A Pregroup Toolbox for Parsing and Building Grammars of Natural Languages

Denis Béchet

LINA CNRS – UMR 6241 – Université de Nantes

2, rue de la Houssinière – BP 92208

44322 Nantes Cedex 03 – France

Denis.Bechet@univ-nantes.fr

Annie Foret

IRISA – Université de Rennes 1

Campus Universitaire de Beaulieu

Avenue du Général Leclerc

35042 Rennes Cedex – France

Annie.Foret@irisa.fr

Abstract. Pregroup grammars are a mathematical formalism in the spirit of categorial grammars. They are close to logical formalisms like Lambek calculus but have a polynomial parsing algorithm. The paper presents a pregroup toolbox for parsing and building grammars of natural languages, including a parser that uses a tabular approch based on majority partial composition.

Keywords. Parser, Pregroups, Lambek Categorial Grammars, Parsing Software, XML Linguistic Ressources, Natural Language Toolbox.

Introduction

Pregroup grammars (PG) [13, 15] have been introduced as a simplification of Lambek calculus [12]. They have been used to model fragments of syntax of several natural languages: English [13, 15, Italian [7], French [1], Turkish [2], German [14, 16], Japanese [6], Persian [18] etc.

They belong to categorial and lexicalized grammatical frameworks : categorial grammars have nice relations to semantic interpretation and lexicalism has many advantages for the definition and acquisition of grammars and for parsing.

Another interest of PG is their order on primitive types, that helps grammar design with natural and compact types (less types) ; this point also allows to combine calculi, both formally and in software[11, 10]. In contrast to some other categorial variants, PG parsing is polynomial ($O(n^3)$).

Based on the PG formalism, and some extensions of it, we have programmed a pregroup toolbox, including a specific parser, and a grammar definition tool. The data are stored in XML format, to allow better interconnexions with other tools. A web version is also provided for parsing with a grammar, either from raw text, from (partially) parenthesized text or from analyzed text (as in XML treebanks).

This article explains the tool characteristics in connection with the underlying formalism, and gives an overview of the toolbox.

Pregroups

Pregroup Grammars

Definition 1 (Pregroup). A pregroup is a structure (P, ≤, ·, l, r, 1) such that (P, ≤, ·, 1) is a partially ordered monoid and l, r are two unary operations on P that satisfy for all element x in P, $x^l x \leq 1 \leq x x^l$ and $x x^r \leq 1 \leq x^r x$.

Let *Type* denote the lists of $p_{i,}^{(ni)}$, for primitive p_i where $p^{(0)} = p$, and $p^{(n)}$ stands for $(p^{(n-1)})^r$ if $n > 0$ and $p^{(n)}$ stands for $(p^{(n+1)})^l$ if $n < 0$.

Definition 2 (Pregroup Grammar). A Pregoup Grammar G is a finite subset of Σ × *Type* (Σ words, G finite). Its language L(G), a subset of Σ^+, is the set of sequence of words such that the concatenation of types entails (≤) the distinguished type s.

Parsing can be based on rewrite rules such as:

$$Xp^{(n)}q^{(n+1)}Y \rightarrow XY \text{ if } p \leq_P q \text{ and } n \text{ is even or if } q \leq_P p \text{ and } n \text{ is odd}$$

Parsing using partial and majority composition.

Rules below proceed by pairs of words (their types are separated by ,) ; thus parsing also provides a binary tree on words.

– [C] (**partial composition**) : for k in N, $X' = p_1^{(n1)} \cdot \cdot \cdot p_k^{(nk)}$, $Y' = q_k^{(nk+1)} \cdot \cdot \cdot q_1^{(n1+1)}$

$$\Gamma , Xp_1^{(n1)} \cdot \cdot \cdot p_k^{(nk)} , q_k^{(nk+1)} \cdot \cdot \cdot q_1^{(n1+1)} Y , \Delta \rightarrow \Gamma , XY , \Delta$$

if $p_i \leq_P q_i$ and n_i is even or if $q_i \leq_P p_i$ and n_i is odd , for $1 \leq i \leq k$.

