously discussed.

2.1. Attestation survey

Each entry of the PBC incorporates a link to the first attestation of the blend (if it is traceable, or a good attestation, if the first one is untraceable or ambiguous). For instance, *escopetarra*, from *escopeta* 'shotgun' + *guitarra* 'guitar' is linked to a Wikipedia entry (i.e., pt.wikipedia.org/wiki/ Escopetarra) that explains the origin of this blend[23].

(20)
escopetarra $_{Nfem}$
meaning = 'guitar built from a modified firearm'
1st base word = *escopeta* $_{Nfem}$ 'shotgun'
1st blend constituent = *escope* $_{C}$
1st base word = modifier
2nd base word = *guitarra* $_{Nfem}$ 'guitar'
2nd blend constituent = *tarra* $_{C}$
2nd base word = head

Access to an attestation is crucial to the identification of the word class of the blend, its morphosyntactic features, such as gender, its meaning, which is related to the retrieval of the base words, and the grammatical structure of each blend, as well as the status of each constituent (clip or word).

2.2. Frequency issues

In general, blends are very low frequency words because they are formed as a creative gesture and not to respond to a specific semantic requirement. However, it is possible to set a difference between blends that, in a contemporary corpus[24], have less than ten tokens (cf. *abreijo* 'hug + kiss'), and those that have already become part of the active Portuguese lexicon, like, for instance, *portunhol* 'mix of Portuguese and Spanish'. Therefore, we have decided to record the frequency value of each blend.
The blends that we have considered in this paper include a considerable number of cases that have 0 records[25] in the corpus NOW (cf. 21a), others that are attested and display less than 100 tokens (cf. 21b). The remaining few are much more frequent[26] (cf. 21c):

(21)
a. *anãofabeto* 0
*agradádiva** 0
*barafundid** 0
*batatalhau** 0
*boaconha** 0
*craquétic** 0
*croissandes** 0
*curibaci** 0
*diplomamata** 0
*escopetarra** 0
*gestemunho** 0
*gestont** 0
*pãodemia** 0

*participassiv** 0
*pirilampisc** 0
*pistralhadora** 0
*preguissons** 0
b. *aminimig** 01
*caligrafeia** 01
*cartomente** 01
*enxadachi** 01
*frangl** 01
*homessensua** 01
*pirilimp** 01
*traficrente** 01
*bótim** 02
*tristemunho** 02
*crentalha** 03
*fabulástic** 03
*impastor** 03
*aborrescente** 04
*cãodidat** 04
*abreijo** 05
*escreviv** 06
*analfabrut** 09
*democradura** 13
*cantriz** 15
*manifestoche** 15
*caipifruta** 30
*apertament** 34
*burquini** 42
*namorid** 59
*chafé** 60
*dramédia** 80
c. *portunho** 250
*cãominhada** 267
*cantautor** 752
*futsa** 20.742

The frequency of the blend's constituents is also registered in the PBC. Since clips are non-morphemic chunks of the base word, the identification of that base word, and, hence, of the meaning it conveys, may be a challenging operation[27].

The constituents of blends that correspond to words are easy to retrieve (cf. *bruto* 'gross' in *analfabruto* 'illiterate and gross'), but the other constituent (i.e., *analfa*) needs to be matched with an existing word[28]. In the case of *analfabruto*, there is only one candidate (i.e., *analfabeto*), which facilitates the understanding of the blend. In other cases, matching the non-word blend constituent with an attested word is not straightforward. The example in (22), i.e., *gestonta*, includes the word *tonta* 'silly (fem)'. The remaining sequence (i.e., *ges*) matches words that belong to three different root families, and in all cases, include the first segment of *tonta*. The adjectival nature of *tonta* constrains the choice of the base word of *ges*, which helps to exclude implausible groupings. Considering word class and agreement requirements, the set of plausible groupings includes only feminine nouns. Therefore, the acceptable options, equally plausible, are *gestão* 'management', *gestante* 'pregnant' and *gestora* 'manager'. Therefore, out of context, *gestonta* is an ambiguous blend[29]. The frequency of these matching words may be relevant for the interpretation of the blend[30]: *gestão* and *gestora* are the best candidates, although *gestante* is the base that was used for the coinage of this blend.

(22)
ges (*tonta*)
1. gest $_{NR}$ (1928)
gestão, gestor(a)/es
2. gest $_{NR}$ (1329)
gesto/s; gestual/is; gesticular
3. gest $_{VR}$ (358)
gestação; gestante/s

Blends that are formed by two clips complexify the matching operation. For example, the context of the noun *abreijo* suggests that it is a sort of salutation, but the segmentation brings other possibilities. The sequences that reach no hits in the corpus NOW, as well as the sequences that have a very large number of hits must be excluded (cf. 23). The sequence *abr-eijo* is, therefore, the best because *eijo* only matches three roots.

