A *blend* corpus:

steps for an experimentally driven description and analysis of *blends* in EP and BP

Alina Villalva (Universidade de Lisboa/CLUL/The Word Lab)

Email: alinavillalva@edu.ulisboa.pt

Rafael Dias Minussi (Universidade Federal de São Paulo/LabLinC/The Word Lab)

Email: rafael.minussi@unifesp.br

Multi-word formation processes, in Portuguese, comprise root-compounding (e.g., *toxico-dependente* 'drug addict', *agridoce* 'sour sweet'), word-compounding (e.g., barco-casa 'houseboat', *guarda-roupa* 'wardrobe', *cantora-atriz* 'singer/actress'), and blending (e.g., *cartomante* 'liar fortune teller', *tristemunho* 'sad testimony', *cantautor* 'singer and composer'). Compound structures are 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., Sandmann, 1989; Gonçalves 2003a and Gonçalves 2003b).

Our research aims primarily to fully understand the nature of blends and blending in Portuguese. The full program contemplates the following steps: (i) formation and annotation of a blend corpus; (ii) application of an offline test that allows to assess the knowledge of the blends by each subject sample of EP and BP subjects; (iii) application of a lexical decision test that provides MRTs used as a baseline for (iv) the output results of a priming + lexical decision test that will allow to compare the relationship of each of the constituents to the blend.

Presently, we have a corpus of 175 blends, which is based on corpora available in the literature and a direct observation of both European and Brazilian Portuguese sources. This is still a work in progress, since politics, advertising and literature are continuously adding new coinages that we wish to incorporate. This corpus is annotated in order to allow for a refined experimental testing and result analysis. The annotation includes:

(i)   the identification of the constituents of the original blending;
(ii)   the POS of the blend and the blend's constituents;
(iii)   the number of syllables of the blend and the blend's constituents;
(iv)   status of each of the constituents of the blends, as a clip (hence, C) or a word (hence, W), which yields three different structures: CW (e.g., *aminimigo* 'friend and enemy'), WC (e.g., *anãofabeto* 'illiterate dwarf') and CC (e.g., *franglês* 'French and English');
(v)   the grammatical role of each constituent, which yields a distinction between head initial (e.g., *cartomante* 'liar fortune teller'), head final (e.g., *tristemunho* 'sad testimony'), and headless (e.g., *cantautor* 'singer and composer') blends.

Based on a subset of this corpus, we have performed a lexical decision test, with EP and BP subjects. In a previous study (e.g., Minussi and Villalva, 2020), we have analyzed the results based on the status of the constituents (cf. (iv) above), which allowed us to conclude that CC blends

facilitate word processing, since MRTs are significantly lower than those obtained by CW and WC blends. The results of the LD test were very consistent, which was quite surprising and somehow counter-intuitive. Therefore, we decided to reanalyze the results focusing, first, on the prosodic structure of the blends, and, secondly, on their internal syntactic structure. Finally, we will cross-analyze the results, taking into account the wordhood status of the constituents, the prosodic nature of the blend, and the syntactic relation of the blend's constituents.

Ultimately, this analysis will allow us to build a strong claim about the word structure of blends, and their relation to compound structures.