Macunaíma as a data source for the distribution and interpretation of bare NPs in Brazilian Portuguese: Corpus preparation and first annotation agreement results

Albert Wall (University of Tübingen, University of Zürich), Niels Ott, Ramon Ziai, Björn Rudzewitz (University of Tübingen)

Why *Macunaíma*?

Macunaíma as a linguistic project

- Mário de Andrade (author) is one of the leading figures in the Brazilian modernist movement and, besides his many interests, has dedicatedly worked on the consolidadtion of a "Brazilian language", closer to everyday usage in Brazil than to the normative grammars from Portugal (cf. e.g. Rodrigues 2013).
- Final classification of the genre of the text by the author in the second edition: **rhapsody**

➔ Narration made for oral presentation, popular/folkloristic themes, "fictitious orality"

Important editorial details

- First version written in December 1926, after a longer period of investigation, first publication in 1928.

- Text refinement and fixation very well-documented: eight editions from the author's life time, two manuscripts, author's comments in letters, newspaper articles, interviews, notes; critical edition (Ancona Lopes 1978).
- "Article omission" in the contexts of interest for this study are never commented or corrected.

➔ Arguably, the article use in Macunaíma is not some editorial problem, but rather intended by the author and represents the contemporary spoken usage as perceived by him (cf. Wall 2013a).

Interesting examples of "bare" NPs

Bare singulars in episodic predicates:

- Abra a porta pra mim entrar!
- Porém **jacaré abriu**? Nem eles! e a cabeça não pôde entrar. (M: 32)

Definite / specific bare singulars:

Macunaíma atirou **a cabeça** por aí, na pressa de matar todos os peixes, **cabeça** caiu numa lapa e juque! mergulhou no rio. (M: 131)

➔ The existence, grammaticality and analysis of such bare singulars (BSs) in BrP is hotly debated in the literature (cf. Wall 2013b, Wall in press).

The annotation: work in progress

Previous example extraction (Wall 2013a)

- **7 occurrences of "jacaré V_{perf}"**: … abriu? / ... acreditou? / ... saiu? (2) / ... achou? (2) / ... (a)fastou?
- "A popular comical expression indicating impossibility" (M: 313) ➔idiomatic, but still flexible and productive!
- **37 occurrendes of "definite / specific" BSs** (26 definite subjects, 20 of them count nouns, 12 animated, all of them anaphoric / already mentioned in the context).
- Claims in the literature about such sentences range from "ungrammatical" (Müller 2002, among others) to "quite frequent" (Barme 2011) ➔ no empirical evidence.

Annotating *Macunaíma (critical edition)*

Automatic pre-processing:

- Automatic pre-annotation of tokens (based on Ziai, 2009), sentences (using OpenNLP), part-of-speech (using OpenNLP+Aelius model), and noun phrase chunks (our own implementation inspired by *Chunking.py* of Aelius)
- Developed a heuristic detector for gender and number of nouns and "bare" attribute of NPs (no article before first noun in chunk)
- POS tag set: modified version of MacMorpho scheme with combined tags for contractions, e.g., PRE_and_ART, definite vs indefinite articles, punctuation.

➔ Markables are nouns (N) and noun phrases (NPs) that must be post-corrected.

Annotation:

- data loaded into Brat Rapid Annotation Tool (Brat) and annotated by two independent human annotators.
- Task: post-correcting Ns and NPs and then selecting properties of Ns and NPs according to a pre-defined scheme.

Annotation scheme for use in *Brat*:

- 2 steps for Ns:
  - number morphology (sg / pl)
  - "denotation" (concr. / abstr.; mass / count)
- 6 steps for NPs:
  - bare (yes / no)
  - presence of modifiers (yes / no)
  - subject / direct object / other syntactic function
  - linear order: before Verb / after Verb
  - "interpretation": refers to object / class / does not refer
  - 6 additional exclusive features regarding NP internal structure and role in discourse:
    - "full DP" (def. / indef. / other determiner)
    - "bare NP" (anaphor / assoc. anaphor / disc. new)

Results: Annotation agreement

Data from a pilot study

- For a pilot study, two human annotators worked on 2307 tokens in 176 sentences of the corpus.
- Annotator A post-corrected/defined 515 NPs, annotator B resulted in 519 NPs.
- Annotator A post-corrected/defined 624 Ns, annotator B resulted in 594 Ns.

Quality of automatic pre-processing:

- On average, the two human annotators agreed with the pre-processing on 60.0% of all NP markables.

➔ The corpus text with its unconventional syntax and high frequency of rare nouns and words in general of course is a challenge for heuristic tagging and chunking

➔ Although it would be desirable to improve the performance, this pre-processing is already quite useful since it saves considerable time for human annotators.

- For all N markables (including proper nouns), the average agreement between the pre-processing and the human annotators was 97.0%.

