AUTOMATIC EXTRACTION OF SUBCATEGORIZATION FRAMES FROM CORPORA IN PORTUGUESE

Leonardo Zilio (PPG-Letras/UFRGS)

Adriano Zanette (PPG-Computação/UFRGS)

Carolina Scarton (CCMC/USP)

OUR GOALS

- To present a tool for automatic extraction of subcategorization frames (SCFs) designed specially for Portuguese

- To show some early results of two different studies which use this tool

WHAT IS AN SCF?

- Syntactic representation of a clause or phrase.

CLAUSE REPRESENTATION

- Clause
  - Marcou o gol que deu sobrevida a o time , deu carrinhos e conduziu a equipe com uma qualidade que nenhum outro jogador apresentou – nem=de=longe .

NP_NP

PHRASE REPRESENTATION

- Phrase
  - Privação de liberdade

SN_SP[de]

OUR EXTRACTOR

- Initially developed by Zanette (2010)
- Improved in 2011-2012 (Zanette et al. 2012)
- Extracts SCFs of clauses
- http://143.107.232.109/scf_port/index.html

HOW IT WORKS

1 - Input: corpora annotated with the parser PALAVRAS (Bick, 2000)

- Dependency trees

2 - Processing of all sentences in the corpora

3 - Extraction of all dependencies of main verbs verbs

4 - Analysis of the relevant dependencies (exclusion of adverbs)

5 – Output: Database of subcategorization frames

WORKS IN PROGRESS

Verb Lexicon

VERB LEXICON

- Building of VerbNet.br (Scarton, 2011)

- Grouping verbs according to syntactic patterns - according to Levin (1993)

- Changed the original frame:
  - O homem quebrou a janela com um martelo

    (The man broke the window with a hammer)
  - **SUBJ[NP] V NP PP[com]**

VERB LEXICON

- Corpora:
  - Lácio-Ref (~9 million words) – (Aluisio 2004)
  - PLN-BR (~26 million words) – (Bruckschen 2008)
  - Revista FAPESP (~6 million words) – (Aziz and Specia 2011)

VERB LEXICON

- Two approaches:
  - Validating a semiautomatic method used to build VerbNet.Br (by using others Computational Lexical Resources)
  - Verb clustering (complete automatic method based on Machine Learning)

VERB LEXICON

- VerbNet.Br:
  - Based on VerbNet (Kipper, 2005)
  - Being built through the alignments among VerbNet, WordNet and WordNet.Br
  - The SCFs are used to validate the candidate members identified by the others resources

VERB LEXICON

- Method for building VerbNet.Br:
  - Identify candidate members to VerbNet classes through use of alignments among VerbNet, WordNet and WordNet.Br
  - For each candidate member, identify the SCFs
  - Compare with the SCFs defined manually for each class

VERB LEXICON

- Verb Clustering (Sun et al., 2010):
  - Use of the syntactic patterns to group verbs together → **MACHINE LEARNING METHODS**
  - Trying to validate Levin's hypothesis → "Verbs that fall into classes according to shared (syntactic) behavior would be expected to show shared meaning components"

VERB LEXICON

- Results
  - Identified:
    - 7.252 verbs
    - 17.448 frames (parameterized by prepositions – frequency higher than 1)
- Verb Clustering:
  - The best result: 42.6% of F-measure (using Spectral Cluster algorithm) for the task in a gold standard with 12 classes of VerbNet (translated from English)
    - The best result for English: 63.3% of F-measure (using Spectral Cluster algorithm)

WORKS IN PROGRESS

Semantic Role Labeling

SEMANTIC ROLES

- The butcher cuts the meat.
  - The butcher = **agent**
  - The meat = **patient/theme**

- I opened the door with a key.
  - I = **agent**
  - The door = **patient/theme**
  - With a key = **instrument**