(23)
a > 1000 hits *breijo* 0 hits
ab > 1000 hits *reijo* 0 hits
abr > 1000 hits *eijo* 3 hits (*beijo*, *queijo, aleijo*)
abre > 472 hits *ijo* > 1000 hits
abrei > 1 hit (*abreijo*) *jo* > 1000 hits
abreij > 1 hit (*abreijo*) *o* > 1000 hits

The context helps to select the base word *beijo* (cf. 24a). The remaining sequence, i.e., *abr* (cf. 24b), matches *abraço* because the target must be a noun and it must be compatible with the salutation meaning suggested by the context.

(24)
a. (*abr*) *eijo*
queij $_{NR}$ (404) *queijo* $_{N}$ 'cheese'
beij $_{NR}$ (346) *beijo* $_{N}$; *beijo* $_{V}$ 'kiss'
aleij $_{VR}$ (5) *aleijo* $_{V}$ 'hurt'
b. abr (eijo)
abril $_{NR}$ (5757) *abril* $_{N}$ 'April'
abr $_{VR}$ (4874) *abrir* $_{V}$ 'open'
abrig $_{VR}$ (686) *abrigar* $_{V}$ 'shelter'
abrang $_{VR}$ (545) *abranger* $_{V}$ 'include'
abraç $_{VR}$ (358) *abraço* $_{N}$; *abraçar* $_{V}$ 'hug'
abrupt $_{ADJR}$ (84) *abrupto* $_{ADJ}$ 'abrupt'

It is worth mentioning that these words allow different interpretations. The PBC only lists possible matchings of clips with base words, and plausible combinations of base words.

2.3. Phonological, phonetic, and prosodic information

The PBC includes two language varieties, European and Brazilian Portuguese, that have different contrasting features in many domains. One of them, that is relevant for the analysis of blends, is the vowel system. Unstressed vowels in European Portuguese tend to be high vowels, whereas Brazilian Portuguese tends to preserve their phonological quality. For that reason, the PBC includes the phonetic transcription of all blends in EP and BP[31], as well as the number of syllables and the stress position. Finally, the morphematic structure of each blend is also included (cf. 25):

(25)
apertamento
#apertament+u#

EP [ɐpɨrtɐ'mẽtu]
BP [aperta'mẽtu]
number of syllables – 5
stress position – penultimate syllable

The same set of features is included in the description of the base word, which allows to check if the blending operation occurs at a morphemic boundary, or not. Truncated sequences are presented inside round brackets, and overlapping segments are marked in bold characters:

(26)
1st base word – *apertado* 'tight'
#a+pert+a+d+u#
EP ɐpɨr (**'t+a**+d+u)
BP aper (**'t+a**+d+u)
number of syllables – 4
stress position – penultimate
2nd base word – *apartamento* 'apartment'
#a+part+a+ment+u#
EP (ɐpɐr) **t+ɐ**+'mẽt+u
BP (apar) **t+a**+'mẽt+u
number of syllables – 5
stress position – penultimate

3. Analysis of the PBC

The analysis of a subcorpus of the PBC, formed by 184 blends, has already allowed to set some hypothesis about the nature of blend structures. This section presents an account of our current findings that are related to linear and grammatical relationships, and to the prosodic profile of blends. This analysis suggests that there is a correlation between cliphood, headedness, and prosodic prominence.

3.1. Linear structure

As mentioned in section 1.1, linear structure is set according to the status of the blend constituents. In the subcorpus of the PBC, the number of CW blends (cf. 27a) is close to the number of CC blends (cf. 27b), and they are both larger than the set of WC blends (cf. 27c). This basic classification will prove, as we will see in a moment, to be quite powerful and far-reaching.

(27)
a. CW 39% *caipi*(*rinha*) $_{C}$ + *fruta* $_{W}$
'drink' + 'fruit'
b. CC 37% *abr*(*aço*) $_{C}$ + (*b*)*eijo* $_{C}$
'hug' + 'kiss'
c. WC 24% *boa* $_{W}$ + (*ma*)*conha* $_{C}$
'good' + 'cannabis'

3.2. Grammatical structure

The grammatical status of the blend constituents (cf. section 1.2) allowed us to find out that almost half of the blends have a coordinated structure (cf. 28a). The remaining units are more often head-final structures (cf. 28b), like morphological compounds, than head-initial structures (cf. 28c), which are closer to the structure of one subtype of morpho-syntactic compounds.

