

Filling gaps in dictionary typologies

ROOTS - a morphological historical root dictionary

Keywords: Morphology, roots, Portuguese, etymology, lexicography, *alto*.

The knowledge of the Portuguese lexicon has still many shortcomings. Some of them are well acknowledged; many other are quite unsuspected. Along the 20th century, Portuguese lexicography lost track of the ongoing research on word diachrony, morphological analysis and lexical semantics, unlike what happened with the lexicographic treatment of other comparable modern languages such as French, Castilian or Italian.

Although Portuguese dictionaries generally include information about etymology and morphological structure, they present it quite inconsistently, since it is generally the output of the accumulation of what can be found in previous dictionaries. This *modus operandi* leaves no room to a systematic analysis of the set of words and word families that form each dictionary's entry list. In fact, most contemporary dictionaries of Portuguese (paper editions and electronic versions as well) obey to conservative models, which tend to incorporate exhaustively the information made available by their predecessors, regardless of the real usage of the words. They devote a very considerable amount of effort to increase the entry list, accommodating neologisms (which are not that frequent) and specialized terms, randomly chosen by dictionary makers. From time to time, major dictionary publishers issue remakes of their own dictionaries for the sake of orthographic updates.

Although users tend to be unaware of their deficiencies, this kind of dictionaries is far from being a useful working tool. In fact, since all previously registered words tend to occur in every new dictionary, without any mention to their usage, users are led to induce that they are equally available, and that is not the case. A different type of dictionary, with a critically selected word list and a thorough lexicographic description, is thus still lacking and still necessary. The project we will present in this paper, ROOTS, has been conceived as a further step to build such a dictionary.

The project ROOTS - a morphological historical root dictionary

ROOTS is a research project that aims to produce a prototype of a specialized dictionary that intends to be a useful tool for linguists, lexicographers, translators and language teachers. More specifically, ROOTS is meant to provide a thorough description of word families, considering their presence in the Portuguese lexicon, from a semantic and a morphological point of view, both in diachrony and in their contemporary usage.

The intent to build a thorough morphological historical dictionary for Portuguese is quite striving on the account of feasibility. It requires a program and a step-by-step progression plan. The project ROOTS embodies an initial stage of this program. It is an exploratory project that aims to design a lexicographic model and to test it with roots of simple words¹. These words are made of unanalysable roots and morphological specifiers (thematic suffixes and inflectional suffixes). Once these simple words are etymologically and semantically described, the model will accommodate their derivatives, thus yielding a dictionary of word families. So, the main goals of this project are the clarification of semantic and morphological issues in the evolution of the Portuguese lexicon, and the assessment of the communicative adequacy of the words that are registered in current Portuguese dictionaries, particularly by signalling obsolete words. Furthermore, if the model is replicated with other languages, multilingual dictionaries will become easier to compile and to validate.

The identification of a set of simple roots is crucial to better understand the lexicon since dealing with a smaller set of units potentiates coherence in the treatment of the data. Furthermore, although no such information is independently available, this project is ground on the assumption that the large amount of words that forms the lexicon of a language (namely Portuguese), regardless of their longer or shorter existence in the lexicon, are made out of a relatively small set of roots.

No existing dictionary provides the type of information that ROOTS is designed to include. In fact, most general dictionaries accumulate information from previous dictionaries and they often present inaccurate information concerning etymology and morphological structures. Even those featured as the best (European and Brazilian) Portuguese dictionaries lack systematic morphological description, historical review and usage assessment. As far as specialized dictionaries are concerned, although a better solution would be desirable, *Corpus Lexicográfico do Português* fulfils the needs that should be addressed by an heritage dictionary. Etymological dictionaries have insufficient information and they mirror the state of the art of the mid 20th century, even if they were produced later. Finally, two specialized dictionaries deserve to be mentioned, although they are obviously outdated (either concerning the coverage, or the methodology adopted). The first one is *Dicionário de Raízes e Cognatos* (Goes 1921), which is based on 19th century lexicographic sources – it is mainly a list of neoclassic loans. The second one is the most important Portuguese morphological dictionary, compiled in Brazil. Unfortunately, it gathered lexical information from untrustworthy sources, such as general language dictionaries (cf. Heckler, Back, Massing 1984-1988). They both offer interesting data, but their consultation also requires extensive critical reading.

ROOTS: lexicographic model

¹ Complex lexicalized words behave simple words. The treatment of this kind of words is left to a latter stage of the program, for the sake of methodological choices.

This project aims to build a specialized dictionary containing a critical selection of a *corpus* of simple words, their lexicographic description and a usage labelling focused on contemporary Portuguese.

Critical word list selection

Being a prime source of lexical information, dictionaries can provide a good documentary basis for the collection of a corpus of simple words, and a corpus of the derivatives that contain the same root. In fact, the background for ROOTS is established by a canon of Portuguese dictionaries, from the 16th to the early 20th century, selected for qualitative reasons, since they all played an important role either for the quality of the information they provided or for the normative role that they assumed. At this early stage, we work with a digital corpus comprising the following pre-modern dictionaries: Cardoso 1569, Barbosa 1611, Bluteau 1712, available online (see *Corpus Lexicográfico do Português*, hence CLP), and Figueiredo 1913, also available online (see *Dicionário Aberto*). In the future, this corpus should be extended to include Morais Silva (namely the 1823 edition)².