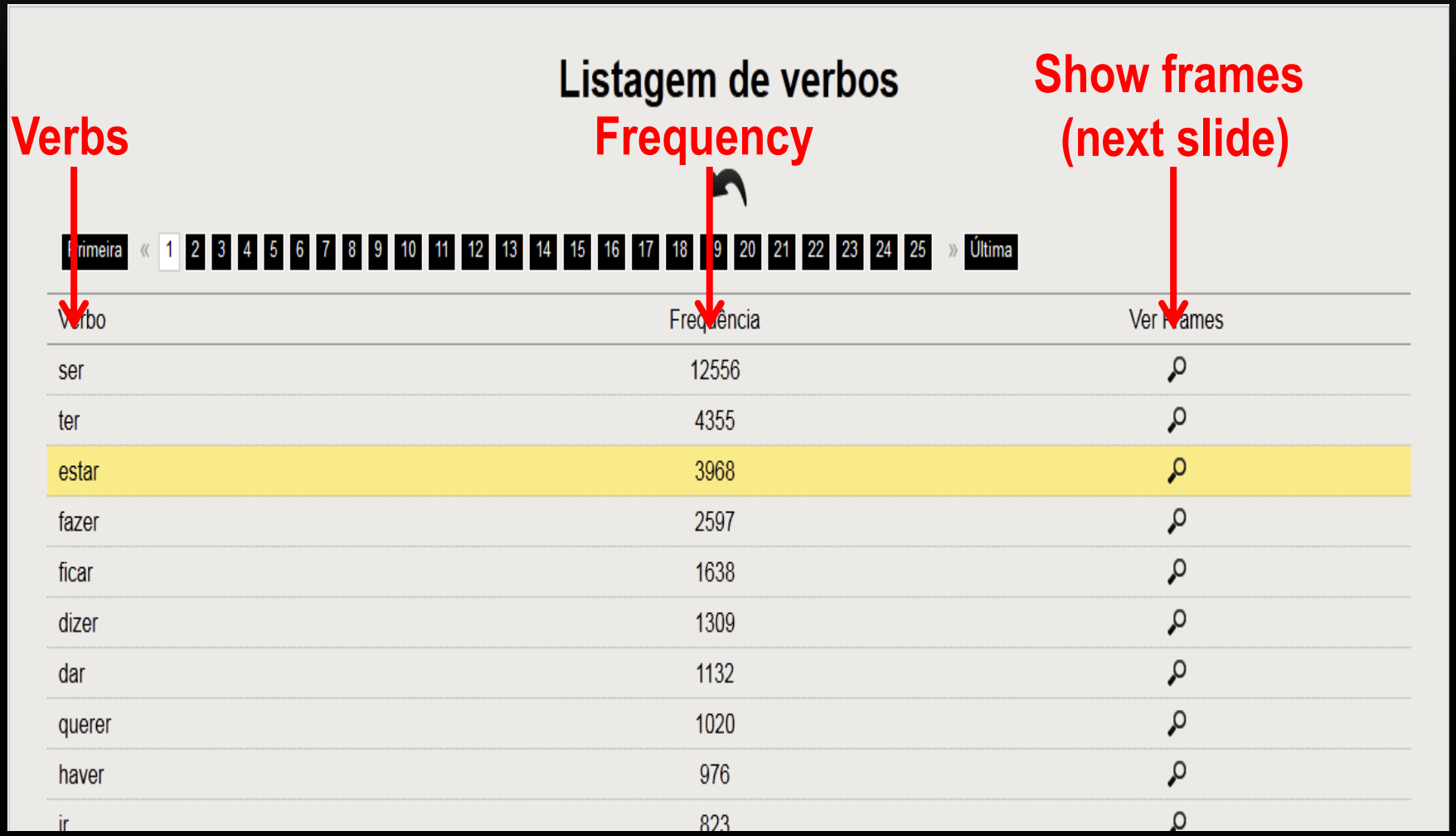
SEMANTIC ROLE LABELING

- Two corpora:
  - Cardiology = 1.5+ million words
  - Newspaper = 1+ million words