(28)
a. HH 48% *cant*(*or*) $_{C}$ + (*au*)*tor* $_{C}$
'singer' + 'author'
b. MH 33% *tris*(*te*) $_{C}$ + (*tes*)*temunho* $_{C}$
'sad' + 'testimony'
c. HM 19% *cartom*(*ante*) $_{C}$ + *mente* $_{W}$
'fortune teller' + 'lies'

3.3. Prosodic prominence

The phonetic and prosodic relationships that hold between blends and their base words (cf. 1.3), allowed us to identify four prosodic patterns: 1BW are blends that inherit the prosodic template of the first base word; 2BW are those that inherit the prosodic template of the second base word; 12BW are blends that inherit the prosodic structure of both base words (a subclass of the previous two types); and the remaining are blends that differ from the prosodic structure of any of the base words (0BW).
The analysis of the PBC revealed that most blends are prosodically related to one of the base words (cf. 29a and 29b), and a small percentage coincides prosodically with both base words (cf. 29c). Blends that diverge prosodically from any of the base words form a smaller set (cf. 29d).

(29)
a. 1BW 39% *imp*(*ostor*) $_{C}$ + *pastor* $_{W}$
'impostor' + 'minister'
b. 2BW 38% *cão* $_{W}$ + (*can*)*didato* $_{C}$
'dog' + 'candidate'
c. 12BW 07% *portu*(*guês*) $_{C}$ + (*espa*)*nhol* $_{C}$
'Portuguese' + 'Spanish'
d. 0BW 16% *agradá*(*vel*) $_{C}$ + *dádiva* $_{W}$
'pleasant' + 'gift'

3.4. Cross-analysis

The analysis of the interplay of the three structural types described above revealed unsuspected regularities. This cross-analysis is still based on the subcorpus of the PBC. The following tables display the partition of structures within the domain of each category.
Table 1 and Table 2 show that each grammatical structure prefers a given linear structure, and each linear structure prefers a given grammatical structure, almost symmetrically (i.e., HH-CC / CC-HH; MH-WC / WC-MH; and HM-CW). The coincidence between head constituents and clips is quite remarkable.

Table 1. grammatical role (H, M) vs. constituent status (W, C)

HH blends 48%	CC	49%	*franglês*
	CW	37%	*cantautor*
	WC	14%	*chafé*
MH blends 33%	WC	46%	*cãominhada*

	CC	33%	*aborrescente*
	CW	21%	*impastor*
HM blends 19%	CW	76%	*cartomente*
	CC	12%	*manifestoche*
	WC	12%	*craquético*

Table 2. constituent status (W, C) vs. grammatical role (H, M)

WC blends 39%	MH	64%	*cãominhada*
	HH	27%	*chafé*
	HM	9%	*craquético*
CC blends 37%	HH	65%	*franglês*
	MH	29%	*aborrescente*
	HM	6%	*manifestoche*
CW blends 24%	HH	46%	*cantautor*
	HM	36%	*cartomente*
	MH	18%	*impastor*

Table 3 and Table 4 show that head constituents often coincide with prosodically prominent base words (i.e., MH-2BW / 2BW-MH; HM-1BW). HH structures split over 1BW and 2BW and almost half of these two categories correspond to HH structures.

Table 3. grammatical role (H, M) vs. prosodic relationship (1BW, 2BW, 12BW, 0BW)

HH blends 48%	1BW	38%	*namorido*
	2BW	35%	*dramédia*
	0BW	17%	*aminimigo*
	12BW	10%	*portunhol*

MH blends 39%	2BW	54 %	*tristemunho*
	1BW	26 %	*impastor*
	0BW	15 %	*agradádiva*
	12BW	5%	*aborrescente*
HM blends 19%	1BW	65 %	*advogata*
	0BW	21 %	*argumentira*
	2BW	15 %	*cãodidato*
	12BW	0%	---

Table 4. prosodic relationship (1BW, 2BW, 12BW, 0BW) vs. grammatical role (H, M)

1BW blends 39%	HH	47%	*namorido*
	HM	31%	*advogata*
	MH	22%	*impastor*
2BW blends 38%	MH	48%	*tristemunho*
	HH	45%	*dramédia*
	HM	7%	*cãodidato*
0BW blends 16%	HH	48%	*aminimigo*
	MH	29%	*agradádiva*
	HM	23%	*argumentira*
12BW blends 7%	HH	75%	*portunhol*
	MH	25%	*aborrescente*
	HM	0%	---

The correlation between heads and clips, and heads and prosodically prominent base words suggests the existence of a similar correlation between clips and prosodically

prominent base words. That correlation can be confirmed in Tables 5 and 6 (i.e., WC-2BW / 2BW-WC; CW-1BW / 1BW-CW).