Roots and derivatives included in our dictionary are strictly those that these sources legitimate. None of the dictionaries produced since the beginning of the 20th century are useful to collect simple words. Typically, the items they add are new derivatives, compounds or loans that can be left out of our search for the time being. Nevertheless, we always consult general contemporary dictionaries such as DLPC, Infopedia, Priberam (European Portuguese) and Houaiss 2009 (Brazilian Portuguese), in order to confirm that these words still have a register in these word lists.

Two written text *corpora* are systematically used: *Corpus de Referência do Português Contemporâneo* (CRPC) allows us to document the occurrence of words in the 20th century; *Corpus do Português* (CdP) is the database that gathers more texts from the 15th century to the 19th century, which is useful information to establish the time span of each word.

Lexicographic description

All simple words are described, following the guidelines of Villalva & Silvestre (2014). The lexicographic description of each word includes an etymologic survey, a morphological analysis and a usage labelling.

Etymological survey

The etymological survey for each root can make use of Machado (1967) or Cunha (1986), which are the reference etymological dictionaries for Portuguese.

² For the time being, Morais Silva can be consulted online at www.brasiliana.usp.br/pt-br/dicionario/.

However, the information they provide often requires further research. The comparison of Portuguese roots with their cognate equivalents in other languages helps to elucidate cases of loan, which are quite commonly found. Therefore, we regularly consult lexicographic *corpora*³. Furthermore, trying to fulfil what is expected of an historical dictionary, we need to sort chronologically the new meanings of polysemic words, through a semantic analysis of textual occurrences.

Morphological analysis

Morphological analysis intervenes to ensure that we appropriately select simple words, like *alto* ‘high, notable’ (and complex lexicalized words, like *altar* ‘altar’, at a later stage), that is, words that contain the root that will define a word family (i.e. *alt-* and *altar-*, respectively). Furthermore, morphological analysis will allow elucidating how the members of a word family (derivatives and modified words) relate to each other: some hang directly on the root, other depend from a simple word, and the remaining words relate a complex word:

- (1) *alt(o/a)* Adjective Root → *altiva* Adjective (fem) → *altivamente* Adverb
 ‘notable’ ‘arrogant’ ‘arrogantly’

Morphological hierarchies of this kind are subordinate to the semantic analysis, but they also help to consolidate it.

Time span and usage labelling

One of the main goals of the project is to identify obsolete words, despite being registered in contemporary dictionaries. In the cases where a derived word does not have a significant occurrence in the CRPC database (inferior to 100, considering that the database has more than 300 million words) or when it is not documented by different textual sources, we will label it as obsolete. Thus, for each word (each meaning of each word, in fact) we establish a tentative time span, according to the registers found in CdP and CRPC. This information is far more stable than the usual identification of a ‘first’ occurrence, which systematically needs to be revised.

Case study: the root *alt-*

³ Namely Tesouro Informatizado da Língua Galega (TILG); Tesouro Medieval Informatizado da Língua Galega (TMILG); *Nuevo Tesoro Lexicográfico de la Lengua Española* (NTLLE); *Le Trésor de la langue française informatisé* (ATILF); *Tesoro della Lingua Italiana delle Origini* (TLIO).

The first stage of ROOTS is devoted to adjectives. So far, we have studied a small number of adjectives that enabled us to test the model of description: *bravo* ‘brave, angry’, *esquisito* ‘weird’, *largo* ‘large’ and *comprido* ‘long’⁴. We have intentionally selected words that have a long existence in Portuguese, which usually enables semantic mutations. Contrastive analysis is quite useful in these cases. In this paper, we will present the study of the root *alto* ‘high, notable’ as a case study and an example of ROOTS work plan, methodology and lexicographic model.

Alto is a very frequent word in Portuguese⁵. *Infopedia* provides the following information:

⁴ See Villalva & Silvestre (2011) for *bravo*, Silvestre & Villalva (2014) and Villalva & Silvestre (in print).