➔ This pre-processing performance was to be expected, since POS tagging works in that ballbark of accuracy.

Agreement on post-corrections:

- The two annotators identically post-corrected/defined 419 NPs (81.0% on average) and 587 Ns (96.4% avg.).

- We present detailed agreement figures on the NPs and Ns common for both annotations:

Inter-annotater Agreement on NP and N attributes:

NP	% of agreement	Cohen's kappa
Bare	95.0	0.88
Linear order	91.4	0.81
Syntactic function	77.3	0.67
Presence of modifiers	**90.9**	0.65
Int. struct. & disc. role	73.5	0.58
Interpretation	73.5	0.53
N		
Number morphology	96.8	0.87
Denotation	**63.4**	0.40

Conclusions and next steps

Interpretation of agreement results

- "Bare", "linear order" and "plural morphology" show a very high annotation reliability of $\kappa > 0.8$.
- "Syntactic function" and "presence of modifiers" have a κ-value slightly below 0.7, a (possible) threshold value for reliability and usefulness in computational studies on discourse (Artstein & Poesio 2008).

➔ Lower performance in the case of "syntactic function" might be due to the sometimes unconventional syntax and style of the corpus text.

➔ Cohen's kappa for "Presence of modifiers" presumably is low due to a highly disproportional occurrence of the two categories, cf. the rather high percentage of agreement.

➔ The same holds for "internal structure & discourse role" and "interpretation".

- "Denotation", although also featuring a pronounced disproportional occurrence between the four categories, obviously contains further problems (cf. rather low percentage of agreement).

➔ the four "canonical" categories are either too coarse-grained for corpus annotation or the annotators' instructions were insufficient.

Conclusions

- Promising results for a pilot annotation study
- Problematic cases are detected and possible sources of the problems are identified.

➔ Further improvement of agreement expected

➔ The annotated corpus can serve as a basis for deeper annotation (modified NPs, noun types (Löbner 2011), denotation classes (Rijkhoff 2002) …).

➔ The corpus allows for automatic search and statistical analysis of a combination of syntactic and semantic features.

➔ The corpus contains interesting information about spoken BrP syntax from the first half of the last century.

Next steps

- Further in-depth analysis of major agreement mismatches
- Development of a more explicit decision guide for problematic cases of semantic and discourse features
- Reconsider "denotation" categories for Ns
- Second (final?) round of annotation
- Try to develop an automatic annotation for "linear order" (and maybe "presence of modifiers")

Acknowledgements: The authors would like to thank **Angelika Gärtner** and **Cristina Lopes-Videira** (University of Tübingen), who performed the independet annotation of the corpus.

References

• Ancona Lopez, T. (ed). 1978. *Mário de Andrade: Macunaíma. O herói sem nenhum caráter*. Edição crítica, LTC Editora: Rio de Janeiro.
• Artstein, R. & Poesio, M. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4): 555-596.
• Barme, S. 2011. Sertanejo não sabe chorar: zum Nullartikel bei Nominalphrasen mit Subjektfunktion im Brasilianischen. *Zeitschrift für romanische Philologie 127*(1), 162-171.
• Löbner, S. 2011. Concept Types and Determination. *Journal of Semantics* 28(3): 279-333.
• Müller, A. L. 2002. Genericity and the denotation of common nouns in Brazilian Portuguese. *D.E.L.T.A.* 18(2): 287-308.
• Rijkhoff, J. 2002. The Noun Phrase. Oxford: OUP.
• Rodrigues, L.G. 2013. A língua brasileira de Mário de Andrade: nacionalismo, literatura e epistolografia. *Todas as Musas* 4(2): 100-116.
• Wall, A. 2013a. „Porém jacaré acreditou?“ Eine kritische Macunaíma-Edition als Glücksfall für die Beschreibung der brasilianischen Nominalphrase. Talk given at the 33. Romanistentag, Würzburg.
• Wall, A. 2013b. The distribution of definite and specific bare nominals in Brazilian Portuguese. In J. Kabatek & A. Wall (eds), *New Perspectives on Bare Noun Phrases in Romance*. Studies in Language Companion Series 141, 223-254. Amsterdam: John Benjamins.
• Wall, A. in press. The role of grammaticality judgments within an integral approach to Brazilian Portuguese bare nominals. In B. Hemforth, B. Schmiedtová & C. Fabricius-Hansen (eds), *Psycholinguistic Approaches to Meaning and Understanding across Language*. Studies in Theoretical Psycholinguistics Series 44. Springer.
• Ziai, R. 2009. A Flexible Annotation-Based Architecture for Intelligent Language Tutoring Systems. Master's thesis, Universität Tübingen, Seminar für Sprachwissenschaft.

• Aelius: http://aelius.sourceforge.net
• Brat: http://brat.nlplab.org/
• OpenNLP: http://opennlp.apache.org