Table 5. constituent status (W, C) vs. prosodic relationship (1BW, 2BW, 12BW, 0BW)

WC blends 39%	2BW	73 %	*cãodidato*
	0BW	23 %	*batatalhau*
	1BW	4 %	*escopetarra*
	12BW	0%	---
CC blends 37%	2BW	44 %	*dramédia*
	1BW	29 %	*namorido*
	12BW	18 %	*portunhol*
	0BW	9 %	*fabulástico*
CW blends 24%	1BW	69 %	*impastor*
	0BW	21 %	*agradádiva*
	2BW	10 %	*bótimo*
	12BW	0%	---

Table 6. constituent status (W, C) vs. prosodic relationship (1BW, 2BW, 12BW, 0BW)

1BW blends 39%	CW	69%	*impastor*
	CC	28%	*namorido*
	WC	3%	*escopetarra*
2BW blends 38%	WC	46%	*cãodidato*
	CC	44%	*dramédia*
	CW	10%	*bótimo*
0BW blends 16%	CW	48%	*agradádiva*

	WC	32%	*batatalhau*
	CC	20%	*fabulástico*
12BW blends 7%	CC	100%	*portunhol*
	CW	0%	---
	WC	0%	---

In sum, the analysis of the relationship between structural types evidenced that heads, clips, and prosodically prominent base words are probably not randomly set. It is premature to present any strong interpretations of these findings that will eventually be checked against the remaining data from the PBC.

4. Experimental analysis

Forming and understanding blends are not symmetrical operations. Blend coinage is based on the manipulation of two words: one of them may be preserved; the other (or both) must be truncated. The output clip is apparently an ad hoc chunk of a word that may be formally close to the base word. Understanding a new blend is a process based on the recovery of missing information: each clip must match a word, and several matching hypotheses may arise. This property of blends justifies the need to elaborate an in-depth linguistic analysis, which we have tried to accomplish in the previous sections, and it also explains the interest they carry for the research on word processing.

This section brings a brief presentation of a previously reported experimental research (cf. Minussi & Villalva 2020): a familiarity test (cf. section 4.1) and a lexical decision test (cf. section 4.2). The results suggest that clips facilitate the processing of blends, since the clipped word is more frequently presented in replies to the familiarity test, and the same happens with reaction time values obtained in the lexical decision test[32].

4.1. Familiarity test

This test was based on a subset of 56 blends. The subjects were young adults (undergraduate students from the University of Lisbon and the Federal University of São Paulo). We asked participants to provide the meaning of each stimulus. The answers were coded to trace the replies that included the first base word, those that included the second base word, and the replies that mentioned both.

Table 7. Familiarity test

Retrieved base word (EP/BP)	Word-Clip	Clip-Word	Clip-Clip
First BW	53%	70%	74%
Second BW	63%	56%	64%

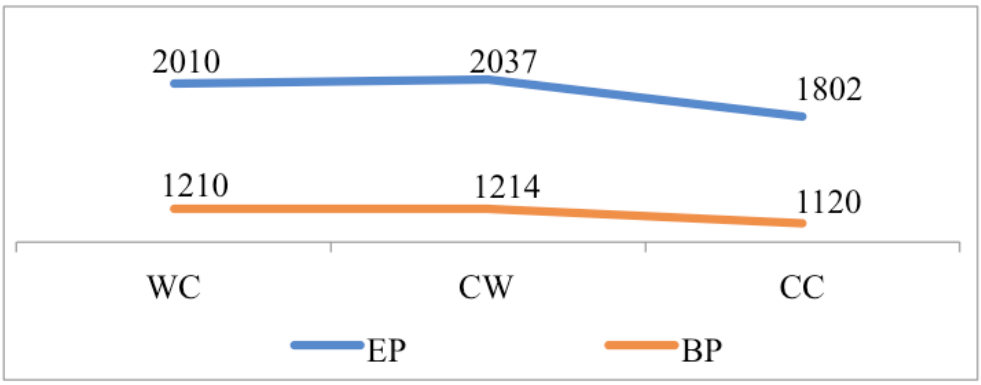
First + second BW	42%	43%	54%

As highlighted in Table 7[33], the results confirm the relevance of the clipped constituents, since they were listed more often than the constituents corresponding to full words. These results suggest that there is a reverse correlation between the visibility of the base word and lexical retrieval. Counter-intuitively, full forms are harder to retrieve than fragmented forms.