⁵ Davies & Preto-Bay (2008) lists the top 5000 most frequent words in the *Corpus do Português*. *Alto* is at the rank 185.

alto (1)

al.to • [ˈaɫu]

adjetivo

1. que tem extensão vertical; que tem altura
2. que está acima do plano em que se encontra o observador; elevado; subido
3. levantado; erguido
4. profundo, intenso
5. ilustre; eminente
6. importante; grave, sério
7. soberbo; altivo
8. excessivo; caro
9. difícil; transcendente
10. arrojado, destemido
11. afastado no tempo, remoto, longínquo

nome masculino

1. dimensão vertical; altura
2. elevação; monte
3. ponto alto; cume; pináculo
4. saliência, protuberância
5. *figurado* céu

advérbio

1. em voz alta, sonoramente, fortemente
2. a grande altura, em lugar elevado

alto e bom som

claramente, com clareza

alto e malo

1. sem distinção, sem escolha
2. à pressa, atabalhoadamente

altos e baixos

1. elevações e depressões
2. *figurado* momentos bons e momentos maus

de alto a baixo


de cima a baixo, totalmente

em alto grau

multíssimo, extraordinariamente

por alto

sem detalhe, sem minúcia

 Do latim *altu-*, «alto»

alto (2)


al.to • [ˈaɫu]

adjetivo

MÚSICA diz-se do som com frequência elevada; agudo

nome masculino

1. MÚSICA forma reduzida de *contralto*, na aceção 1
2. MÚSICA registo de falsete masculino utilizado na música pré-clássica
3. MÚSICA (*instrumento de cordas*) ver [viola](#)
4. MÚSICA instrumento de sopro da família dos saxofones

 Do italiano *alto*, «idem»

alto (3)

al.to • [ˈaɫu]

nome masculino


ato de suspender o movimento; paragem

alto!

exclamação que se usa para impor paragem ou suspensão de movimento e para exprimir desacordo, pare!, basta!

alto lá!

não diga mais!, basta!

 Do alemão *halt*, imperativo de *halten*, «parar»

In this entry, we find a musical term (*alto* 2), which we will not consider here, and a homonym, which is in fact a different word (*alto* 3), and that is why we will disregard it as well. We will also skip the expressions included in 1, since they deserve a description on their own right, but not here. Thus, we will consider the adjective, the noun and the adverb *alto*, from *alto* 1, which are presented with an equivalent status, all of them being anchored in the Latin *altu-*. This is probably not the intended interpretation of this entry and in fact it is not difficult to find out that no *altu-* noun or *altu-* adverb ever existed in Latin. We could alternatively deduce that these three Portuguese words were derived from the Latin adjective, but no demonstration of that hypothesis has been found and it is quite absurd to admit it, since conversion processes that turn adjectives into nouns and adjectives into adverbs are quite frequent in contemporary Portuguese. The list of meanings (let us focus on the adjective for now) is also quite difficult to interpret: the first meaning uses a derivative (i.e. *altura* ‘height’) to explain its input, but the most difficult part is to decide what kind of hierarchy presides this ordering (frequency, antiquity, etc.). Some of these meanings are difficult to recognize in contemporary Portuguese (cf. *alto* = *profundo* ‘deep’) or independently of a specific collocation (cf. *alto* = *caro* ‘expensive’ vs. *preço alto* = ‘expensive’).

Another issue concerns the location of complex words that contain the root *alt-*. Alphabetically ordered paper dictionaries offer us the possibility to easily trace words beginning with *alt-* but the same is not true for roots that have been prefixed; some electronic dictionaries have a word list searching tool that emulates searching facilities of paper dictionaries, but some don’t, and a few dictionaries allow word searches with wildcards that facilitate the location of the root in non-initial positions. For Portuguese, one such dictionary is the electronic version of the 1913 edition of Cândido de Figueiredo. The availability of this kind of search tools is far from being the solution we need. In fact, no available dictionary for Portuguese includes a morphological encoding that would allow to trace words containing a given root. They just allow to search graphic strings. The output of this search will be a word list that includes words containing the target root (e.g. *altivo* ‘arrogant’) and words containing a homographic string of characters (e.g. *esmalte* ‘enamel’), which have to be discarded.

Figueiredo registers 340 words that contain the sequence *alt* in initial or medial position⁶, but not all of them contain the root *alt-*. A huge time consuming manual selection will set apart words that are related to the root *alt-* from those that randomly contain the sequence <alt>. If we accept the definitions proposed by this lexicographer, we obtain a set of 92 words somehow related to the root *alt-*.

<i>alta</i> _N	<i>altear</i>	<i>altipotente</i>	<i>alto</i>	<i>enaltar</i>
<i>alta-roda</i>	<i>alteável</i>	<i>altirrostro</i>	<i>alto-alegrense</i>	<i>enaltecer</i>
<i>altabaixo</i>	<i>alteza</i>	<i>altíssimo</i> _{ADJ}	<i>alto-alemão</i>	<i>enaltecimento</i>
<i>altabrava</i>	<i>altibaixa, o(s)</i>	<i>altíssimo</i> _N	<i>alto-alentejano</i>	<i>exaltação</i>
<i>altaforma</i>	<i>altícomo</i>	<i>altissonância</i>	<i>alto-beirão</i>	<i>exaltadamente</i>

⁶ This sequence can not occur in final position, since no Portuguese words end in <t>.

