HAL
open science

Concordance Comparison as a Means of Assembling Local Grammars

Juliana P C Pirovani, Elias de Oliveira, Eric Laporte

To cite this version:

Juliana P C Pirovani, Elias de Oliveira, Eric Laporte. Concordance Comparison as a Means of Assembling Local Grammars. Aline Villavicencio; Moreira Viviane; Alberto Abad; Helena Caseli; Pablo Gamallo; Carlos Ramisch; Hugo Ricardo Gonçalo Oliveira; Gustavo Henrique Paetzold. Computational Processing of the Portuguese Language. 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings, 11122, Springer, pp.57-65, 2018, Lecture Notes in Artificial Intelligence, 978-3-319-99721-6. 10.1007/978-3-319-99722-3_6 . hal-01890393

HAL Id: hal-01890393

https://hal.science/hal-01890393v1

Submitted on 8 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Concordance Comparison as a Means of Assembling Local Grammars

Juliana P. C. Pirovani[1][0000−0002−4157−6503], Elias de Oliveira[1], and Eric Laporte[2][0000−0002−0984−0781]

[1] Universidade Federal do Espírito Santo - UFES
Av. Fernando Ferrari, 514, 29075-910 Vitória, ES, Brazil
jupcampos@gmail.com, elias@lcad.inf.ufes.br
[2] Université Paris-Est,
LIGM, UPEM/CNRS/ENPC/ESIEE, Champs-sur-Marne, 77420, France
eric.laporte@univ-paris-est.fr

Abstract. Named Entity Recognition for person names is an important but non-trivial task in information extraction. This article uses a tool that compares the concordances obtained from two local grammars (LG) and highlights the differences. We used the results as an aid to select the best of a set of LGs. By analyzing the comparisons, we observed relationships of inclusion, intersection and disjunction within each pair of LGs, which helped us to assemble those that yielded the best results. This approach was used in a case study on extraction of person names from texts written in Portuguese. We applied the enhanced grammar to the Gold Collection of the Second HAREM. The F-Measure obtained was 76.86, representing a gain of 6 points in relation to the state-of-the-art for Portuguese.

Keywords: Concordance · Local Grammar · Named Entity Recognition

1 Introduction

Named Entity Recognition (NER) involves automatically identifying names of entities such as persons, places and organizations. Person names are a fundamental source of information. Many applications seek information on individuals and their relationships, e.g. in the context of social networks. However, extracting this type of Named Entity (NE) is challenging: person names are an open word class, which includes many words and grows every day [8].

"*A good portion of NER research is devoted to the study of English, due to its significance as a dominant language that is used internationally*" [15, page 470]. An influential impetus to the development of systems for this purpose in Portuguese came with the HAREM [14, 9] events, a joint assessment of the area organized by Linguateca [7]. The annotated corpora used in the first and second HAREM, known as the Golden Collection (GC), serve as a reference for recent works on Portuguese NER.

The main approaches used to develop NER systems involve (i) machine learning, whereby systems learn to identify and classify NEs from a training corpus,

(ii) the linguistic approach, which involves manual description of rules in which NEs can appear, and (iii) a hybrid approach that combines both previous methods.

"*Local grammars (LG) are finite-state grammars or finite-state automata that represent sets of utterances of a natural language*" [6, page 1]. They were introduced by Maurice Gross [5] and serve as a way to group phrases with common characteristics (usually syntactic or semantic). Describing rules in the form of LGs for the construction of Information Extraction (IE) systems requires human expertise and training in linguistics; little computational aid for this task is available.

A method for constructing LGs around a keyword or semantic unit is presented by [6]. LGs for extracting person names from Portuguese texts were presented in [3] and [11]. In the Second HAREM [9], the Rembrandt system, which uses grammar rules and Wikipedia as sources of knowledge [4], ranked best for the 'person' category. A comparison between four tools to recognize NEs in Portuguese texts [2] suggested that the rule-based approach is the most effective for person names. Recently, LGs have been successfully integrated in a hybrid approach to Portuguese NER [12].

This paper describes how to use the Unitex concordance comparison tool [1] as an aid to constructing an LG. Our point of departure was a set of LGs to identify person names in Portuguese texts. By comparing concordances obtained from them, we found some relationships between them in terms of set theory. Taking into account these relationships, we picked the best LGs and combined them in order to achieve better performance.