– [@] (**majority composition**) : if (moreover) the result is not greater than the biggest argument (the width of |XY| must be less or equal to the maximum of the width of XX′ and the width of Y′Y).

Pregroup extended with iteration types

For iteration types p* [5], the parser is also based on partial composition rules:

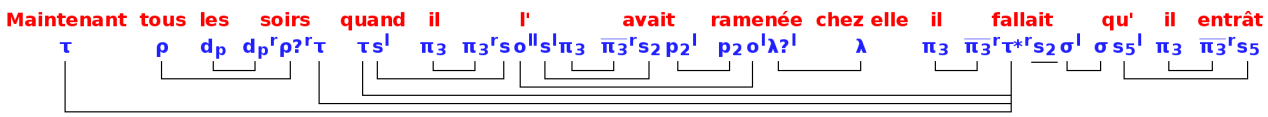
– [C] (**partial composition**) : for X′ Y' ≤ Z′ , with Z′ empty (1) or $a^{(k+1)}$:

$$\Gamma , XX' , Y'Y , \Delta \rightarrow \Gamma , XZ'Y , \Delta$$

– [@] (**majority composition**) : if (moreover) at most a half of one of the two type strings (XX′ or Y′Y) is in the result (|X′| ≥ |X| or |Y′| ≥ |Y|).

Example 3. Let us see the following sentence, in French, taken from “Un amour de Swann” by M. Proust: *Maintenant, tous les soirs, quand il l’avait ramenée chez elle, il fallait qu’il entrât.* [1]

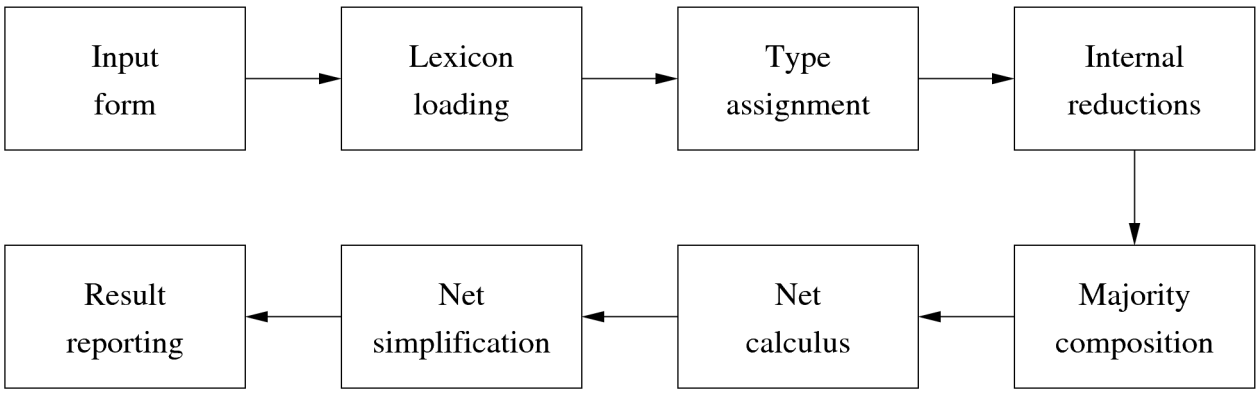
Below is a proof of correctness of assignment of types to its fragment. The primitive types used in this proof are: π_3 = third person (subject - also with a bar), p_2 = past participle, o = object, s = sentence, s_5 = subjunctive clause, with $s_5 \leq s$, σ = complete subjunctive clause, τ = adverbial phrase, λ = locative, d_p = plural determiner, ρ = special adjective. $A^?$ denotes an optional simple type. This grammar assigns s to the following sentence[1]:



[1][Now, every evening when he took back her to her home, he ought to enter.]

