Lexical analysis of pre and post revolution discourse in Portugal

Michel Généreux, Amália Mendes, L. Alice Santos Pereira, M. Fernanda Bacelar do Nascimento

Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal

Abstract

This paper presents a lexical comparison of pre (1954-74) and post (1974-94) revolution parliamentary discourse in four comparable sub-corpora extracted from the Reference Corpus of Contemporary Portuguese (CRPC). After introducing the CRPC, including annotation and meta-data, we focus on a subset of the corpus dealing with parliamentary discourses, more particularly a time frame of forty years divided into four comparable sub-corpora, each covering a ten-year period, two pre revolution and two post revolution. We extract lexical density information as well as salient terms pertaining to each period to make a comparative evaluation of the periods. Our results show how a linguistic analysis essentially based on the use of simple n-gram statistics can produce key insights into the use, change and evolution of the Portuguese language around a critical time period in its history.

1. Introduction

This paper presents the Reference Corpus of Contemporary Portuguese and how we use it to compare the lexicons of pre (1954-74) and post (1974-94) revolution parliamentary discourse in Portugal. The question we are addressing is to what extent a political change (Portuguese revolution in 1974) is reflected in a change in vocabulary usage in speeches of the national assembly. The period covered by our corpora, 1954-1994, was chosen in accordance to the political situation in Portugal over this period of time. The date of the revolution, April 25th 1974, marks a deep change in the political regime, when a dictatorship who lasted almost 50 years was replaced by a democratic state. The first period, from 1954 to 1963, follows the second World War and brought some innovation, but is mainly marked by the McCarthyism, appreciated by the dictator Salazar, and also by the beginning of the war for liberation in the African territories occupied by Portugal. The following period, from 1964 to 1974, was dominated by the colonial wars, especially in Angola, Guinea-Bissau and Mozambique, and had a very negative imprint on the Portuguese population, leading to an increase in revolutionary activities, and an increasingly violent repression by the political police of the regime. During these two periods, there were no free elections and the political regime was based on the autocratic power of Salazar. The Parliament, called at the time *Assembleia Nacional*, could only discuss the legislation proposals of the Government, and political parties, like the Communist and Socialist Parties, could not openly exercise their activities. After the revolution, in 1974, there was a strong rupture with the ideology of the dictatorship, the colonial wars ended abruptly and the African colonies gained their independence. The political parties were legalized and the first free election took place exactly one year after the revolution. In the period between 1984 and 1994, the Portuguese State entered into a stable democracy and joined the European Community.

1.1. The CRPC corpus

The CRPC is the result of numerous efforts at the Centro de Linguística da Universidade de Lisboa (CLUL)[1] to produce an electronically based linguistic corpus containing, after cleaning, 301 million tokens, taken by sampling from several types of written texts (literature, newspapers, science, economics, law, parliamentary debates, technical and didactic documents, etc.) as well as spoken texts, both formal and informal (2,5M tokens). These samplings pertain to national and regional varieties of Portuguese European Portuguese and also Portuguese spoken in Brazil, in the countries where Portuguese is the official language (Angola, Cape Verde, Guinea-Bissau, Mozambique, São Tomé and Principe, East-Timor), and in Macao and Goa. From a chronological point of view, our corpus contains texts from the second half of the XIX century up until 2008, albeit mostly after 1970 (**?**; **?**). Therefore, the CRPC is very-well suited for comparative studies.

The compilation of the CRPC started in 1988 and its main goals are to keep an up-to-date and balanced version of the corpus, disseminate information related to it and make it available on-line so that the resource is friendly and easily accessible. The CRPC is a resource and knowledge database made of authentic linguistic documents, organized in an electronic format accessible to researchers, teachers, and translators and to all, national and foreign, working on the Portuguese language to whom there is a need for reliable linguistic data. These specific linguistic resources constitute an essential prerequisite for a large number of research projects and several types of development and applications.

