Programa de Pós-Graduação em Informática
Universidade Federal do Espírito Santo (UFES)

Concordance Comparison as a Means of Assembling Local Grammars

Juliana P. C. Pirovani
Elias de Oliveira
Eric Laporte

INTRODUCTION

- **Local Grammars** (LGs) are one means of representing contextual rules.

- **Named Entity Recognition** (NER) is to automatically identify and classify named entities into predefined categories such as persons, organizations and places.
  - **LGs** have been successfully integrated in a hybrid approach to Portuguese NER [1].

- Describing rules and constructing LGs requires human expertise.

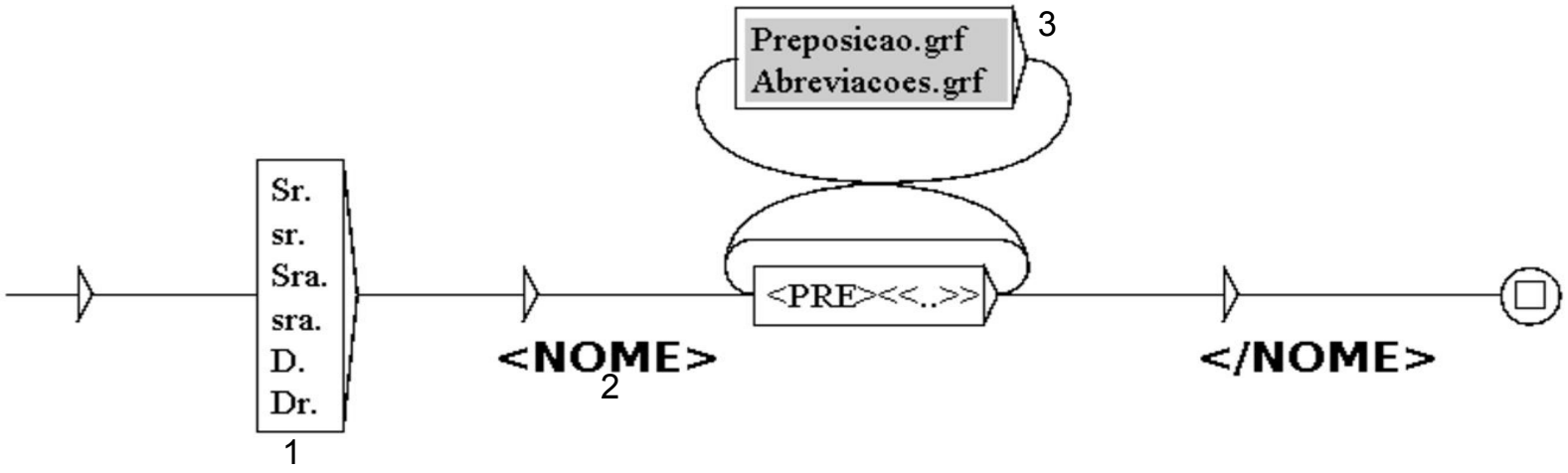
OBJECTIVES

- Describe how to use the Unitex [2] concordance comparison tool as an aid to construct an LG.

- Present a case study on extraction of person names from texts written in Portuguese.

METHODOLOGY

- LGs were created and processed with Unitex
  - A local grammar is represented as a set of one or more graphs (Local Grammar Graphs – LGG).
  - A list of occurrences is referred to as a concordance.



1. Mr./Mrs./... 2. Name 3. Preposition/abbreviation

METHODOLOGY

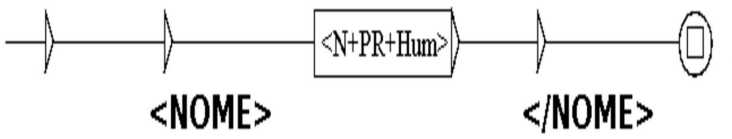
- Comparison of concordances (Concordiff) in Unitex
  - ConcorDiff program compares two concordance files line by line and shows their differences.

tros, James Brown e <NOME>Michael Jackson</NOME> ?{S} Há br
tros, James Brown e <NOME>Michael Jackson</NOME> ?{S} Há br
ntre o Holocausto e <NOME>Luther King</NOME>, remodelaram n
ntre o Holocausto e <NOME>Luther</NOME> King, remodelaram n
dios !!! </P> <P> O <NOME>Antonio Ricardo</NOME> e mais uma
uma força para o ' <NOME>Chico Buarque</NOME> ' (Israel),