Parsing using majority partial composition : PPQ

A Cocke-Younger-Kasami (CYK) algorithm for pregroup grammar can be developed. The granularity of this algorithm is words (or entries if the lexicon assigns also types to a sequence of words like “pomme de terre” (potato in French)). This method presented in [4] has been implemented into a tabular parser together with other components:



Input form. This form selects one of the dictionnary, asks for the input string (alternatively, one may choose one or all samples that are associated to the selected dictionnary).

Lexicon loading.

```
<?xml version="1.0" encoding="UTF-8"?>
<grammar>
 <pregroup>
  <order inf="n" sup="n-bar"/>
...
 </pregroup>
 <sentence type="s"/>
 <lexicon>
  <w><word>whom</word>
     <type><simple atom="q’"/>
           <simple atom="o" exponent="-2"/>
           <simple atom="q" exponent="-1"/>
     </type>
  </w>
...
 </lexicon>
</grammar>
```

Grammars are described in XML files. A grammar defines a partial order on basic types, a set of basic types that are considered to form the correct sentences and a lexicon that associated to an entry (a list of tokens) a set of types. It is also possible to describe special entries with a regular expression that is useful for instance for the class of numerical number or the class of proper noun (that starts with an upper case letter). To improve the efficiency of this step that may be very long if the lexicon is big (Lefff 2.5.5 [19] has 534753 entries – the PPQ XML lexicon is a file whose size is 31,367,146 bytes), a compressed text format or a SQLite database may be used. With an indexed table, this kind of big lexicon is accessed very quickly (less than a second rather than several tens of seconds). In fact, the parser do not load all the lexicon. It selects the entries that correspond to the input string.

Type assignment to words/entries. This step assigns to each list of tokens of the input string a set of pregroup types. The input string is split into tokens using spaces and several regular expressions

For instance“l’homme” (the man in French) is segmented into two tokens: “l’” and “homme”. If the string is split in n tokens, there are n × (n + 1)/2 possible entries. Each one is searched in the lexicon and defines the initial value of the parsing matrix that computed the types associated with each segment of tokens of the input string.

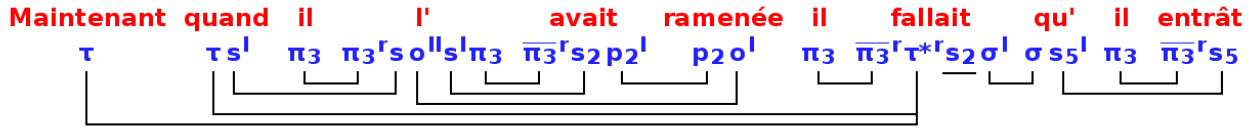
Majority partial composition of sequences of entries.

The Matrix Content

cell 1-4 q'			
cell 1-3 q' o^{ll} p_2^{l}	cell 2-4 q o^{l}		
cell 1-2	cell 2-3 q p_2^{l}	cell 3-4	
cell 1-1 q' o^{ll} q^{l}	cell 2-2 q p_2^{l} π_2^{l}	cell 3-3 π_2	cell 4-4 p_2 o^{l}
whom	have	you	seen

This step computes the parsing matrix with the result of majority partial composition (rather than using production rules of the Chomsky normal form of a context-free grammar). Of course, because we also want to describe the pregroup nets, the matrix is in fact a complex directed acyclic graph. This matrix may be displayed by the parser at the end of the report. Here, the matrix of ”whom have you seen” as input string for the Test grammar is displayed.

Net calculus. This step computes representation of the analyses of the parser. They are called pregroup nets. In a net, each entry is associated to a pregroup type and the link represents the different axioms that associates two by two the simple types of the net. This representation is close to a dependency tree except that the structure is a graph rather than a tree. Moreover, with the introduction of iterative simple types [5], a simple type can be connected to more that one other simple type as the following example shows.



Net simplifications. This step simplifies the net by suppressing the optional simple types that are not used and by taking into account cut annotations (see next section).

Result reporting. This step puts together all the results and presents them with different formats. Actually, there are three possible output formats: an HTML format useful for a web server, a text output that is suitable for terminal and an XML format that may be used if the output needs to be processed by another program.