4.2. Lexical decision test

The lexical decision test was based on the same subset of the blend corpus and a similar sample of EP and BP subjects[34]. The results also help to consolidate the previous findings, since CC blends (e.g., *franglês* 'French and English') facilitate word processing - median reaction time, in these cases, is significantly lower than for CW (e.g., *aminimigo* 'friend and enemy') and WC (e.g., *anãofabeto* 'illiterate dwarf') blends. The contrast between WC and CC blends is statistically significant (p=0,016 EP and p=0,018 BP), and the same occurs between CW and CC blends (p=0,003 EP and p=0,007 PB). The graph in Figure 1 shows that the results are consistent in both EP and BP. Curiously, median reaction time is much higher in EP (2010ms for WC, 2037ms for CW and 1802ms for CC), than in BP (1210ms for WC, 1214ms for CW and 1120ms for CC), which may be related to the availability of phonetic clues, that, as mentioned above, differ in the two language varieties.

Figure 1. Lexical decision test



Concluding remarks

Building a blend corpus for Portuguese is a prerequisite for further research that will ultimately lead to a thorough analysis of this kind of words and for the study of word processing and lexical access.

The nature of the PBC is primarily a consequence of a blend analysis, but ultimately, it has allowed to unveil previously unsuspected structural correlations. In fact, the corpus analysis allowed us to hypothesize that most blends fall into a small number of prototypical structures:

1. Coordinated blends are optimally formed by the concatenation of two clips, and they are prosodically related to the second base word (cf. *dramédia*). They may also be formed by clip-

word structures, that are either prosodically equivalent to the first base word (cf. *traficrente*), or prosodically unrelated to any of the base words (cf. *aminimigo*).

2. Head-final blends are optimally formed by word-clip structures, and prosodically they are also typically related to the second base word (cf. *cãominhada*).
3. Finally, head-initial blends are optimally formed by clip-word structures, and they are prosodically related to the first base word (cf. *cartomente*).

These findings were experimentally corroborated by a familiarity test and a lexical decision test. Both tests produced an unexpected outcome, since clipped constituents were more frequently stated in the replies to the first test, and CC blends were apparently easier to process.

All these findings need to be cross-checked by future research, but the hypothesis that remains solidly on the table is that head constituents can more easily be clipped, and clipped constituents help the speakers to locate the head of the blend, which is crucial for assigning them a meaning.

BIBLIOGRAPHY

Andrade K. E. (2008). *Uma Análise Otimalista Unificada para Mesclas Lexicais do Português do Brasil*. Dissertação de Mestrado em Língua Portuguesa. Rio de Janeiro: UFRJ.

Beliaeva N. (2019). "Blending in morphology", in *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press. DOI: 10.1093/acrefore/9780199384655.013.511.

Cook P. & Stevenson S. (2010). "Automatically identifying the source words of lexical blends in English", *Computational Linguistics* 36 (1): 129-149. DOI: 10.1162/coli.2010.36.1.36104.

Gonçalves C. A. (2003). "Blends lexicais em português: não-concatenatividade e correspondência", *Veredas* 7. 1-2: 149-167.

Gries S. T. (2004a). "Isn't that fantabulous? How similarity motivates intentional morphological blends in English", in M. Achard & S. Kemmer (eds.) *Language, culture, and mind*. Stanford, CA: Center for the Study of Language and Information (CSLI), 415-428.

Gries S. T. (2004b). "Shouldn't it be breakfunch? A quantitative analysis of blend structure in English", *Linguistics* 42.3: 639-667. DOI: 10.1515/ling.2004.021.

Gries S. T. (2012). "Quantitative corpus data on blend formation: Psycho- and cognitive-linguistic perspectives", in V. Renner, F. Maniez & P. J. L. Arnaud (eds.) *Cross-disciplinary perspectives on lexical blending*. Berlin: De Gruyter Mouton: 145-167. DOI: 10.1515/9783110289572.145.

Johnson R. L., Slate S. R., Teevan A. R. & Juhasz B. J. (2019). "The processing of blend words in naming and sentence reading", *The Quarterly Journal of Experimental Psychology* 72(4), 847-857. DOI: 10.1177/1747021818768441.

Jurado A. B. (2019). "A study on the 'wordgasm': the nature of blends' splinters", *Lexis* [online] 14. DOI: 10.4000/lexis.3916.

Kubozono H. (1990). "Phonological constraints on blending in English as a case for phonology-morphology interface", *Yearbook of Morphology* 3, 1-20. DOI: 10.1515/9783112420744-001.