This article is organized as follows. Section 2 presents the methodology used in this work. The results of the study are presented in Section 3, and Section 4 presents conclusions and avenues for future research.

2 The Methodology

The input to our experiment was a repository of small LGs to recognize person names. Some were obtained from the literature (e.g. those presented in [3]) and we created others.

All of these LGs were created and processed with Unitex [1], an open-source system initially developed at University of Paris-Est Marne-La-Valle in France. A local grammar is represented as a set of one or more graphs referred to as Local Grammar Graphs (LGG). Unitex allows for creating LGGs, preprocessing texts, applying dictionaries to texts, applying LGs to extract information, generating concordances and comparing concordances.

The LGG shown in Fig. 1 recognizes honorific titles such as *Sr.*, *Sra.* and *Dr.* ("Mr.", "Mrs.", "Dr.") followed by words with the first letter capitalized, as identified by the code `<PRE>` in Unitex dictionaries. The `<<..>>` after `<PRE>` denotes the application of a morphological filter to words with the first letter capitalized, indicating that they must include at least two characters. This prevents the recognition of definite articles at the beginning of sentences, for example.

Between the capitalized words, prepositions or abbreviations may occur and are recognized by two graphs, *Preposicao.grf* and *Abreviacoes.grf*, which have been created separately and are included as subgraphs. Examples of phrases recognized by the graph (occurrences) include *Sra. Joana da Silva* and *Dr. Antônio de Oliveira Salazar*. A list of occurrences accompanied with one line of context is referred to as a concordance.

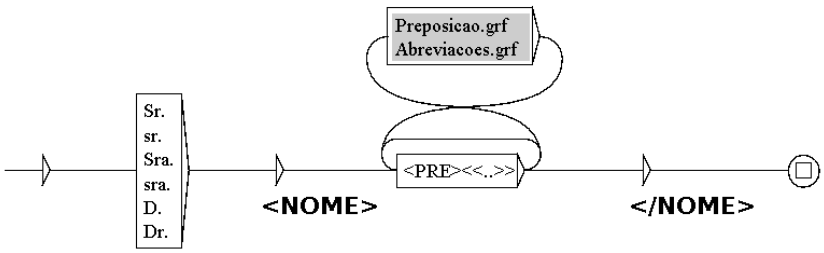

Fig. 1. LGG G_1 (ReconheceFormasDeTratamento.grf)

Unitex allows for attaching outputs to graph boxes. Outputs are displayed in bold under boxes. In Fig. 1, `<NOME>` ("name") and `</NOME>` shown under the arrows represent such outputs. Unitex inserts them into the concordance when a graph is applied in the "MERGE with input text" mode. Thus, the identified names appear enclosed in these XML tags in the concordance file.

The LGs of the repository are small but can be combined to compose a larger grammar to identify person names.

We applied the LGs of the repository to the Golden Collection (GC) of the Second HAREM, producing a concordance file for each LG. We used Portuguese and English dictionaries because several English names appear in GC texts.

The GC of the Second HAREM [9] is a subset of 129 annotated texts. These texts have different textual genres and are written in European or Brazilian Portuguese. The HAREM classifies ten categories of NEs: abstraction, event, thing, place, work, organization, person, time, value, and other. Person names, the focus of this work, are classified as a subtype within the 'person' category and are represented by the code PERSON (INDIVIDUAL). In the GC of the Second HAREM, 1,609 NEs are annotated with this code.

2.1 Concordance comparison

We compared all the concordances pairwise (every pair of files) using the ConcorDiff concordance comparison tool provided by Unitex. This tool can be applied to any pair of concordance files, provided they are in the Unitex format, which is publicly documented in the manual [10].

The Unitex ConcorDiff program compares two concordance files line by line and shows their differences. The result is an HTML page that presents alternate lines of both concordances and that leaves an empty line when an occurrence appears in only one of them. An example is presented in Fig. 2. The lines with a pink background shading (lines 1, 3, 5 and 7) are from the first concordance (the first parameter to ConcorDiff), and those with a green background shading (lines 2, 4 and 6) are from the other concordance (the second parameter to ConcorDiff).

