**Tokenization for Occitan**

**Marianne Vergez-Couret, 09/11/2016, ANR RESTAURE**

The programme is an adaptation from the perl programme to tokenize texts in French made by Tanguy et Hathout (2007) in its extended version (that is to say with a list of exceptions). It is based on the recognition of word delimiters. Some are considered reliable - end of line, punctuation marks (comma, semicolon, colon), brackets. The other ambiguous typographical elements - space, apostrophe, dash, period, comma - are processed more precisely with a list of rules described above:

|  |  |  |  |
| --- | --- | --- | --- |
| Case | Regular expression | Examples | Tokenization |
| -n-  -z-  -u- | (-[nzu]-) | a-n-aquel (à ce) | a  -  n  -  aquel |
| d'  l'  m'  n'  s'  t'  ç'  c'  b'  z'  u' | ([dlmnst]\') | l'arbre (l'arbre) | l'  arbre |
| qu'  ns'  vs' | (\pL\*[qnv][us]\') | qu'arribi (j'arrive) | qu'  arribi |
| ent'anar | (\pL\*[e][n][t]\') | ent'anar (pour aller) | ent'  anar |
| [chiffres].[chiffres]  [lettres].[lettres] | ([\pL\d]+\.([\pL\d]+\.)+) | con.hession (confession) | con.hession |
| -vos  -nos  -ne  -se  -tu  -te  -la  -lo  -li  -las  -los  -me  -u | -(ne|[st][eu]?'?|l[ao']s?|li|me|d'|en|[nv]os|u) | tirar-u (le tirer) | tirar  -u |
| 'u  'us  'n  'v  'ns  'vs  'm  't  'i  's  'ac | \'([unv]) | que'ns cau (il nous faut) | que  'ns  cau |

Tokens that should not be segmented are recorded in an exception file named exceptions\_occitan.txt, such as “pr'aquò” or “tre que”.