Minussi R. D. & Nóbrega V. (2014). "A interface sintaxe-pragmática na formação de palavras: avaliando os pontos de acesso da Enciclopédia na arquitetura da gramática", *Veredas* 18.1: 161-184.

Minussi R. D. & Villalva A. (2020). "Reconhecimento e acesso lexical dos blends em Português Europeu e Português Brasileiro", *Todas as Letras* 22.1: 1-14.

Piñeros C. E. (2004). "The creation of portmanteaus in the extragrammatical morphology of Spanish", *Probus* 16(2): 203-240. DOI: 10.1515/prbs.2004.16.2.203.

Pereira I. (2013). "Cruzamento vocabular em português", in F. Rainer, M. Russo & F. Sánchez Miret (eds.) *Actes du XXVIIe Congrès international de linguistique et de philologie romanes*, Nancy: ATILF/SLR. [www.atilf.fr/cilpr2013/actes/section-3/CILPR-2013-3-Pereira.pdf, 27-09-18].

Plag I. (2003). *Word-formation in English*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511841323.

Prearo-Lima R. (2019). "Blends lexicais e neologismos: alguns conceitos e problematizações", *Entrepalavras* 9.3: 38-56.

Renner V. (2019). "French and English Lexical Blends in Contrast", *Languages in Contrast* 19, 1: 27-47. DOI: 10.1075/lic.16020.ren.

Renner V. (2022). "Blending", in P. Ackema, S. Bendjaballah, E. Bonet & A. Fábregas (eds.), *Wiley Blackwell companion to morphology*. Hoboken, NJ: Wiley.

Rio-Torto G. (2014). "Blending, cruzamento ou fusão lexical em português: padrões estruturais e (dis)semelhanças com a composição", *Revista Filologia e Linguística Portuguesa* 16.1: 7-29.

Villalva A. & Gonçalves C. A. (2016). "The phonology and morphology of word formation", in L. Wetzels, J. Costa & S. Menuzzi (eds.) *The Handbook of Portuguese Linguistics*. Oxford: Wiley Blackwell: 167-187. DOI: 10.1002/9781118791844.

Villalva A. & Minussi R. D. (2019). "Word formation, processing, and lexical access: blends and compounds in European and Brazilian Portuguese". 12th Mediterranean Morphology Meeting. DOI: 10.13140/RG.2.2.23539.58407.

NOTES

1. Assuming that blends, like compounds, are multi-word lexical units, we will systematically bring compounds to the discussion.

2. Neoclassical roots are recent loans from Latin (e.g., *cid*) or Ancient Greek (e.g., *hidr*) that occur only in complex words (derivatives, like *hídrico* 'hydric'; or compounds, like *raticida* 'raticide' or *hidrossolúvel* 'water soluble'). They are semantically equivalent to vernacular roots that occur in simplex (e.g., *água* 'water') as well as in complex words (e.g., *aguado* 'watery').

3. This vowel is /ɔ/ in European Portuguese (henceforth EP), and /o/ in Brazilian Portuguese (henceforth BP).

4. Only one case of a three-constituent blend was found in Portuguese: *curibacil* is a noun formed by ***curi****oso* 'nosy', ***ba****baca* 'fool' and *imbe****cil*** 'imbecile'. Cf. forum.wordreference.com/threads/tipos-de-alunos.2829 866/#post-15055660 [05/10/2021].

5. For a recent discussion on splinters, see Jurado (2019).

6. We will not discuss this process in this paper.

7. If these constituents were in fact a root, a stem or a word, the output forms would correspond to compound structures.

8. ADJR=adjective root; ADJS=adjective stem; ADJ=adjective; NR=noun root; NS=noun stem; N=noun; VR=verb root; VS=verb stem; V=verb.

9. We use the term clip to refer to truncated words independently of their morphemic nature. For other discussions about splinters and clips see Beliaeva 2019.

10. Section 3 (I) will bring information on the percentual weight of each of these categories in the PBC.

11. The structure of blends has also been recently discussed by Renner 2019 and 2022.

12. Henceforth MH.

13. Henceforth HM.

14. Section 3 (II) will bring the information on the percentual weight of each of these categories in the PBC.

15. See Kubozono 1990, Gries 2004a and 2012, and Piñeros 2004, for more discussions on prosodic aspects of blends.

16. When the second constituent does not include the final sequence (e.g., *futsal*), the stress is assigned according to the general stress assignment system, but there are very few blends of this sort.