<i>altamente</i>	<i>alticornífero</i>	<i>altissonante</i>	<i>alto-comissário</i>	<i>exaltado</i>
<i>altanado</i>	<i>altifalante</i>	<i>altissonantemente</i>	<i>alto-duriense</i>	<i>exaltador</i>
<i>altanar-se</i>	<i>altiloquência</i>	<i>altissono</i>	<i>alto-falante</i>	<i>exaltamento</i>
<i>altanaria</i>	<i>altiloquente</i>	<i>altista</i>	<i>alto-forno</i>	<i>exaltar</i>
<i>altaneiramente</i>	<i>altiloquia</i>	<i>altitonante</i>	<i>alto-minhoto</i>	<i>pernaltas</i>
<i>altaneiro</i>	<i>altilóquio</i>	<i>altitude</i>	<i>alto-navarro</i>	<i>pernalteira</i>
<i>altania</i>	<i>altiloquo</i>	<i>altívago</i>	<i>alto-relevo</i>	<i>pernalteiro</i>
<i>altar</i>	<i>altimetria</i>	<i>altivamente</i>	<i>alto-ribatejano</i>	<i>pernalto</i>
<i>altar-mor</i>	<i>altimétrico</i>	<i>altivar</i>	<i>altor</i>	<i>pernaltudo</i>
<i>altareiro</i>	<i>altímetro</i>	<i>altivez</i>	<i>altosa</i>	<i>planalto</i>
<i>altarista</i>	<i>altimurado</i>	<i>altiveza</i>	<i>altura</i>	<i>ribalta</i>
<i>altavela</i>	<i>altinopolense</i>	<i>altivo</i>	<i>burro-alto</i>	<i>sobreexaltar</i>
<i>alteação</i>	<i>altinopolitano</i>	<i>altivolante</i>	<i>contralto</i>	<i>superexaltado</i>
	<i>altiplano</i>	<i>altívolo</i>	<i>cornialto</i>	

A subset (marked in grey in the table above) is formed by 55 morphological compounds that will not be considered here, since they have involve at least another root (e.g. *planalto* = *plan* + *alt*). Thus, only 39 words are eligible members of the *alt*-root family:

<i>alta_N</i>	<i>altar</i>	<i>altíssimo_{ADJ}</i>	<i>altivo</i>	<i>exaltação</i>
<i>altamente</i>	<i>altareiro</i>	<i>altíssimo_N</i>	<i>alto</i>	<i>exaltadamente</i>
<i>altanado</i>	<i>altarista</i>	<i>altista</i>	<i>altor</i>	<i>exaltado</i>
<i>altanar-se</i>	<i>alteação</i>	<i>altitude</i>	<i>altosa</i>	<i>exaltador</i>
<i>altanaria</i>	<i>altear</i>	<i>altivamente</i>	<i>altura</i>	<i>exaltamento</i>
<i>altaneiramente</i>	<i>alteável</i>	<i>altivar</i>	<i>enaltar</i>	<i>exaltar</i>
<i>altaneiro</i>	<i>alteza</i>	<i>altivez</i>	<i>enaltecer</i>	<i>sobreexaltar</i>
<i>altania</i>		<i>altiveza</i>	<i>enaltecimento</i>	<i>superexaltado</i>

Since Figueiredo is presently a centennial dictionary, we crosschecked this list with the word list of more recent dictionaries (*DLPC*, *Infopedia* and *Priberam*). The comparison brought very few additions (four) and it revealed that not so many words have since been excluded. Crosschecking this final list with frequency data from CRPC proved that a significant amount of its members either have no records or they have a very low frequency, suggesting a very peculiar and restricted usage.

The following table comprises 43 words, including those that come from Figueiredo's list, those that appeared in later dictionaries, as well as their registers in all these dictionaries and their frequency in CRPC:

Figueiredo 1913	DLPC 2001	Infopédia 2014	Priberam 2014	CRPC 2014
<i>alta_N</i>	✓	✓	✓	6.684
<i>altamente</i>	✓	✓	✓	8.211
<i>altanado</i>	✗	✓	✓	0
<i>altanar</i>	✓	✓	✓	0
<i>altanaria</i>	✓	✓	✓	14
<i>altaneiramente</i>	✗	✗	✓	7
<i>altaneiro</i>	✓	✓	✓	43
<i>altania</i>	✗	✓	✗	0
<i>altar</i>	✓	✓	✓	1417
<i>altareiro</i>	✗	✗	✓	0
<i>altarista</i>	✗	✗	✓	0
<i>alteação</i>	✗	✓	✓	0
✗	<i>alteado</i>	✗	✓	28

