Annotating highly inflected languages from Brazil: From Portuguese to Kadiwéu
Charlotte Galves (Unicamp), Filomena Sandalo (Unicamp), Luiz Veronesi (Unicamp/USP), Glauber Romling da Silva (Unicamp)

We propose for Kadiwéu, a polysynthetic language of Brazil, an extension of the POS annotation of the Tycho Brahe Annotated Corpus of Historical Portuguese (www.tycho.iel.unicamp.br/~tycho/corpus) – henceforth TBC, which consists in tagging both words and morphemes, yielding a two-level annotation. The tagging of words is necessary to generate the syntactic parsing that is missing from the current corpuses of Brazilian native languages. The morphological tagging is also crucial for polysynthetic languages since it allows searching for grammatical properties encoded by the morphemes. This is a pioneer proposal since it is the first time an American Indian language will be part of a Corpus allowing grammatical searches that include morphological and syntactic information.

For the TBC, in order to treat the rich inflectional morphology of Portuguese, the POS tagging system of the Penn Parsed Corpora of Historical English (Kroch and collaborators 2000, 2004, 2010) was adapted in such a way that tags can be articulated, with a basic one, corresponding to the category of the word (VB, D, N, NPR, ADJ), and one or more secondary tags which encode morphological properties (-D, -UM-F, -F, -G) (cf. Britto, Finger, Galves 2002)

Pelejou/VB-D a/D-F armada/N de/P Holanda/NPR com/P uma/D-UM-F fight-PST the army of Holland with a esquadra/N d@/P @a/D-F armada/N Real/ADJ-G de/P Castela/NPR squadron of the army royal of Castela

In a language like Kadiwéu, the information conveyed by morphology is too rich to be treated this way. In these languages, except for some rare cases of porte-manteau morphemes and suppletion, the correspondence between form and features is one-to-one, and can be encoded in a single tag. In a two-level tagging system, this can be represented as in (2), generated by the structure in (3):

(1) ‘Natigide jaaGa jatematiGa natematigigi ica maleka iliidaGadi aneotedoGoji oko’

Now we are going to tell the story of when God created people’

(3)

<struct>
<tokens>
	<tt pos="1" value="Natigide" m="ADV"/>
	<tt pos="2" value="jaaGa" m="AUX"/>
	<tt pos="3" value="jatematiGa" m="VB"/>
	<st value="j-" m="ISBJ"/>
	<st value="atemati" m="ROOT"/>
	<st value="-Ga" m="PL"/>
</tt>
	<tt pos="4" value="natematigigi" m="N"/>
<st value="n-" m="ANT"/>
</struct>
The text structure used for Kadiwéu's sample is the same structure used for Portuguese in the TBC. This structure breaks the texts into tokens (t) and works with chunks (ck) referring to token positions. With this structure it is possible to do both chained and overlapped chunks. Chunks are composed by any group of sequential tokens, and may represent a single sentence, a paragraph, a page, any quoted information for discursive purposes and even syntactic information like the tokens that compose a projection. The smallest piece of information is the split token (st) that may represent any segmented token. Tokens can be segmented for edition purposes when two words were written together and need to be split for tagging and parsing.

For the system, tokens and split tokens are equivalent: they all are treated like tokens. In the Kadiwéu structure proposed in (3), the smallest tokens are the morphemes. From a linguistic point of view, in a language like Kadiwéu, words and morphemes are treated the same way since the list of tags includes both tags associated with words (in the example above: ADV, AUX, VB, N, D and C) and tags associated with morphemes (1SBJ, ROOT, PL, ANT, NOM, MASC, 3SBJ, TRANS).

The automatic tagging process will consist of two levels: first, as for languages like Portuguese, the tagger will be run at the level of the whole sentences, in order for each word to be assigned a POS tag. At the second level, the process will be run inside of each relevant word, assigning tags to the morphemes.

The system presented above will be the basis for the syntactic annotation of texts, projected from the word-level tags, as for the languages currently annotated, but without loss of the detailed morphological information typical of polysynthetic languages. This will allow for searches combining the three levels of annotation.