Two examples of contrastive studies based on comparable corpora partially extracted from CRPC are the results obtained under the scope of the projects VARPORT-Contrastive Analysis of Portuguese Varieties[2], and African Varieties of Portuguese[3], both on the analysis of geograph-

[1] `http://www.clul.ul.pt/english/sectores/linguistica_de_corpus/projecto_crpc.php`

[2] VARPORT is a joint project of CLUL and UFRJ-Rio de Janeiro, Brazil `http://www.letras.ufrj.br/varport`

[3] `http://www.clul.ul.pt/english/sectores/`

ical varieties of Portuguese and, in the case of VARPORT, combining a diachronic approach (**?**; **?**).
Once the corpus collected, our methodology for segmenting and processing the corpus follows widely accepted principles and its recent update is largely inspired by (**?**). The written corpus, which is the focus of this paper, contains 368k files, a large number of them extracted from potentially noisy web sources (html, asp, sgml, php).

1.2. CRPC Meta-data

The richness of the meta-data included in the CRPC allows us to select a subset of documents suitable for our comparative study. Here we describe the main meta-data in some detail to give insights in the variety of information available and how the CRPC can be tailored to our needs. Each document in the CRPC is first classified according to a broad categorization distinguishing written from spoken materials, to which a specific set of meta-data applies. Written texts are classified in terms of analytic meta-data regarding source, text type (book, review, newspaper, parliamentary discourses, etc.) and topic. For each major type a particular combination of text-descriptive features is assigned: for example, the set of descriptive meta-data for newspapers includes information on the sections, while for didactic books it covers the course name and the curricular year. Other general descriptive meta-data address a set of bibliographic information like title, editor, country of edition, date of edition and the author's name. Since the corpus covers different time periods and national varieties of Portuguese, a set of descriptive meta-data give detailed information on the year and country of birth of the author, as well as on his first language and on the country whose variety he represents (for example, some authors born in Portugal and whose first acquired variety might be European Portuguese are in fact living in Mozambique and their works are to be classified as pertaining to the Mozambique variety in the corpus). Other descriptive meta-data focus on the file properties: its name, size in tokens, location in the corpus directories. And finally editorial meta-data describe the status of the file in terms of its correction and normalization (e.g, there are two levels of correction for texts that are digitalized: corrected and revised). The meta-data are stored in an Excel database and have recently been revised regarding the main fields. The meta-data will soon migrate to a MySQL database.

1.3. Cleaning the corpus

The CRPC has been cleaned using the publicly available software *NCLEANER* (**?**). In the first step, *NCLEANER* removes HTML tags and produces segments, essentially paragraphs. To remove segments which contains unwanted texts (boilerplate, announcements, spam, etc), the second step requires a language model that can be produced by training the system on a relatively small number of annotated documents (with *good* and *bad* segments). We have trained *NCLEANER* on 200 documents selected randomly in the CRPC. The segments produced by the first *NCLEANER* step were annotated as being either *good* or *bad*. This resulted in 4986 *good* segments and 1474 *bad* segments, that we used to train *NCLEANER* with no text normalization (-m 0) to preserve accented characters. The language model created was evaluated using ten-fold cross validation on all 6460 annotated segments, obtaining a F-score of 90%. This language model was used to clean the entire CRPC, which shrank from 433 to 301 millions token. This cleaned corpus was used for our diachronic study. The cleaned corpus has been POS annotated with Treetagger(**?**).

linguistica_de_corpus/projecto_variedades_africanas.php

2. Experiments: diachronic variation of Portuguese around the revolution

In this section we present the experiment aiming at discovering how political and social turmoil initiated by the revolution in 1974 in Portugal changed the discourse in parliamentary sessions of the Portuguese national assembly. The general idea is to compare sub-corpora representing parliamentary discourses in four consecutive decades around the revolution in Portugal with one reference corpora (RC). In what follows we first describe the sub-corpora of the CRPC used, then the approach adopted and finally the results.