x	<i>alteador</i>	✓	✓	1	
x	<i>alteamento</i>	✓	✓	55	
	<i>altear</i>	✓	✓	23	
	<i>alteável</i>	x	x	0	
	<i>alteza</i>	✓	✓	297	
	<i>altíssimo</i>	✓	✓	1299	
	<i>altíssimo_N</i>	✓	✓	132	
	<i>altista</i>	✓	✓	166	
	<i>altitude</i>	✓	✓	2.086	
x	x	<i>altitudinal</i>	x	0	
	<i>altivamente</i>	✓	✓	53	
	<i>altivar</i>	x	x	0	
	<i>altivez</i>	✓	✓	180	
	<i>altiveza</i>	x	x	0	
	<i>altivo</i>	✓	✓	174	
	<i>alto_{ADJ}</i>	✓	✓	50.604	
	<i>altor</i>	x	x	0	
	<i>altosa</i>	x	✓	0	
	<i>altura</i>	✓	✓	70.768	
	<i>enaltar</i>	x	x	0	
	<i>enaltecer</i>	✓	✓	675	
	<i>enaltecimento</i>	✓	✓	50	
	<i>exaltação</i>	✓	✓	1095	
	<i>exaltadamente</i>	x	✓	19	
	<i>exaltado</i>	✓	✓	896	
	<i>exaltador</i>	x	✓	4	
	<i>exaltamento</i>	✓	✓	5	
	<i>exaltar</i>	✓	✓	533	
	<i>sobreexaltar</i>	✓	✓	0	
	<i>superexaltado</i>	x	x	0	
	38 words	24 words	33 words	34 words	26 words

The analysis of this data allows us to conclude that contemporary dictionaries have a quite stable entry list since the beginning of the 20th century. Around half of the words in this table (22 out of 43) are present in all of these four dictionaries. According to frequency information from CRPC, two of them are very frequent words (cf. 2a); six other are quite frequent (cf. 2b); eight words have a low frequency (cf. 2c); five have a very low frequency (cf. 2d); and two of them are inexistent in this corpus (cf. 2e).

- (2) a. *altura* CRPC = 70.768
alto/a(s)_{ADJ} CRPC = 50.604
b. *altamente* CRPC = 8.211
alta_N CRPC = 6.684
altitude CRPC = 2.086
altar CRPC = 1.417
altíssimo_{ADJ} CRPC = 1.299
exaltação CRPC = 1.095
c. *exaltado* CRPC = 896
enaltecer CRPC = 675
exaltar CRPC = 533
alteza CRPC = 297
altivez CRPC = 180
altivo CRPC = 174
altista CRPC = 166

	<i>altíssimo</i> _N	CRPC = 132
d.	<i>enaltecimento</i>	CRPC = 50
	<i>altaneiro</i>	CRPC = 43
	<i>altear</i>	CRPC = 23
	<i>altanaria</i>	CRPC = 14
	<i>exaltamento</i>	CRPC = 5
e.	<i>altanar</i>	CRPC = 0
	<i>sobreexaltar</i>	CRPC = 0

Furthermore, five words from Figueiredo's list have been excluded in subsequent dictionaries. None of them occurs in the contemporary *corpus*, which may suggest that they are obsolete words, but tagging words as obsolete requires other tools, as we will see later. This status is probably adequate for four of them (cf. 3a), but not for the last one (cf. 3b). Since *superexaltado* is a superlative adjective, which we do not expect to typically occur in written texts, its absence in CRPC is probably due to the design of the corpus and not to the absence of the word in contemporary European Portuguese. Notice that, unlike those words in (3a), *superexaltado* is a complex compositional word, a kind of word that dictionary makers may decide to include or leave out of the word list.

(3)	a.	<i>altivar</i>	CRPC = 0
		<i>altiveza</i>	CRPC = 0
		<i>altor</i>	CRPC = 0
		<i>enaltar</i>	CRPC = 0
	b.	<i>superexaltado</i>	CRPC = 0

Six other words have been excluded only by the DLPC. Three of them (cf. 4a) have no records in the contemporary *corpus*, which should be interpreted as in the previous case: they may be obsolete words. The remaining two (cf. 4b) are quite infrequent in the corpus, but they are compositional complex words, which typically have a low frequency usage: speakers when needed form them, and their interpretation is fully predictable. As we said above, dictionary makers can choose to include them in the word list, or not, but the decision must be systematically respected. DLPC is not systematic, since it excludes *exaltadamente* and *exaltador*, but it includes *exaltado* and *enaltecimento*, for instance, as all the other dictionaries do, and it also includes *altivamente*, which is excluded only by *Infopedia* (cf. 4c). Words of this kind are not obsolete words:

(4)	a.	<i>altanado</i>	CRPC = 0
		<i>altosa</i>	CRPC = 0
		<i>alteação</i>	CRPC = 0
	b.	<i>exaltadamente</i>	CRPC = 19
		<i>exaltador</i>	CRPC = 4
	c.	<i>altivamente</i>	CRPC = 53

Furthermore, five words from Figueiredo's list fail to be present in DLPC and one of the other dictionaries: *Infopedia* in (5a) and *Priberam* in (5b). These cases are not very different from those just mentioned: either they are obsolete or they don't need a dictionary register. The problem is that all these dictionaries make random choices of what to include and what to let out.