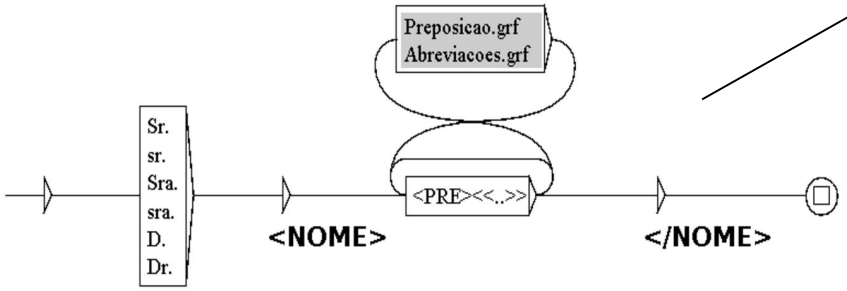
Blue: Occurrences common to the two concordances
Red: Occurrences that overlap only partially
Green: Occurrences that appear in only one of the two concordances
Purple: Identical occurrences with different outputs inserted

METHODOLOGY

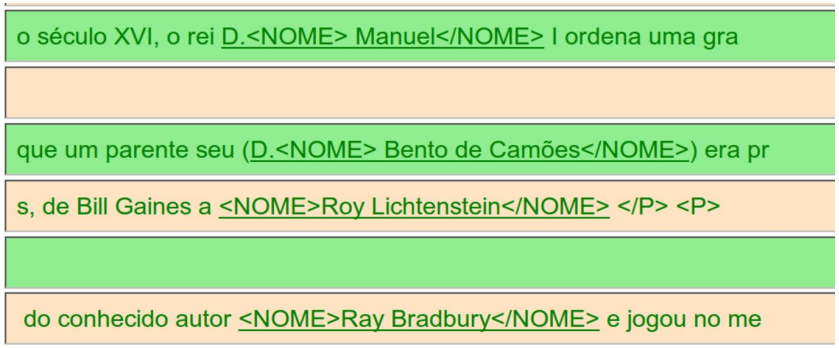
- Composition of LG from concordance comparisons
  - LGGs: G1 and G2
  - Respective concordance files: C1 and C2
  - C1 x C2: file that shows the differences between C1 and C2



LGG G1 – ReconheceNomesCompostos.grf



LGG G2 – ReconhecePronomesDeTratamento.grf

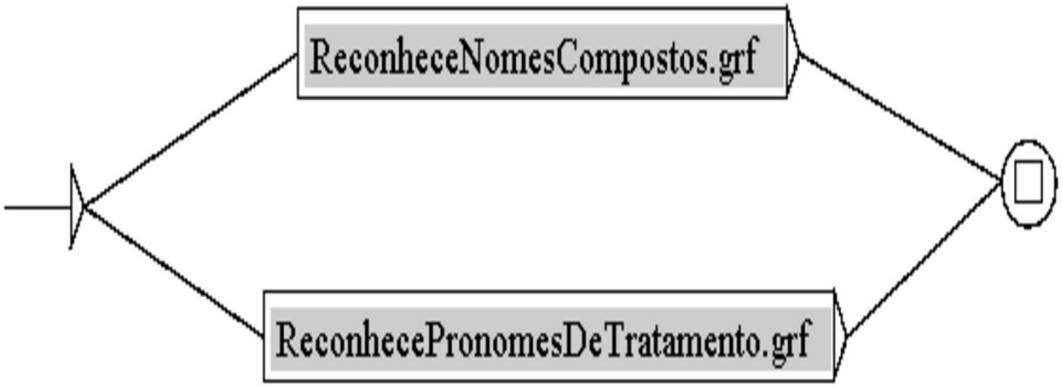


Part of the concordance C1 x C2

It represents the disjunction of the two concordances.

METHODOLOGY

- Composition of LG from concordance comparisons
  - If all the lines of the comparison are in green and distributed between the two background colors, C1 and C2 are disjoint sets: G1 and G2 are worth retaining as subgraphs of a grammar.



METHODOLOGY

- Experiment

1) Construction of LGGs repository in Unitex.

2) Application of the LGGs to the GC of the Second HAREM.

3) Comparison of the concordances obtained by the Concordiff program of Unitex.

4) Analysis of the files generated by ConcorDiff.

5) Composition of the LG to recognize person names.