2.1. The sub-corpora

The CRPC corpus includes parliamentary speeches from the 19th and the 20th centuries. To examine changes that occurred in the parliamentary sessions at the time of the revolution, we have limited ourselves to a period of 40 consecutive years, spamming from 1954 to 1994. In order to make a pre/post revolution comparison, the 40 years were divided into 4 ten year periods with an approximately equal number of tokens: 1954-63 (PER1), 1964-74 (PER2), 1974-84 (PER3) and 1985-94 (PER4). The 1974 split was made on April 25th, when the dictatorship ended. Each of the four 10-year periods was made of a random selection of parliamentary speeches from the CRPC pertaining to that period. A reference corpus (RC), built from a random selection of files pertaining to the written CRPC that originates from Portugal, serves as a basis for the comparisons. It must be said that because the CRPC itself has more documents after than before the revolution, the RC also shares this characteristic, which means that even though it does not affect the interpretation of relative values, absolute values should be interpreted with caution. Table 1 gives more detailed information about the corpora, where there appears to be no significant distribution discrepancy between the reference corpus (RC) and the four sub-corpora.

Info	RC	PER1	PER2	PER3	PER4
Nb doc.	10k	6k	6k	7k	8k
Types	116k	70k	73k	61k	58k
Tokens	5768k	3643k	3698k	3589k	3552k
V	16%	14%	14%	17%	17%
ADV	5%	5%	5%	5%	5%
NOM	32%	31%	30%	33%	34%
ADJ	8%	10%	11%	7%	7%

Table 1: Lexical density

2.2. The approach

Table 1 already provides some useful information to compare the periods. However, providing a more exhaustive comparison requires an analysis of the statistics of words and multi-word expressions (MWs in short), as in (**?**). Statistics about words are rather straightforward to collect, while MWs and statistics about them are more challenging to acquire. We investigated two methods for MW extraction. We did not lemmatize the texts, so we performed extraction directly on lexis. Our first approach is based on the one presented in (**?**). The advantage of this approach is that it can generate automatically a list of MWs with little supervision. This approach first uses the log odds ratio measure to compare token and document frequencies between a target corpus (here one of the four periods) and a reference corpus (here RC) producing a list of candidate unigrams. Then a list of connectors is collected from the reference corpus by looking for words and bigrams that frequently occur between the candidate unigrams (e.g. *de*, *a*, *para o*) which are subtracted from the list of stop-words for Portuguese. Starting with bigrams, a procedure is applied recursively to find MWs that must satisfy a number of linguistic (they contain at least one candidate unigram but no stop words and no connectors at the edges or adjacent to each other) and statistical (they satisfy a threshold of frequency and cannot be part of a longer term with frequency close to their own) constraints. Applying this method to each of our four periods provided us with a list of MWs of variable quality[4], although a vast majority of them has a positive log-odds ratio (salience). As this automated list did not provide enough information to form a solid basis for a comparative evaluation of the four periods, we turned our efforts towards the extraction of all n-grams ($n < 6$) from the texts of the periods, the only constraint being that no stop-words should appear at the edge of the n-grams. We also use the log odds ratio as a statistical measure of salience or prominence of a MW, so the salience for each expression was then computed and sorted from the highest to the lowest. Those lists were then inspected by humans top-down, so that potentially more informative expressions were examined first. The BootCat (**?**) tool was used and adapted for our needs in both approaches. Other approaches generating lists of keywords are possible, for example (**?**).