(5)	a.	<i>altaneiramente</i>	CRPC = 7
		<i>altareiro</i>	CRPC = 0
		<i>altarista</i>	CRPC = 0
	b.	<i>altania</i>	CRPC = 0
		<i>alteável</i>	CRPC = 0

Finally, only four different words appear in more recent dictionaries. We might think that these words were neologisms, because these dictionaries are quite recent (or recently updated), but they are not and they are also quite infrequent in the corpus. Once again, these are complex compositional words that do not need to be present in general dictionaries:

(6)	<i>alteado</i>	CRPC = 29
	<i>alteamento</i>	CRPC = 55
	<i>alteador</i>	CRPC = 1
	<i>altitudinal</i>	CRPC = 2

Now that we have a list of words that contain the root *alt-*, we must decide if they are all members of the same word family, or not. Word family is a concept that needs to be handled with care for two main reasons⁷. The first one is related to the fact that there is no consensual understanding of what a word family might include (e.g. compound words, inflectional paradigms, etc.). In this project, word families include simple words (e.g. *alto* 'high'), documented compositional derivatives (e.g. *altura* 'height'; *altivez* 'pride') and documented compositional modified words (e.g. *altíssimo* 'very high'). Consequently, word families in ROOTS exclude (cf. 7a) lexicalized derivatives that will head a family of their own (usually these words are loans) and (cf. 7b) words that are cognates but in contemporary Portuguese they are morphologically unrelated (most frequently, the root has a different form) – they will also head a family of their own:

- (7) a. the derivative lacks a base form:
 **altano* → *altaneiro* 'proud'
 no such suffix is available in contemporary Portuguese:
 alt[itude]* 'altitude'
 alt[ar]* 'altar'
 no such prefix is available:
 **[ex]altar* 'to exalt'
 words that have undergone a semantic shift:

⁷ See Bauer and Nation (1993) for further discussion.

altosa ‘long wool’

- b. *alça* ‘handle’ ← *alçar* ‘to elevate’ < Lat. **altiāre*
outeiro ‘hill’ < Sp. *otero* < Lat. *altarīu*

These exclusion criteria reduce the *alt-* word family to a set of 23 words:

<i>alt-</i> word family	other word families
<i>alta</i> _N	> <i>altaneiro</i> _{ADJ}
<i>altamente</i>	↳ <i>altaneiramente</i> _{ADV}
<i>alteação</i>	> <i>altanaria</i> _N
<i>alteado</i>	> <i>altania</i> _N
<i>alteador</i>	> <i>altanar</i> _V
<i>alteamento</i>	↳ <i>altanado</i> _{ADJ}
<i>altear</i>	> <i>altar</i> _N
<i>alteável</i>	↳ <i>altareiro</i> _N
<i>alteza</i>	↳ <i>altarista</i> _N
<i>altíssimo</i>	> <i>altitude</i> _N
<i>altíssimo</i> _N	↳ <i>altitudinal</i> _{ADJ}
<i>altista</i>	> <i>altosa</i> _N
<i>altivamente</i>	> <i>exaltar</i> _V
<i>altivar</i>	↳ <i>exaltação</i> _N
<i>altivez</i>	↳ <i>exaltado</i> _{ADJ}
<i>altiveza</i>	↳↳ <i>superexaltado</i> _{ADJ}
<i>altivo</i>	↳↳ <i>exaltadamente</i> _{ADV}
<i>alto</i> _{ADJ}	↳ <i>exaltador</i> _N
<i>altor</i>	↳ <i>exaltamento</i> _N
<i>altura</i>	↳ <i>sobreexaltar</i> _V
<i>enaltar</i>	
<i>enaltecer</i>	
<i>enaltecimento</i>	

The second issue related to the concept of word family is that sharing a root is not the only relevant feature wrt to the relationships these words have with each other (cf. Williams 1981). In this project, we decided to show the hierarchy of morphological relationships, in order to distinguish words that are derived or modified from the root (cf. 8a) from words that are derived or modified from previously derived or modified words (cf. 8b):

- (8) a. *alto* → *altura* ‘high → height’
alto → *altíssimo* ‘high → very high’
- b. *altivo* → *altivez* ‘arrogant → pride’
altíssima → *altíssimamente* ‘very high → very highly’

The output of this morphological hierarchy is as follows:

- (9) *alt*_{ADJR}
↳ *alto/a*_{ADJ} CRPC = 50.604
↳↳ *altamente*_{ADV} CRPC = 8.211
↳↳ *alta*_N CRPC = 6.684

↳ <i>altíssimo/a</i> _{ADJ}	CRPC = 1.299
↳ <i>altíssimo</i> _N	CRPC = 132
↳ <i>altivo/a</i> _{ADJ}	CRPC = 174
↳↳ <i>altivamente</i> _{ADV}	CRPC = 53
↳↳ <i>altivez</i> _N	CRPC = 180
↳↳ <i>altiveza</i> _N	CRPC = 0
↳↳ <i>altivar</i> _V	CRPC = 0
↳ <i>altista</i> _{ADJ}	CRPC = 166
↳ <i>altura</i> _N	CRPC = 70.768
↳ <i>alteza</i> _N	CRPC = 297
↳ <i>altor</i> _N	CRPC = 0
↳ <i>altear</i> _V	CRPC = 23
↳↳ <i>alteação</i> _N	CRPC = 0
↳↳ <i>alteado/a</i> _{ADJ}	CRPC = 29
↳↳ <i>alteador</i> _N	CRPC = 1
↳↳ <i>alteamento</i> _N	CRPC = 55
↳↳ <i>alteável</i> _{ADJ}	CRPC = 0
↳ <i>enaltar</i> _V	CRPC = 0
↳ <i>enaltecer</i> _V	CRPC = 675
↳ <i>enaltecimento</i> _N	CRPC = 50