RESULTS

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

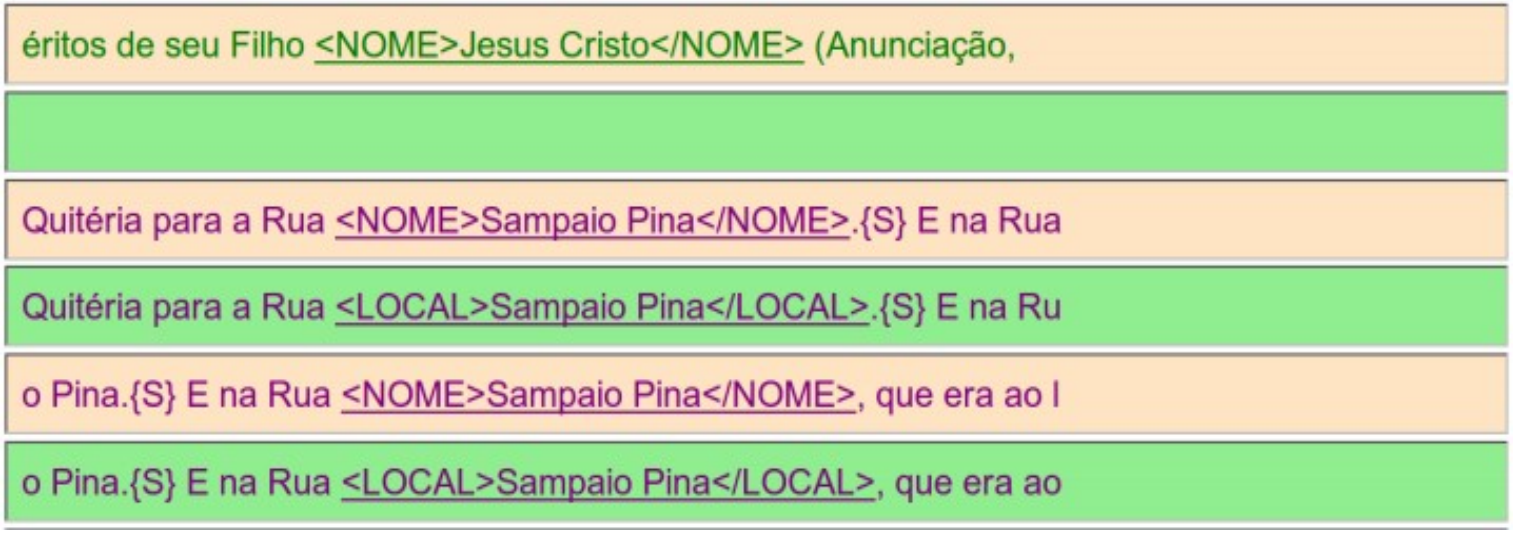
[1] $C_X \sim C_Y \Leftrightarrow (n = m$ and $\forall i \ \ x_i$ overlaps $y_i)$.

RESULTS

- Relation: Disjunction with partial overlapping of occurrences
  - Consequence: LG must be used to extract long occurrences as this results in a higher probability that the occurrences are really a name.

RESULTS

- Relation: Intersection
  - Consequence: Both LGGs should be retained and the ambiguity should be treated afterward Unitex.



RESULTS

- Initial results
  - Precision: 59.06%
  - Recall: 55.22%
  - F-Measure: 57.07%
- Some improvements have been made at LG
  - Recognition of the honorific titles as part of the name
  - Recognition of the ‘position’ subcategory of the ‘person’ category

RESULTS

- Rembrandt vs. final LG: PERSON(INDIVIDUAL)

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

- Systems in AMARAL (2014) vs. final LG: PERSON(*)

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

CONCLUSIONS

- We presented the use of the Concordiff program of the Unitex as a computational aid in manual composition of LGs.
- We presented the description of the main set-theoretic relationships that we could use to assembling LGs.
- Future Works
  - Addition of rules for recognizing other types of NEs
  - Study of the feasibility of building elementary LGGs automatically or semi-automatically from examples

REFERENCES

- [1] 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).
- [2] Unitex (2018), http://unitexgramlab.org/, acesso em: 02/09/2018
- [3] 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).
- [4] Paumier, S.: Unitex 3.1 user manual (2016), http://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf
- [5] Cardoso, N.: Rembrandt-reconhecimento de entidades mencionadas baseado em relações 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).
- [6] Amaral, D.O., Fonseca, E.B., Lopes, L., Vieira, R.: Comparative analysis of portuguese named entities recognition tools. In: 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).

THANK YOU!

jupcampos@gmail.com