2.3. Results and Analysis

2.3.1. N-grams sorted by salience

The results of the n-grams lists sorted according to salience provided most information for identifying significant word forms or MWs in the different periods. For example, the adjective *ultramarina* 'overseas' which qualified territories ruled by Portugal but which were located outside its European frontiers, as well as services or institutions in those territories, show relatively high salience in corpora 1 and 2 (values 3.9 and 3.8, respectively). The same is true for MWs related to Portugal's colonies, which gained their independence after the revolution of 1974, like *territórios ultramarinos* 'overseas territories' (3.6 and 3.3) and, indirectly, *missão civilizadora* 'civilizing mission' (5.8 and 3.3) and for MWs related to the concept of 'corporatism', defended by the regime before 1974, like *nosso corporativismo* 'our corporativism' (5.9 and 4.3), *enquadramento corporativo* 'corporatist frame' (5.5 and 2.5). Other MWs are, on the contrary, highly salient in corpus 3, from 1974 to 1984, like *democraticamente eleitos* 'democratically elected' (5.2), related to the new democratic state, and *prédios nacionalizados* 'nationalized buildings' (2.8), evoking the high number of nationalization of industries and holdings after the revolution.

2.3.2. Diachronic contrast in collocational profile

After identifying a unit as significant based on salience, in many cases a more detailed analysis showed a different collocational profile (**?**) of the lemma for each period. This collocational profile can be analysed in certain cases as related to semantic prosody, i.e., the notion that words associate with collocates that belong to a specific semantic set and that particular collocations receive specific attitudinal semantics. We will discuss three cases, *comunista* 'communist', *democracia* 'democracy' and its derived forms, and the adverb *publicamente* 'publicly'. The collocates of *comunista* revealed quite opposite perspectives regarding this ideology. In corpora 1 and 2 we find *bloco comunista* 'communist bloc', *China comunista* 'communist China' and *propaganda communist* 'communist propaganda', while in corpora 3 and 4 we encounter *Juventude comunista* 'communist Youth', *Partido Comunista* 'Communist Party', *comunistas portugueses* 'Portuguese communists', *nós comunistas* 'we communists', among many others. The collocational profile in the first two periods reflects an ideology which is alien to the Portuguese state at that point in time and contrasts deeply with the high level of involvement of the later collocates. A similar contrast is found with the noun *democracia* and the adjective *democrático* 'democratic'. The two occur in exactly two MWs in corpora 1 and 2: *chamadas democracias* 'so called democracies' (salience 3.3), *totalitarismo democrático* 'democratic totalitarianism' (salience 1.01), but are highly frequent in MWs in corpora 3 and 4, and, as the following selection shows, with a quite different semantic prosody: *regras democráticas* 'democratic rules', *ética democrática* 'democratic ethic', *governo democrático* 'democratic government', *sociedades democráticas* 'democratic societies', *escola democrática* 'democratic school', *jovem democracia* 'young democracy', *democracia avançada* 'advanced democracy', *democraticamente legitimados* 'democratically legitimized'. Finally, in corpora 3 and 4, the adverb *publicamente* 'publicly' occurs in the highly salient MW *aqui publicamente* 'here publicly' (which is part of a larger expression starting with a declarative verb: *declaro/afirmo aqui publicamente* 'I declare/affirm here publicly') and is very productive in collocations of the last two periods: *denunciar pub-*

[4]One problem is the overly significant number of proper names. In fact, (**?**) reports a precision of 73% with a recall of 68% for English, and a precision of 32% with a recall of 5% for Italian. However, the quality of the texts on which extraction was performed was somewhat lesser than that of the CRPC, not the least because the English and Italian texts were all harvested from the web.

licamente 'to denounce publicly', *anunciar publicamente* 'to announce publicly', *assumidas publicamente* 'assumed publicly', etc. But there are no well formed MW with this adverb in the first two corpora. The diachronic contrast that we observed in the collocations behaviour answers the need for more diachronic studies of semantic prosody: "A diachronic approach, on the other hand, could try to establish how the meaning of the unit changes over the years or centuries, or it could investigate how words bestow meanings upon each other over time within that unit" (**?**) and is an interesting subject to further explore in the analysis of the 4 comparable corpora.