17. See Plag 2003, Gries 2004a, and Renner 2022, for other discussions on this subject.

18. Length is measured according to the number of syllables.

19. Section 3 (III) will bring the information on the percentual weight of each of these categories in the PBC.

20. The following examples include blends formed by reputed authors from four different Portuguese speaking countries:

1. José Luandino Vieira, 2009 (Angola)

gestemunho < *gesto* 'gesture' + *testemunho* 'testimony'

(books.google.pt/books?

id=tUCshXZTTwAC&pg=PT18&lpg=PT18&dq=gestemunho&source=bl&ots=9mon09A6vm&sig=ACfU3U0QQK8Ffjuha0oSIRssBebGJZRnfw&

PT&sa=X&ved=2ahUKEwiLipje4rjzAhUMxIUKHYDTCqUQ6AF6BAgSEAM#v=onepage&q=gestemunho&f=false)

2. Millôr Fernandes, 1968 (Brazil)

cartomente < *cartomante* 'fortune teller' + *mente* 'he/she ies'

(ronaldofranco.blogspot.com/2010/02/dicionarios-de-millor-fernandes.html)

3. Urbano Tavares Rodrigues, 1970 (Portugal)

escreviver < *escrever* 'to write' + *viver* 'to live'

(www.estantevirtual.com.br/sebonovafloresta/rodrigues-urbano-tavares-ensaios-de-escreviver-2760602607?show_suggestion=0)

4. Mia Couto, 2001 (Mozambique)

pirilampiscar < *pirilampo* 'firefly' + *piscar* 'to blink'

(bibliotecaweb20.pbworks.com/w/file/fetch/86022520/Mia-Couto-O-Gato-e-o-Escuro.pdf)

21. See, among others, Gonçalves 2003, Andrade 2008, Pereira 2013, Minussi & Nóbrega 2014, Rio-Torto 2014, Prearo-Lima 2019.

22. See, for instance:

foicebook < *foice* 'sickle' + *facebook* (foicebook.blogspot.com/)

familícia < *família* 'family' + *milícia* 'militia'

(www.dicionarioinformal.com.br/famil%C3%ADcia/)

23. The segmentation of *escopetarra* presented in (20) obeys to the criteria discussed in section 1.1: the last segment of the first constituent, which is a noun stem, is phonetically identical to the first segment of the second constituent. According to the criteria presented above, the overlapping segment is assigned to the second constituent, to coincide with the syllable boundary. Notice that the final vowel of *escopeta* is a thematic index that can never be stressed. Therefore, the stressed vowel in *escopetarra* comes from *guitarra*.

24. The corpus used for the evaluation of frequency is a subcorpus of *Corpus do Português* (Corpus Now), that has a coverage of more than a billion words, collected between 2012 and 2019.
25. This is quite expectable. Although the corpus NOW has a good language coverage, it includes only a selection of documents.
26. The asterisk following each blend indicates that the frequency value corresponds to a search that includes all inflected forms (i.e., *cãodidat** = *cãodiadato*, *cãodidata*, *cãodidatos*, *cãodidatas*).
27. Cook and Stevenson (2010) present a statistical approach to the problem of the identification of the source words that is based on the syllable structure.
28. We have used the *Corpus do Português Brasileiro,* available at www.lexicodoportugues.com/, to find the matching candidates.
29. As mentioned in the previous section, the context is crucial to the interpretation of the blend. However, in cases such as *gestonta*, when several matching options are available, ambiguity may persist even in context. This is way other matching possibilities must be considered.
30. The results of the familiarity test presented in section 4 corroborates this claim.
31. EP transcriptions are based on the phonetic transcription of the base words provided by *Infopédia,* which corresponds to the standard EP pronunciation. BP transcriptions are based on the pronunciation of a native speaker.
32. For more information on blend processing see Johnson R.L., Slate S.R., Teevan A.R., & Juhasz B.J. (2019). According to these authors, very little is known about complex word processing. Among the main results, the study showed that blends were processed more slowly than control words.
33. Each cell in this table corresponds to the percentage of the total number of answers. Notice that some answers (i.e., first base word or second base word, and first + second base words) were not mutually exclusive.
34. This test used Psychopy for data collection, and SPSS for the statistical analysis of the results.