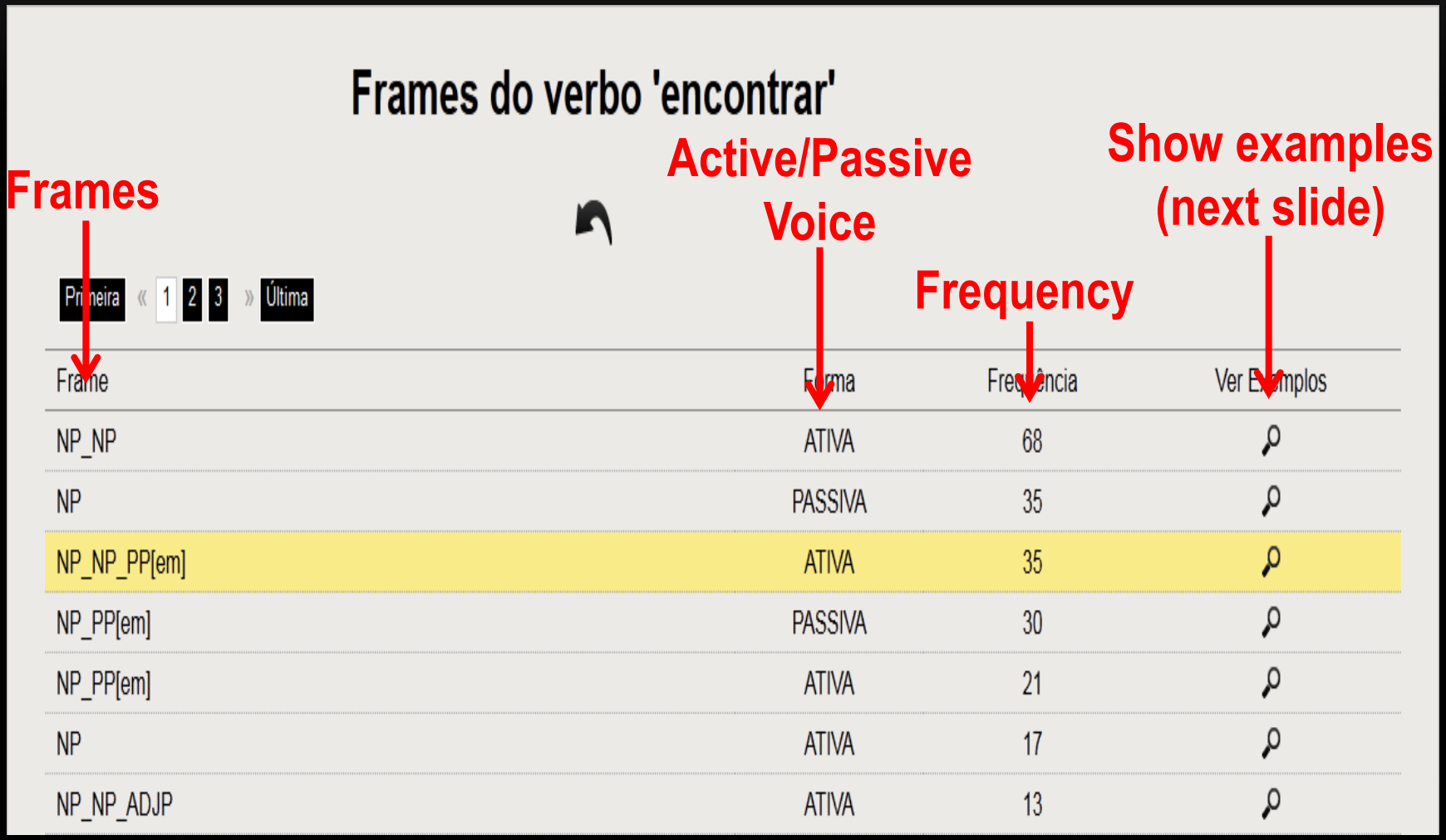
- Semantic roles from the works of Brumm (2008) and Gelhausen (2010)