2.3.3. N-grams sorted by differential

The n-grams lists with statistical measures for salience, produced for the period before and after the revolution, give us interesting results, as we can see from the examples above. However, our search for significant MWs in the n-grams lists has shown that their salience values are not necessarily very high in any of the four periods: for example, *causa nacionalista* 'nationalist cause' (1.01 and 1.12), *nacionalismo* 'nationalism' (-0.1 and -0.7), *colónia* 'colony' (0.0 and -1.7). The analysis of the full data shows that it is important to look not only to the salience in each period, but mainly to the contrast between saliences in periods 1-2 and in periods 3-4. Under this contrastive approach, the word form *colónia*, with low salience for periods 1 and 2, becomes much more prominent because its salience decreases significantly in the last period, when Portuguese colonies had gained their independence, and the same accentuated decrease is true for the MWs *colónia portuguesa* 'Portuguese colony' and *territórios ultramarinos* 'overseas territories' (see Figure 1). Other significant word forms in periods 1 and 2 show moderate salience, but a strongly contrastive behaviour over the four periods, like the noun *corporação* 'corporation' and the feminine adjective *católicas* 'catholics' (see Figure 2).

Instead of looking for significant MWs in one or more periods, a different approach is to produce lists of word forms which do not occur in corpora 1 and 2 and do occur in 3 and 4. This immediately singles out new word forms appearing after the revolution, like *Parlamentar* 'Parliamentary', *Constitucionais* 'constitutionals', *Liberdades* 'liberties', *Esquerda* 'Left', as well as terms designing new political parties (*PSD*, *CDS*, *PCP*, *Socialista*), and terms for new concepts like *computador* 'computer' and *euros* 'euros'. The results obtained showed that the first two periods shared a common lexicon, just as the last two periods, and that the best approach was to contrast the salience in the first two corpora with the last two. This led us to produce new statistics sorted according to the difference between saliences 1-2 and 3-4, which we call differential (so differential = S1+S2-S3-S4). The top of the list highlights MWs with a strong contrast between a high salience in 1-2 and a low one in 3-4, and the bottom of the list shows exactly the opposite. An intuitively significant word form in corpora 1-2, like *indígenas* 'indigenous', is difficult to single out based on its individual salience in each period, yet it receives a high differential of 15.5. We present in Table 2 a sample of the top of the list for unigrams and bigrams and in Table 3 a sample of the bottom of the list for unigrams and bigrams: the first column is the differential, followed by columns for the salience in periods 1, 2, 3 and 4. At the top of the unigrams list are word forms like *metrópole* 'metropolis', *Corporativa* 'Corporative', *ultramar* 'Portuguese colonies', *províncias* 'provinces', very significant terms under the dictatorship, while at the bottom are word forms like *Democracia* 'Democracy', *comunista* 'communist', *parlamentar* 'parliamentary', *quórum* 'quorum', highly representative of the politics after the Revolution. This is certainly the most promising approach for the analysis of the four corpora.

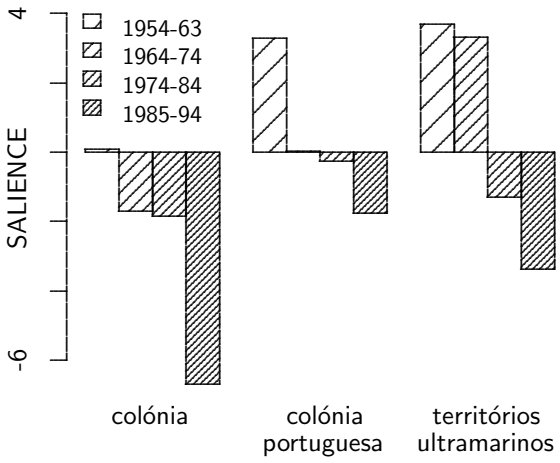

Figure 1: Diachronic behaviour of *colónia*, *colónia portuguesa* and *territórios ultramarinos*