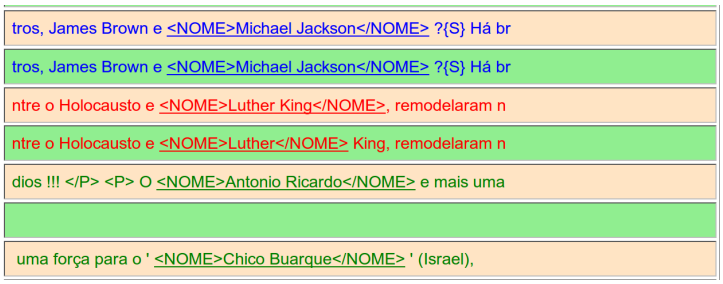

Fig. 2. Part of a concordance comparison file

Lines in blue characters (lines 1 and 2) are the occurrences common to the two concordances. In the example shown in Fig. 2, this means that both LGs recognized *Michael Jackson*. Lines in red characters (lines 3 and 4) correspond to occurrences that overlap only partially, which is the case, for instance, when an occurrence in a concordance is part of an occurrence in the other. In the example, an LG recognized *Luther King*, and the other recognized *Luther*. Lines in green characters (lines 5 and 7) are the occurrences that appear in only one of the two concordances. *Antonio Ricardo* and *Chico Buarque* were recognized only by the first LG. Lines in purple characters indicate identical occurrences with different outputs inserted, which does not happen in this example.

We then analyzed the files generated by ConcorDiff.

2.2 Composition of LG from concordance comparisons

Let G_X and G_Y be two LGs, and let C_X and C_Y the respective concordance files obtained by applying them to the same corpus. Thus, C_X is the set of occurrences identified by G_X, and C_Y is the set of occurrences identified by G_Y. Let $C_X \times C_Y$ be the file that shows the differences between concordances C_X and C_Y and is obtained through the ConcorDiff program of Unitex. In $C_X \times C_Y$, the elements x_1, x_2, ..., x_n of C_X are displayed on a pink background, while the elements y_1, y_2, ..., y_m of C_Y are displayed on a green background. It may exist between C_X and C_Y some relationships of the set theory, such as inclusion,

intersection or disjunction, and these relationships can be observed by analyzing $C_X \times C_Y$.

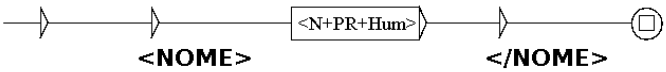

Fig. 3. LG G_2 (ReconheceNomesCompostos.grf)

Consider, for example, LGs G_1 (Fig. 1) and G_2 (Fig. 3). G_2 recognizes person names stored in dictionaries, through dictionary codes `N+PR` for proper names and `Hum` for nouns referring to human beings. Multiword person names such as *Marilyn Monroe, Cameron Diaz* and *Albert Einstein* are recognized by this LG after applying the English dictionary to the input text.

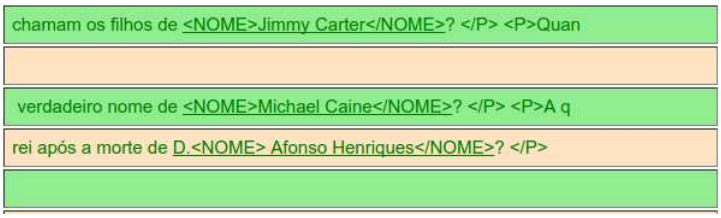

Fig. 4. Part of the concordance comparison $C_1 \times C_2$

Figure 4 shows part of the concordance comparison $C_1 \times C_2$. The first line, y_1, includes the name *Jimmy Carter* recognized by G_2. The first line displayed on a pink background, x_1, includes the name *Afonso Henriques* occurring after *D.* and recognized by G_1. Since lines in green characters are occurrences identified by only one of the two graphs, the first two occurrences were identified by G_2 only, and the last one by G_1 only. If all the lines of the comparison are in green characters and distributed between the two background colors, C_1 and C_2 are disjoint sets: thus, both LGs G_1 and G_2 are worth retaining as subgraphs of a grammar because they recognize different names.

Table 1 summarizes the main set-theoretic relationships identified. Each situation has a consequence in terms of priority between LGs, for example: G_X can be discarded if G_Y is retained. After analysing relationships between all pairs of LGs, we selected a subset of LGs and combined them into a larger LG (30 LGGs) by invoking them in a main graph.