INTERFACE

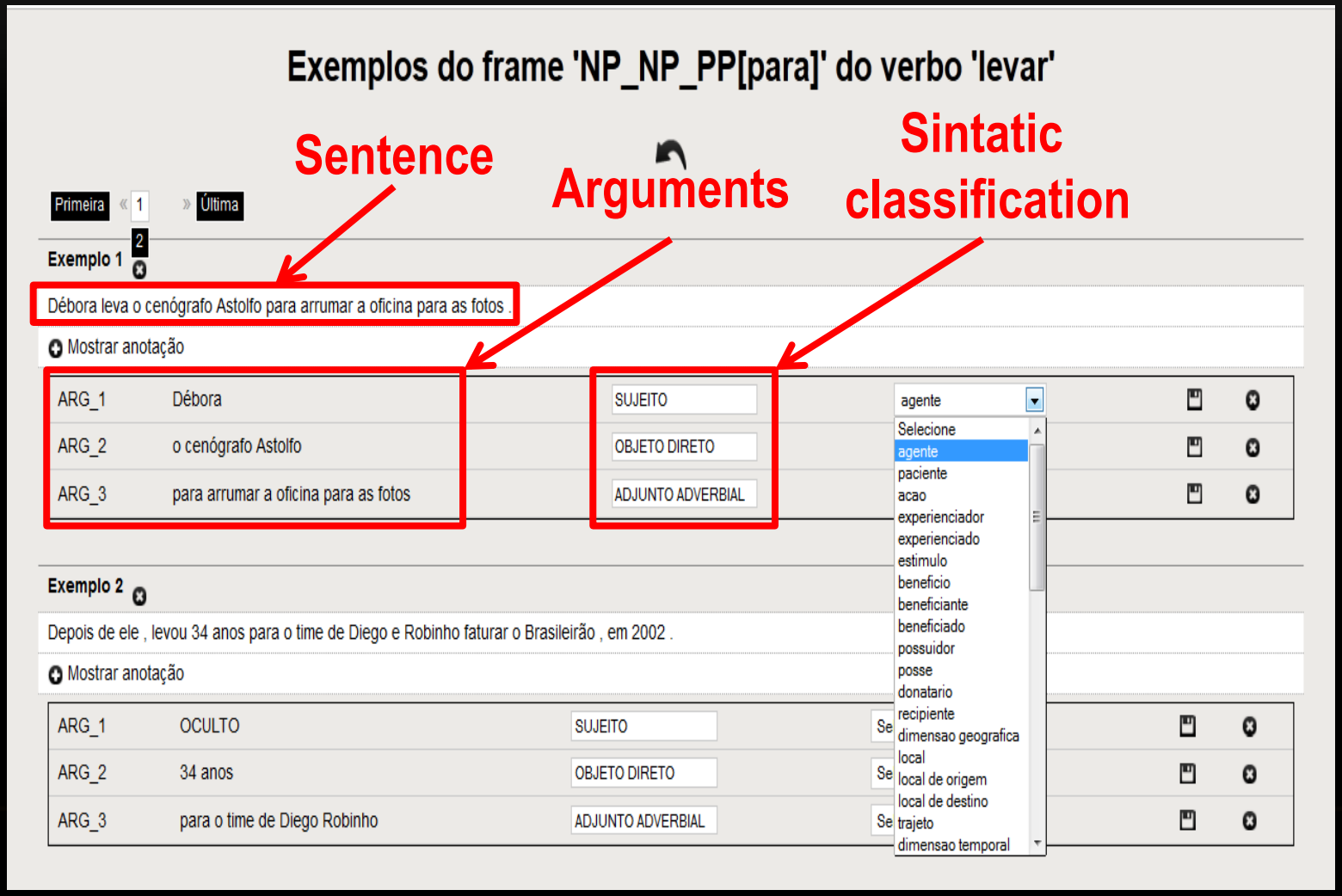
PHP-INTERFACE 1 – LIST OF VERBS



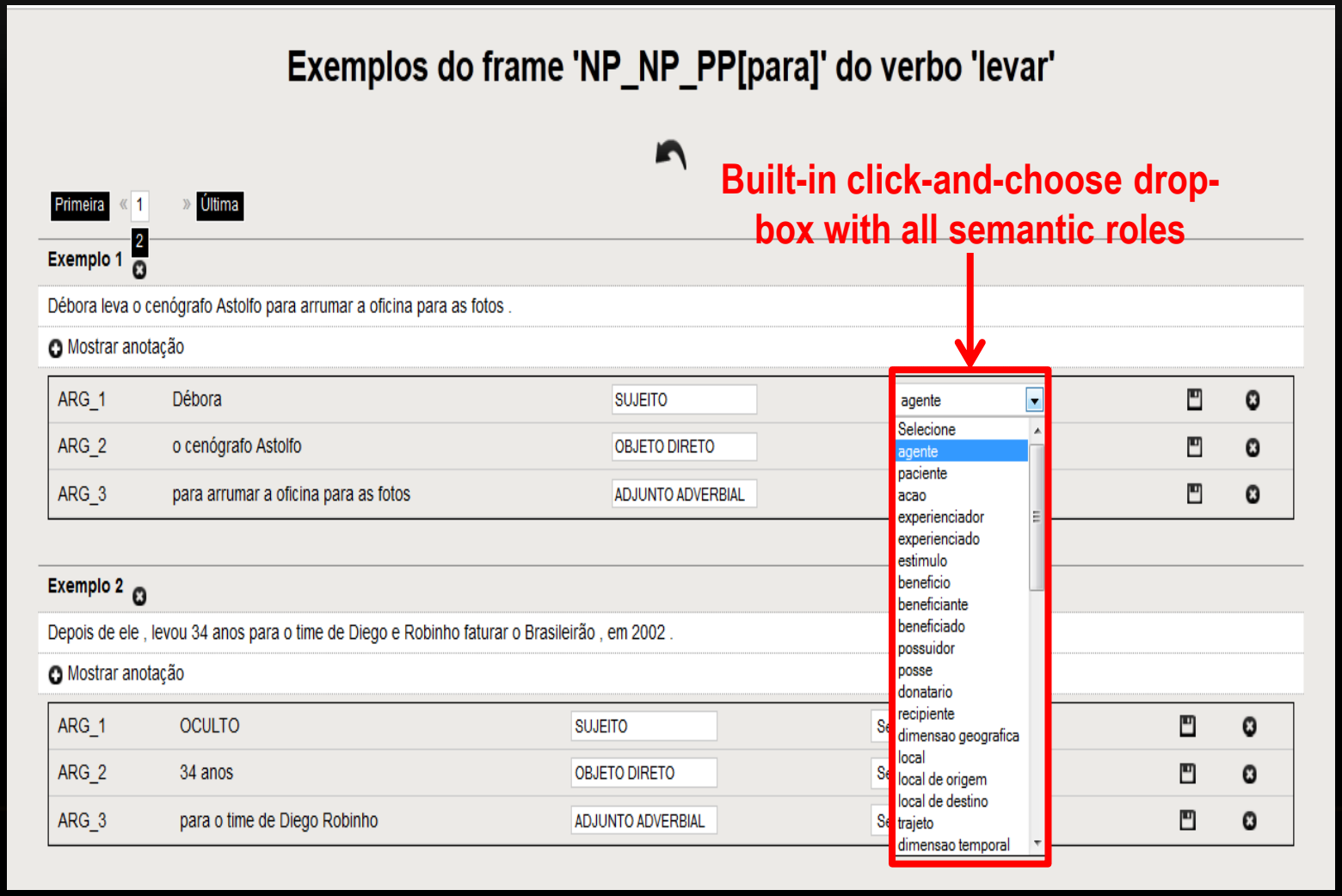
PHP-INTERFACE 2 – LIST OF FRAMES



PHP-INTERFACE 3 – LIST OF EXAMPLES



SEMANTIC ROLE LABELING



CURRENT SEMANTIC ROLE LABELING

- 46 semantic roles (Brumm 2008; Gelhausen 2010)