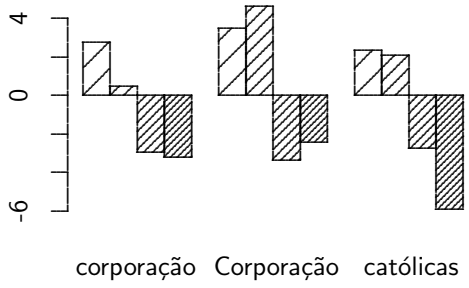

Figure 2: Diachronic behaviour of *corporação* and *católicas*

2.3.4. Other diachronic contrastive patterns

The analysis of the results has mostly showed a strong contrast in lexicon between periods 1-2 and periods 3-4, delimited by the revolution of April 74. Further analysis also reveals other lexical patterns distinguishing period 3 from

Diff.	S1	S2	S3	S4
unigrams				
metrópole 'metropolis'				
24.21	4	4	-8	-9
Corporativa 'Corporative'				
24.08	4	3	-7	-10
ultramar 'Portuguese colonies'				
22.41	4	4	-7	-7
províncias 'provinces'				
22.36	4	4	-7	-9
Colonização 'Colonization'				
16.13	4	2	-4	-6
ultramarinos 'overseas'				
15.41	4	4	-3	-5
bigrams				
Câmara Corporativa 'Corporate Board'				
23.93	4	3	-7	-10
Educação Nacional 'National Education'				
21.66	4	4	-6	-8
ordem administrativa 'administrative order'				
18.73	5	2	-6	-5
Previdência Social 'Social Security'				
18.44	4	3	-6	-6
espaço português 'Portuguese space'				
16.85	1	5	-5	-5
Fomento Nacional 'National Development'				
16.81	4	2	-6	-4

Table 2: Differential: salience high in 1-2 and low in 3-4

Diff.	S1	S2	S3	S4
unigrams				
Democracia 'Democracy'				
-17.63	-6	-8	3	1
deputado 'deputy'				
-18.43	-8	-8	1	1
CEE 'EEC'				
-19.26	-10	-9	0	1
Comunista 'Communist'				
-21.24	-9	-11	2	0
abstenções 'anstentions'				
-22.56	-8	-11	2	2
Parlamentar 'Parliamentary'				
-23.05	-13	-8	2	1
quórum 'quorum'				
-24.78	-10	-9	2	3
bigrams				
sociedade democrática 'demoocratic society'				
-11.34	-5	-5	2	0
pré escolar 'preschool'				
-12.09	-5	-7	0	1
salário mínimo 'minimum wage'				
-12.31	-7	-5	1	0
partidos políticos 'political parties'				
-14.35	-7	-5	2	0
vamos votar 'we will vote'				
-20.61	-10	-7	1	2
Partido Comunista 'Communist Party'				
-21.48	-9	-11	2	0

Table 3: Differential: salience low in 1-2 and high in 3-4

period 4. The salience of many MWs first decreases from 1 to 2, then increase in 3 and decrease again in 4, as exemplified by *reforma agrária* 'agrarian reform', *Democracia* and *Comunista* in Figure 3. These cases point to an abrupt change in the Parliamentary lexicon related to an equally abrupt political event in 1974, with new parties, ideologies and their application in society, but also to a progressive decrease of radicalism in period 4, when the Portuguese society settled in a stable democracy. In the case of *guerra colonial* 'colonial war', a slightly different pattern appears, with a small increase of salience in 2, followed by a strong increase in 3 and a final decrease in period 4 (see Figure 4). The last observation can be explained by the ongoing wars and the gain of independence. In some very specific cases, there is no contrast between 1-2 and 3-4, but rather between the first three periods and period 4, starting in 1985. An example is the masculine and feminine form of the adjective *Europeu, Europeia* 'european' and the currency *euro* (see Figure 5). This pattern is not related to the pre and post revolution, but instead to the integration of Portugal in the European Community in 1985.

