

LGC, September 10-14, 2013, Universidade do Algarve, Portugal



Frequencies of occurrence of entries and subcategorization frames in LGLex lexicon with **IRASubcat**

Elsa Tolone & Romina Altamirano

FaMAF, Universidad Nacional de Córdoba, Argentina elsa.tolone@univ-paris-est.fr; romina.altamirano@gmail.com



We present a method for enlarge a lexicon (with frequencies information), that is useful for parsing and others NLP applications. We show an example enlarging the verbal LGLex lexicon of French, using several corpora extracted from the evaluation campaign for French parsers Passage. To do that, we use the results of the FRMG parser with IRASubcat, a tool that automatically acquires subcategorization frames from corpus in any language and that also allows to complete an existing lexicon. We obtain the frequencies of occurrence for each input and each subcategorization frame for 14,068 distinct lemmas.



MIN =- NThim	ŀ	N0 =: N-hum	N0 =: Nnr	Ppv	Ppv =: se figé	Ppv =: en figé	Ppv =: les figé	Nég	<ent></ent>	N0 V	N0 être V-ant	N1 =: Nhum	N1 =: N-hum	N1 =: le fait Qu P	Ppv =: lui	Ppv =: y	N0hum V W sur ce point	[extrap]	<opt></opt>
+		-	-	<e></e>	-	-	-	-	renaître	+	+	-	+	-	-	+	-	-	Max renaît au bonheur de vivre
+		-	-	se	+	-	-	-	rendre	+	-	+	+	+	-	+	+	÷	Max s'est rendu à mon opinion
+		-	-	se	+	-	-	-	rendre	+	-	+	-	-	-	-	-	-	Le caporal s'est rendu à l'ennemi
+		-	-	<e></e>	-	-	-	-	renoncer	-	-	+	÷	-	-	+	-	-	Max renonce à son héritage

http://infolingu.univ-mlv.fr/english > Language Resources > Lexicon-Grammar > Download

Gross, 1975; 1994

EXPERIMENT

To do the experiment with IRASubcat and the *LGLex* lexicon of French, we must:

- choose a corpus with millions of words, also we just only need a small part of this corpus for the experiment.
- parse the corpus with the FRMG parser, with and without the LGLex lexicon (i.e. only with the Lefff lexicon) results with FRMG-*LGLex* and with FRMG-Lefff.
- convert both the processed corpus and the *LGLex* lexicon into XML format, required by IRASubcat.
- use IRASubcat in order to add the frequencies of occurrence extracted from the big corpus into the LGLex lexicon.

• a tool that acquires subcategorization information about the behaviour of any tag class (e.g., part of speech, syntactic function, etc.) or combination of them, from corpora. • takes as input a corpus in XML format.

• the output is a lexicon, also in XML format, where each of the verbs under inspection is associated to a set of subcategorization patterns. The lexicon also provides information about frequencies of occurrence for verbs, patterns, and their co-occurrences in corpus.

• allows to integrate the output lexicon with a preexisting one, merging information about verbs and patterns with information that had been previously extracted, possibly from a different corpus or even from a hand-built lexicon.

http://www.cs.famaf.unc.edu.ar/~romina/irasubcat/

Altamirano & Alonso Alemany, 2010

FUTURE WORK

• The processed corpus is the results of the FRMG parser with *LGLex* lexicon, so it could find wrong sense.

• The next step is to consider the information on realizations, that we must extract from processed corpus, but it is not a straightforward task.

- Then we have to use the FRMG parser with Lefff lexicon only, without the LGLex lexicon influences the results.
- We could also use IRASubcat with another parser which is statistical, such as MaltParser, MSTParser, or Berkeley Parser.
- And we could do a comparison using the original lexicon and the enlarged lexicon with that different parsers to verify that the accuracy is better using more information.

RESULTS

Corpus

CPJ (Corpus Passage Jouet) with 100K sentences of AFP (Agence France-Presse), Europarl, Wikipedia and Wikisources, extracted from the corpus of the evaluation campaign (in 2009) for French parsers Passage.

<dictionary></dictionary>
<pre><entry count_oc_verb="1" verb="acheverV_1_1"></entry></pre>
<tag different_patterns="4" name="fs"></tag>
<pattern <="" count_w_verb="0" id="['obj', 'suj']" td="" total_count="1001"></pattern>
rejected_patterns_freq_test="NO">
<pattern <="" count_w_verb="0" id="['obl', 'suj']" td="" total_count="214"></pattern>
rejected_patterns_freq_test="NO">
<pattern <="" count_w_verb="1" id="['obl2', 'suj']" td="" total_count="325"></pattern>
rejected_patterns_freq_test="NO">
<pattern <="" count_w_verb="0" id="['obl', 'obl2']" td="" total_count="0"></pattern>
rejected_patterns_freq_test="NO">
representation</td

·/ • • • • •

<ids_from></ids_from>	
<entry total_count="1" verb="</td><td>acheverV_1_1"></entry>	
<tag name="</td><td>fs"></tag>	
<pattern< td=""><td>id="['obj', 'suj']"></td></pattern<>	id="['obj', 'suj']">
<s_list< td=""><td>;></td></s_list<>	;>
[]	
<td>t></td>	t>
<td></td>	
<pattern< td=""><td>id="['obl', 'suj']"></td></pattern<>	id="['obl', 'suj']">
<s_list< td=""><td>;></td></s_list<>	;>
[]	
<td>t></td>	t>
<td></td>	
<pattern< td=""><td>id="['obl2', 'suj']"></td></pattern<>	id="['obl2', 'suj']">
<s_list< td=""><td>;></td></s_list<>	;>
['12']	
<td>t></td>	t>
<td></td>	
<pattern< td=""><td>id="['obl', 'obl2']"></td></pattern<>	id="['obl', 'obl2']">
<s_lis<sup>+</s_lis<sup>	
[]	
<td></td>	
<td></td>	

Number of occurrences of verbs

LGLex with

frequencies

syntactic lexicon in

IRASubcat format

				verb	or number of verbs	$total_count$
					$\hat{\mathbf{e}}\mathbf{tre}_{2}$	63
			po	ouvoirV_1_88	60	
NI-	umber of o	oourrano	d	$evoir_{}V_1_38$	37	
1N		ccurrenc		faire2	22	
	of patt	erns	0	lireV_9_130	19	
	or part			vo	$uloir_{}V_{-}15_{-}82$	17
					2	16
	pattern	$total_count$		av	$ m oir_{}V_37E_10$	13
	['obj', 'suj']	1001 325 214 142			2	12
	['obl2', 'suj']				3	10
	['obl', 'suj']				4	9
	['att', 'suj']				3	8
	['loc', 'suj']	92			8	7
	['obi'. 'sui']	91			12	6
					14	5
					30	4
					63	3
					192	2
					740	1
					$13 \ 043$	0