- Annotation of 4 verbs in both corpora:
  - encontrar [to find]
  - levar [to take/carry]
  - receber [to receive]
  - usar [to use])

- Test in a small set of verbs

RESULTS

- Annotation of 482 frames
- 138 diferent semantic roles configurations

CURRENT DEVELOPMENTS

- Too many roles, some are not used or are too specific
- Change of the semantic roles set
- Testing of the set applied at the VerbNet (Kipper 2005)

ACKNOWLEDGEMENTS

- This research was partly funded by the following agencies:
  - CNPq
  - FAPESP
- Institutes:
  - NILC – ICMC-USP
  - IL-UFRGS
  - Inf-UFRGS

REFERENCES

- ALUISIO, S.; PINHERO, G. M.; MANFRIM, A. M. P.; OLIVEIRA, L. H. M. de; GENOVES JR., L. C.; TAGNIN, S. E. O.: The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. In The Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Lisboa, Portugal, 1779-1782.
- AZIZ, W. and SPECIA, L.: Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation, 2011. In Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, Cuiaba, Brasil.
- BICK, Eckhardt. (2000) *The Parsing System* PALAVRAS*: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press. Disponível em: <http://beta.visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf>
- BRUMM, Torben. (2008) *Erstellung eines Systems thematischer Rollen mit Hilfe einer multiplen Fallstudie*. Studienarbeit, 103p. Betreuer: Tom Gelhausen. Diponível em: http://www.ipd.uka.de/Tichy/theses.php?id=135
- BRUCKSCHEN, M., MUNIZ, F., SOUZA, J. G. C., FUCHS, J. T., INFANTE, K., MUNIZ, M., GONÇALVES, P. N., VIEIRA, R. e ALUISIO, S. M. Anotaçãoo Linguística em XML do Corpus PLN-BR, 2008. Série de Relatorios do NILC. NILC-TR-09-08, 39 p.

REFERENCES

- GELHAUSEN, Tom. (2010) *Modellextraktion aus natürlichen Sprachen*: Eine Methode zur systematischen Erstellung von Domänenmodellen. Karlsruhe: KIT Scientific Publishing. Dissertation, Karlsruher Institut für Technologie. Disponível em: <http://digbib.ubka.uni-karlsruhe.de/volltexte/documents/1437903>
- KIPPER, K. (2005) VerbNet: a broad-coverage, comprehensive verb lexicon. University of Pennsylvania. Tese de doutorado orientada por Martha S. Palmer.
- Scarton, C.: VerbNet.Br: construção semiautomática de um léxico computacional de verbos para o português do Brasil, 2011. In Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, Cuiaba, Brasil.
- SUN, L.; KORHONEN, A.; POIBEAU, T.; MESSIANT, C.: Investigating the cross-linguistic potential of VerbNet: style classification, 2010. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, 1056-1064
- ZANETTE, Adriano. (2010) *Aquisição de* Subcategorization Frames *para Verbos da Língua Portuguesa*. Projeto de Diplomação. UFRGS. Orientadora: Aline Villavicencio.
- ZANETTE, Adriano; SCARTON, Carolina; ZILIO, Leonardo (2012) Automatic extraction of subcategorization frames from corpora: an approach to Portuguese. In: *Proceedings of PROPOR 2012* - Demonstration Session. Coimbra, Portugal.

MUITO OBRIGADO!

Leonardo Zilio

lzilio@ig.com.br