ABSTRACTS

Multi-word formation processes in Portuguese comprise root-compounding (e.g., *toxicodependente* ‘drug addict’, *agridoce* ‘sour sweet’), word-compounding (e.g., *barco-casa* ‘houseboat’, *guarda-roupa* ‘wardrobe’, *cantora-atriz* ‘singer/actress’), and blending (e.g., *cartomente* ‘lying fortune teller’ < *cartomante* ‘fortune teller’ + *mente* ‘he/she lies’, *tristemunho* ‘sad testimony’ < *triste* ‘sad’ + *testemunho* ‘testimony’, *cantautor* ‘singer and composer’ < *cantor* singer + *autor* ‘composer’). Compound structures have been quite thoroughly described by several authors (e.g., Villalva & Gonçalves 2015), whereas blending has garnered some controversial and even contradictory analyses. Some authors claim that blends and compounds have similar structures, while others consider that they have completely different structures (e.g., Gries 2004b, Minussi & Nóbrega 2014).
This paper focuses on the description and analysis of an annotated corpus of Portuguese blends, and on the presentation of experimental evidence that aims to assess the knowledge of these words by native young adult European and Brazilian Portuguese speakers.
The first section is devoted to the presentation of Portuguese blends, namely regarding the status of their constituents and their linear relationships, grammatical structure, and the phonetic/prosodic relationship that holds between the blends and their base words. The second section

focuses on the description of the Portuguese blend corpus. Although new coinages make this a never-ending work in progress, it is presently formed by (circa) 300 blends, collected in a variety of sources, from literary texts to political and social commenting, advertising, or even individual ad hoc instances. The information gathered in the corpus comprehends all the features considered in the analysis of the blends, presented in section 1, and it also provides data on attestation and frequency. The third section presents an analysis of a sub-corpus (184 forms), related to their linear structure, grammatical and prosodic relationships, and a cross-analysis that suggests that there is a strong link between headedness and cliphood. Finally, the fourth section offers a brief presentation of experimental work that suggests that clips facilitate the processing of blends, since the clipped word is more frequently used in replies to a familiarity test, and the same happens with reaction time values obtained in a lexical decision test.

La formation de mots qui contiennent plus d'un lexème, en portugais, comprend la composition morphologique (p. ex., *toxicodependente* 'toxicomane', *agridoce* 'aigre-doux'), la composition morphosyntaxique (p. ex., *barco-casa* 'péniche', *guarda-roupa* 'garde-robe', *cantora-atriz* 'chanteuse et actrice'), et la formation de mots-valises (p. ex., *cartomente* 'diseuse de bonne aventure qui ment' < *cartomante* 'diseuse de bonne aventure' + *mente* 'il/elle ment', *tristemunho* 'triste témoignage' < *triste* 'triste' + *testemunho* 'témoignage', *cantautor* 'chanteur et compositeur' < *cantor* 'chanteur' + *autor* 'compositeur'). Les structures composées du portugais sont décrites en détail par plusieurs auteurs (p. ex., Villalva et Gonçalves 2015), tandis que la formation de mots-valises a reçu des analyses controversées et même contradictoires : certains auteurs affirment que la structure des mots-valises et des composés est similaire, tandis que d'autres considèrent qu'elles sont très différentes (p. ex., Gries 2004b, Minussi et Nóbrega 2014).
Cet article se concentre sur la description et l'analyse d'un corpus annoté de mots-valises, en portugais, et la présentation de quelques preuves expérimentales qui visent à évaluer la connaissance de ces mots par les jeunes adultes natifs lusophones, portugais et brésiliens. La première section est consacrée à la présentation des mots-valises du portugais, notamment en ce qui concerne le statut de leurs constituants et leurs relations linéaires, leur structure grammaticale et la relation phonétique/prosodique entre les mots-valises et les mots-base. La deuxième section se concentre sur la description du corpus de mots-valises du portugais. Bien qu'il s'agisse d'un travail sans fin, le corpus est, à présent, formé de (environ) 300 mots-valises, recueillis dans une variété de sources, des textes littéraires aux commentaires politiques et sociaux, à la publicité ou même aux contributions individuelles ad hoc. Les informations recueillies dans le corpus comprennent toutes les propriétés prises en compte dans l'analyse présentée à la section 1 et fournissent également des données sur attestations et fréquence. La troisième section présente une analyse d'un sous-groupe du corpus (184 formes), qui considère la structure linéaire et les relations grammaticales et prosodiques. Une analyse croisée suggère qu'il y a un fort lien entre la tête du mot-valise et le constituant-fragment. Enfin, la quatrième section offre une brève présentation de travaux expérimentaux antérieurs, qui suggèrent que les mots tronqués facilitent la compréhension des mots-valises, puisque le mot tronqué est plus fréquemment utilisé dans les réponses à un test de familiarité. Il en va de même avec les valeurs de temps de réaction obtenues dans un test de décision lexicale.

INDEX

Mots-clés: mots-valises, composés morphologiques, composés morpho-syntactiques, portugais
Keywords: blends, root compounds, word compounds, Portuguese

AUTHORS

ALINA VILLALVA

Universidade de Lisboa, Portugal

RAFAEL DIAS MINUSSI

UNIFESP, Brazil