Table 1. Main relationships observed through concordance comparison

Relation	Situation	Character color	Consequence
Inclusion	$C_X \subset C_Y$	Blue and green (on green background)	Keep G_Y
	$C_Y \subset C_X$	Blue and green (on pink background)	Keep G_X
Intersection	$C_X = C_Y$	Blue	Keep or G_X or G_Y
	$C_X = C_Y$ with different outputs	Violet	Analyze ambiguity
	$C_X \cap C_Y \neq \emptyset$	Blue and green (on different backgrounds)	Keep G_X and G_Y
Disjunction	$C_X \cap C_Y = \emptyset$, with $C_X = \emptyset$	Green (on green background)	Keep G_Y
	$C_X \cap C_Y = \emptyset$, with $C_Y = \emptyset$	Green (on pink background)	Keep G_X
	$C_X \cap C_Y = \emptyset$	Green (on different backgrounds)	Keep G_X and G_Y
Disjunction with partial overlapping of occurrences	$C_X \cap C_Y = \emptyset$, with $C_X \sim C_Y$[1]	Red	Keep G_X if $\forall i\ \lvert x_i \rvert > \lvert y_i \rvert$, keep G_Y if $\forall i\ \lvert x_i \rvert < \lvert y_i \rvert$
	$C_X \cap C_Y = \emptyset$, with $\exists i\ \exists j\ \ x_i$ overlaps y_j	Red and green (on identical background)	Keep G_X and G_Y if the occurrences in green characters are relevant. If not, keep only the LG that matches larger occurrences

[1] $C_X \sim C_Y \Leftrightarrow (n = m$ and $\forall i\ \ x_i$ overlaps $y_i)$.

3 Results and Discussion

We could not compare the performance of the obtained LG to the initial set of small LGs, since this set does not make up a single annotator together. Instead, we simply evaluated two annotators, one based on the obtained LG and another on an enhanced version of it, and we compared the results to those of Rembrandt, as a widely known reference.

We applied the obtained LG to the HAREM corpus and generated an XML file with the identified NEs, annotated according to directives of the Second HAREM. Parts of the person names identified by LG that appear isolated in the text are also annotated.

This file was submitted to SAHARA [13] for performance evaluation. SAHARA is an online system for automatic evaluation for HAREM, which computes the precision, recall and F-measure of an NER system after the user configures the evaluation and submits XML-annotated files.

The results obtained by applying the LG to the GC of the Second HAREM were 59.06% for precision, 55.22% for recall and 57.07 for F-measure.

Then, we employed manual strategies to improve the performance of the LG. In the Second HAREM, some words in lowercase letters should form part of NE[3]. For example, the honorific titles recognized by LGG in Fig. 1 and the person's social position that appears before the name. In an example provided by HAREM,[4] *A rainha Isabel II surpreendeu a Inglaterra* "Queen Elizabeth II surprised England", not only the name *Isabel*, but the whole phrase *rainha Isabel II* "Queen Elizabeth II" should be labeled as a person name.

We adapted the LGG ReconheceFormasDeTratamento.grf to address this issue by simply shifting the tag (`<NOME>`) before the honorific title in the graph, so that the title belongs to the tagged NE. Furthermore, we also used these words in lowercase letters to recognize the 'position' subcategory of the 'person' category, represented by PERSON(POSITION), and to recognize person names with a noun of social position in the left context.

The results obtained by the final LG are presented in Table 2. They were obtained with SAHARA by selecting the custom setting PERSON(INDIVIDUAL). This table also shows measures computed by SAHARA for Rembrandt, the system with the best performance for the 'person' category of the Second HAREM.

Table 2. Results considering PERSON(INDIVIDUAL): Rembrandt vs. final LG

System	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
Rembrandt	79	64.08	70.76
LG	79.75	74.18	76.86

The LG outperfoms Rembrandt. The recall of the LG is approximately 10 percentage points above that of Rembrandt.

Although our LG recognizes only the 'individual' and 'position' subtypes of the 'person' category, its evaluation was also carried out using SAHARA for all types of categories by selecting the PERSON(*) setting. A comparison of the obtained results with the results of the four tools presented in [2] for the 'person' category is shown in Table 3.

The LG has a better precision. However, as expected, it has a lower recall as it identifies fewer types of NEs: only two subtypes of the 'person' category ('individual' and 'position') are recognized, whereas the other systems recognize eight subtypes. We believe that with the addition of rules to the LG in order to recognize other subtypes of the 'person' category, the recall could be further increased, improving the LG approach even more as compared to other tools.