Now that the word family is defined, we must evaluate the usage of each of its members. Frequency figures (from CRPC) shed some light on the set of words that are more or less generally used, but used or rarely used words require further research, since they do not form an homogeneous set. This survey can help to set apart the set of words that have a real existence from those that have a merely fictitious survival in dictionaries, as a result of editorial options rather than a linguistic validation.

Words that were used in past synchronies, but are no longer in use will be marked as obsolete (cf. 10a); words that have a dictionary register but no matches in the corpus will be marked as undocumented (c. 10b). The first set includes old words and also, quite frequently, phonetic or morphological old variants of words that are still available; the second set includes possible words and morphological duplicates.

- (10) a. *altiveza* ‘pride’
altor ‘hight’
cf. *altivez* ‘pride’
cf. *altura* ‘hight’
- b. *altivar* ‘to make arrogant’
alteável ‘that can be elevated’
alteador ‘that elevates’
alteação ‘elevation’
enaltar ‘to make high’
cf. = *alteamento*
cf. = *enaltecer*

The last step is probably the most decisive to the design of our dictionary prototype. Simple words often have a polysemic character, often related to the fact that they have long been present in the lexicon. New meanings may accumulate with old meanings, and some old meanings may also become obsolete. The morphological hierarchy presented in (9) does not incorporate semantic information, which is an essential component of this lexicographic model. The list of meanings that can be found in the source dictionaries often includes proper meanings and meanings that depend on collocations, as well as meanings that correspond to common usage and other that

don't. Hence, we decided to extract the relevant meanings of each simple word that can be documented in a text corpus, namely *Corpus do Português*.

Once all the meanings of a given root have been deciphered, we can link each complex word to the relevant meaning and, thus, doublecheck the semantic analysis as well as the chronological marking. In the case of *alt-*, the output of the morphosemantic analyses, and a brief etymological description, is as follows:

alt- < Lt. *altu*-ADJ (de *altu-* 'alimentado, crescido', participio de *alo*, *-ere* 'alimentar')

1. <i>alto/a</i> _{ADJ}	profundo 'deep'	13th - 18th
<i>altíssimo/a</i> _{ADJ}	muito profundo 'very deep'	17 th
<i>altamente</i> _{ADVloc}	profundamente 'deeply'	15 th - 17 th
<i>altura</i> _N	profundidade 'depth'	13 th - 19 th
<i>alto</i> _N	mar profundo 'deep see'	14 th -
2. <i>alto/a</i> _{ADJ}	grande 'tall'	13th -
↳ <i>alto</i> _{ADV}	(num tom) elevado 'loudly'	16 th -
↳ <i>altamente</i> _{ADVintens}	muito 'very'	17 th -
<i>altíssimo/a</i> _{ADJ}	muito alto 'very tall'	19 th -
<i>alto</i> _N	cimo 'top'	15 th -
<i>alta</i> _N	cimo 'top'	13 th -
<i>alta</i> _N	subida 'rise'	19 th -
↳ <i>altista</i> _{ADJ}	em subida 'rising'	20 th -
<i>alta</i> _N	permissão 'consent'	18 th -
<i>alteza</i> _N	qualidade do que é alto 'height'	14 th , 18 th
<i>altura</i> _N	qualidade do que é alto 'height'	15 th -
<i>altear</i> _V	tornar mais alto 'make high(er)'	19 th -
3. <i>alto/a</i> _{ADJ}	ilustre 'notable'	14th -
<i>altíssimo/a</i> _{ADJ}	muito ilustre 'very notable'	17 th -
↳ <i>altíssimo</i> _N	deus 'god'	15 th -
<i>altamente</i> _{ADV}	de um modo ilustre 'in a notable way'	14 th -
<i>altivo</i> _{ADJ}	superior (positivo) 'superior (positive)'	16 th -
↳ <i>altiveza</i> _N	qualidade do que é superior (positivo) 'quality of being superior (positive)'	17 th -
↳ <i>altivez</i> _N	qualidade do que é superior (positivo) 'quality of being superior (positive)'	17 th -
↳ <i>altivo</i> _{ADJ}	superior (negativo) 'superior (negative)'	17 th -
↳ <i>altiveza</i> _{SUBS}	qualidade de ser superior (negativo) 'quality of being superior (negative)'	16 th - 17 th
↳ <i>altivez</i> _{SUBS}	qualidade de ser superior (negativo) 'quality of being superior (negative)'	18 th -
<i>alteza</i> _N	qualidade do que é ilustre 'quality of being notable'	14 th - 18 th
↳ <i>alteza</i> _N	pessoa ilustre 'notable person'	15 th -
<i>altura</i> _N	'quality of what is notable'	14 th -
<i>enaltar</i> _V	tornar notável 'to make notable'	19 th
<i>enaltecer</i> _V	tornar notável 'to make notable'	19 th -