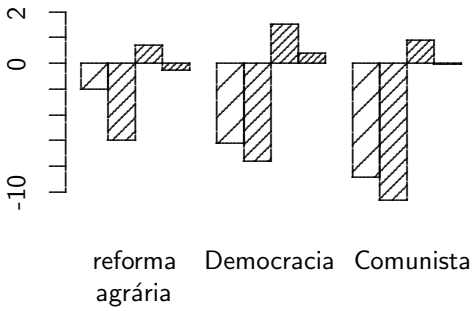

Figure 3: Diachronic behaviour of *reforma agrária*, *Democracia* and *Comunista*

The cases discussed above were either salient in one or two corpora or prominent in terms of differential and could be easily recognized as pertinent because they refer to political realities well known by the Portuguese population. This allowed us to evaluate different approaches towards the identification of lexical units undergoing change. But among the more prominent lexical units we also found words or expressions which seemed to us less obviously representative of any of the periods under study. For example, the high number of terms related to agriculture, heavy industries and sports in corpora 1 and 2 requires a collaborative study with experts in recent Portuguese history, sociology and political science. The specialized nature of our corpora requires a terminological approach: these four corpora are not rep-

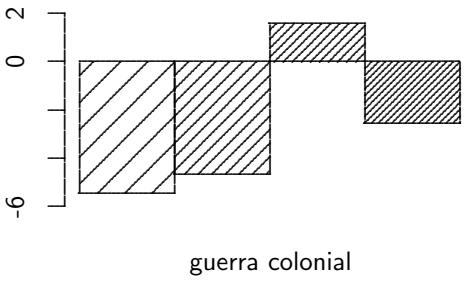

Figure 4: Diachronic behaviour of *guerra colonial*

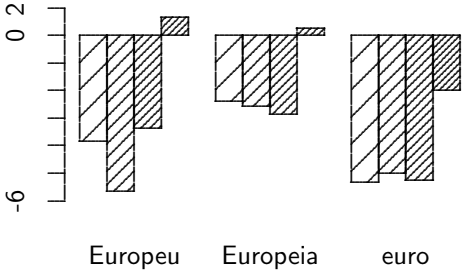

Figure 5: Diachronic behaviour of *Europeu*, *Europeia* and *euro*

resentative of how people talked in the streets or wrote in the newspapers before and after the revolution, but instead are representative of how members of the Parliament expressed themselves in these periods. Statistical information extracted from the corpora is not always in accordance to our intuition as to how a specific word should behave: for example, in the case of *colónia* and *colónia portuguesa* the fact that these MWs, whose usage was disapproved before the revolution (the official term being *província ultramarina*), have the same salience in periods 2 and 3 is unexpected and needs further investigation.

3. Conclusion and Future work

We have presented the Portuguese corpus CRPC and the challenges we met in preparing and organising the corpus. After some work on cleaning the CRPC, we explored the diachronic variation of Portuguese during the revolution through careful inspection of an exhaustive list of n-grams. Our main findings are that the most effective method to identify salient lexical units is to compare the four corpora, either by contrasting the lexicon which only occurs in the pre or the post revolution periods, or, and this gives even more interesting results, by using the differential values in corpora 1-2 and 3-4. A follow-up is to contrast the collocates of MWs which are significant in one of the periods. This methodology pointed to lexical units undergoing strong diachronic variation during the periods under study. Several patterns of change were identified: in many cases, the contrast is between the first two corpora and the last two, but the presence of other significant lexical units have shown that there are significant differences in lexical behaviour between the two corpora pre-revolution and even more significant ones between the two periods following the revolution. Future work should look into the optimization of the identification of significant word forms in each period and consider an interdisciplinary approach with social sciences (politics, history, sociology) and law to fully explore the lexicon. Contrasts between the two pre-revolution periods should be fully explored to identify shifts in the dictatorship's ideological and social politics in a time where resistance gained influence. Likewise, a more in depth analysis is required to evaluate lexical changes between period 3, when the revolutionary process gains its full expression, and period 4 when Portugal gradually settles in an established democracy. The methodology produced important and interesting results over the comparable sub-corpora and proved very productive. We plan to apply it to other periods covered by the CRPC.