**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 consolidation 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/tolerationistic themes, “futilous earthly”

**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 project, 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 para mim entrar! - Abra a porta para eu entrar! (M: 32)
- Porém *jácaré abriu*? Nem eles! e a cabeça não pode entrar. (M: 131)

**Definite / specific bare singulars:**
- Macunaíma atira a cabeça por aí, na pressa de matar todos os peixes, *caica* caiu numa lapa e *jugue* 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).

**Previous example extraction (Wall 2013a)**
- 7 occurrences of “jacaré *Vou*”: ... abriu? / ... acredito? / ... saiu? (2) / ... achou? (2) / ... (a)estou? ...
- “A popular comic expression indicating impossibility” (M: 313) - idiomatical, but still flexible and productive!
- 37 occurrences 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 “definite” 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 ballpark of reliability.

**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-annotator Agreement on NP and N attributes:**

<table>
<thead>
<tr>
<th>NP</th>
<th>% of agreement</th>
<th>Cohen’s kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bare</td>
<td>95.0</td>
<td>0.88</td>
</tr>
<tr>
<td>Linear order</td>
<td>91.4</td>
<td>0.81</td>
</tr>
<tr>
<td>Syntactic function</td>
<td>77.3</td>
<td>0.67</td>
</tr>
<tr>
<td>Presence of modifiers</td>
<td>90.9</td>
<td>0.65</td>
</tr>
<tr>
<td>Int. struct. &amp; disc. role</td>
<td>73.5</td>
<td>0.58</td>
</tr>
<tr>
<td>Interpretation</td>
<td>73.5</td>
<td>0.53</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number morphology</td>
<td>96.8</td>
<td>0.87</td>
</tr>
<tr>
<td>Denotation</td>
<td>63.4</td>
<td>0.40</td>
</tr>
</tbody>
</table>

**Conclusions and next steps**

**Interpretation of agreement results**
- “Bare”, “linear order” and “plural morphology” show a very high annotation reliability of κ > 0.8.
- “Syntactic function” and “presence of modifiers” have a κ-value slightly below 0.7, a (possible) threshold valid for reliability and usefulness in computational studies on discourse (Arstein & 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), discourse 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 (first round of annotation)
- Try to develop an automatic annotation for “linear order” (and maybe “presence of modifiers”)
References


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