So, the adjective *alto* has three meanings. Meaning 1 and 2 have older registers, but only the second remains available. Meaning 3 corresponds to a metaphorical diversion from meaning 2. Notice that the first meaning of *alto*, which is 'deep' is lost in contemporary usage (although we can not deduce this from CRPC frequencies, which do not differentiate meanings). The only exception is *alto*_N, meaning 'deep see', which is lexicalized, and still available. Finally, derivatives may be exclusive from a given meaning (cf. *alto*_N 'deep see', from meaning 1; *altista*_{ADJ} 'rising', from meaning 2; *enaltecer*_V 'to make notable', from meaning 3), but they may also be polysemic, along the lines that the root is (cf. *altura*_N 'depth', 'height', 'importance').

Conclusion

To our knowledge, ROOTS is an innovative project, which includes a new lexicographic model. Its main purpose is to provide an adaptive tool for the description of the core lexicon of a given language. Portuguese is a good case study, because there is an obvious deficit on such kind of specialized dictionaries, but it is also interesting because of its position in the space of romance languages. In fact, the lexicographic model designed for ROOTS facilitates the detection of contrastive issues, such as the direction of loans and semantic divergencies.

References

- L. Bauer & P. Nation (1993) Word Families. *International Journal of Lexicography*, Vol. 6 No. 4 (253-279).
- C. Goes (1921) *Dicionário de Raízes e Cognatos da Língua Portuguesa*. Belo Horizonte: Paulo Azevedo & Cia.
- E. Heckler, S. Back, E.R. Massing (1984-1988) *Dicionário Morfológico da Língua Portuguesa*. São Leopoldo: Unisinos.
- J. P. Silvestre & A. Villalva (2014). A morphological historical root dictionary for Portuguese. In A. Abel, C. Vettori & N. Ralli, eds. (2014) *Proceedings of the International Congress EURALEX XVI: The User Focus*. Bolzano (967-978).
- J. P. Silvestre & A. Villalva (in print). Mutations lexicales romanes: *esquisito, bizarro et comprido*. InnTrans: *Innsbrucker Beiträge zu Sprache, Kultur und Translation*. Frankfurt am Main: Peter Lang Verlag.
- A.Villalva & J. P. Silvestre (2011) De *bravo* a *brabo* e de volta a *bravo*: evolução semântica, análise morfológica e tratamento lexicográfico de uma família de palavras. *ReVEL* 9, 17. Online: www.revel.inf.br/files/artigos/revel_17_de_bravo_a_brabo.pdf [09/10/2014]
- A.Villalva & J. P. Silvestre (2014) *Introdução ao Estudo do Léxico. Descrição e Análise do Português*. Petrópolis: Vozes.
- E. Williams (1981) On the notions 'lexically related' and 'head of a word'. In *Linguistic Inquiry* Vol. 12, No. 2 (245-274).
- CLP - *Corpus Lexicográfico do Português*. Universidade de Aveiro - Centro de Linguística da Universidade de Lisboa Online: clp.dlc.ua.pt
- CRPC - *Corpus de Referência do Português Contemporâneo*. Centro de Linguística da Universidade de Lisboa. Online: www.clul.ul.pt/pt/recursos/183-reference-corpus-of-contemporary-portuguese-crpc
- CdP - *Corpus do português: 45 million words, 1300s-1900s*. Online: www.corpusdoportugues.org [10/09/2014].
- C. de Figueiredo (1913). *Novo Dicionário da Língua Portuguesa*. Porto: Typ. da Empr. Litter. e Typographyca.
- A. Houaiss & M. Villar, (2009). *Dicionário Houaiss da Língua Portuguesa*. Rio de Janeiro: Objetiva.

Infopédia. Dicionário da Língua Portuguesa da Porto Editora. Online: www.infopedia.pt/lingua-portuguesa/ [10/09/2014].

TLFI - Le Trésor de la langue française informatisé. Online: atilf.atilf.fr/tlf.htm [10/09/2014].

NTLLE - Nuevo Tesoro Lexicográfico de la Lengua Española. Online: ntlle.rae.es/ntlle/SrvltGUILoginNtllle [10/03/2014].

Vocabulário Ortográfico do Português. Online: www.portaldalinguaportuguesa.org/vop.html [10/03/2014].

TILG - Tesouro Informatizado da Língua Galega

TMILG - Tesouro Medieval Informatizado da Língua Galega

TLIO - Tesoro della Língua Italiana delle Origini

Cardoso 1569, Barbosa 1611, Bluteau 1712, available online (see *Corpus Lexicográfico do Português*, hence CLP), and Figueiredo 1913, also available online (see *Dicionário Aberto*). In the future, this corpus should be extended to include Morais Silva (namely the 1823 edition)