[3] http://www.linguateca.pt/aval_conjunta/HAREM/minusculas.html
[4] http://www.linguateca.pt/aval_conjunta/HAREM/ExemplarioSegundoHAREM.pdf

Table 3. Results considering PERSON(*): Systems in [2] vs. final LG

Systems	Precision (%)	Recall (%)	F-Measure (%)
NERP-CRF	57	51	54
Freeling	55	61	58
Language-Tasks	63	62	62
PALAVRAS	61	65	63
LG	81	60	69

4 Conclusions

This paper presented the use of the Unitex concordance comparison tool as a computational aid in manual composition of LGs. We used this tool for the composition of an LG to identify person names in texts written in Portuguese. The same methodology can be applied to the construction of LGs for other purposes.

Table 1 was created by listing the main set-theoretic relationships (inclusion, intersection and disjunction) that we could observe when analyzing concordance-comparison files generated by Unitex. Taking into account these relationships, we could produce a more compact and easily understandable grammar. We could also observe that a concordance offers an overview of what a LG recognizes in a specific corpus, allowing ambiguities and false positives to be identified.

The results of out final LG show its potential for NE extraction. It performed better (gain of 6 points) than Rembrandt, the system with the best performance for the 'person' category in the Second HAREM, when evaluating the 'person' category, 'individual' subtype, for which it was created.

As avenues for future work, we plan to apply the LG approach to other corpora of texts written in Portuguese, and to assess performance with a corpus not used in the construction of the LG. Moreover, we may add rules for recognizing other types of NEs. We also intend to study the feasibility of building elementary LGGs automatically or semi-automatically from examples, with the goal of minimizing human effort during construction. The concordance comparison tool presented in this article might facilitate the automation of decision-making for this purpose.

References

1. Unitex (2018), `http://unitexgramlab.org/`, acesso em: 02/03/2018
2. Amaral, D.O., Fonseca, E.B., Lopes, L., Vieira, R.: Comparative analysis of portuguese named entities recognition tools. In: Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 2554–2558. European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)
3. Baptista, J.: A local grammar of proper nouns. In: Seminários de Linguística. vol. 2, pp. 21–37. Faro: Universidade do Algarve (1998)

4. Cardoso, N.: Rembrandt-reconhecimento de entidades mencionadas baseado em relaações e análise detalhada do texto. In: In Cristina Mota and Diana Santos (eds.). Desafios na Avaliaação Conjunta do Reconhecimento de Entidades Mencionadas. vol. 1, pp. 195–211. Linguateca (2008)
5. Gross, M.: The construction of local grammars. In ROCHE, E.; SCHABÈS, Y. (eds.). Finite-state language processing, Language, Speech, and Communication, Cambridge, Mass. pp. 329–354 (1997)
6. Gross, M.: A Bootstrap Method for Constructing Local Grammars. In: Bokan, N. (ed.) Proceedings of the Symposium on Contemporary Mathematics, pp. 229–250. University of Belgrad (1999)
7. Linguateca: (2018), http://www.linguateca.pt, acesso em: 02/03/18
8. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT press (1999)
9. Mota, C., Santos, D.: Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM. Linguateca (2008), https://www.linguateca.pt/LivroSegundoHAREM/
10. Paumier, S.: Unitex 3.1 user manual (2016), http://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf
11. Pirovani, J.P.C., de Oliveira, E.: Extração de Nomes de Pessoas em Textos em Português: uma Abordagem Usando Gramáticas Locais. In: Computer on the Beach 2015. pp. 1–10. SBC, Florianópolis, SC (March 2015)
12. Pirovani, J.P.C., de Oliveira, E.: CRF+LG: A Hybrid Approach for the Portuguese Named Entity Recognition. In: International Conference on Intelligent Systems Design and Applications (ISDA 2017). Delhi, India (2017)
13. SAHARA: (2018), http://www.linguateca.pt/SAHARA/, acesso em: 02/03/2018
14. Santos, D., Cardoso, N.: Reconhecimento de Entidades Mencionadas em Português: Documentação e Actas do HAREM, a Primeira Avaliação Conjunta na Área. Linguateca (2007), http://www.linguateca.pt/aval_conjunta/LivroHAREM/Livro-SantosCardoso2007.pdf
15. Shaalan, K.: A survey of arabic named entity recognition and classification. Computational Linguistics **40**(2), 469–510 (2014), https://doi.org/10.1162/COLI_a_00178