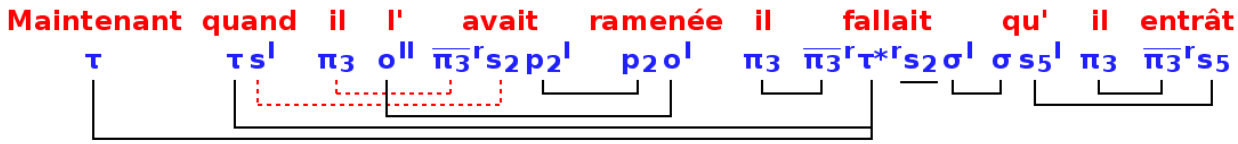
Parsing with added cuts

Grammars based on free pregroups even with optional and iterative simple types are context free. Thus, for several complex syntactical constructions some types must include a way to cross part of the environment. This is particularly the case for non projective dependencies. For instance, the French clitics are placed between the verb and the subject: In “il la mange” (he is eating it) “la” is between “il” and “mange”. The previous example shows such a construction:

The clitic *l'* is assigned $\pi_3{}^{r}so^{ll}s^{l}\pi_3$. The main simple type is o^{ll} . The rest enables the crossing of two

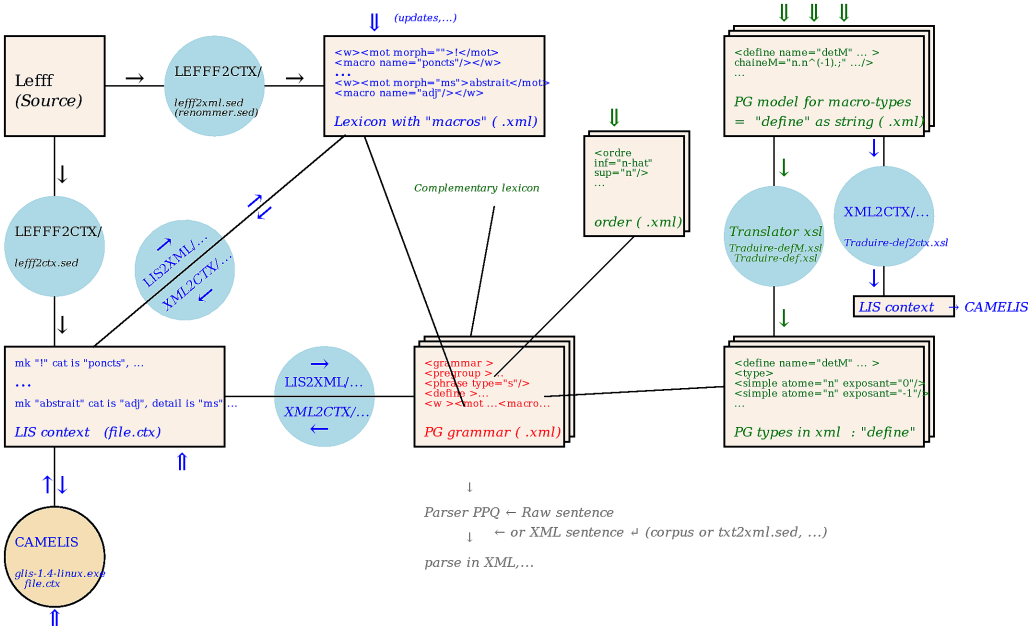
axioms (one corresponding to π_3^r and π_3, the other to s and s^l). To solve this problem, the pregroup parser PPQ can use *cut* annotations on the types assigned to special words like clitics or adverbs. These annotations that can be seen as a limited form of semantical interpretation replace two normal axiom links by a long distance axiom link. The previous French example has such annotations for the type associated to the clitic *l'*. Here two cuts have been added, one between π_3^r and π_3 and one between s and s^l . Thus on the net, the axiom between π_3 of *il* and π_3^r of *l'* and the axiom between π^3 of *l'* and π_3^r of *avait* are replaced by a single long distance axiom between π_3 of *il* and π_3^r of *avait*. Another long distance axiom is created for the other cut that links s^l of *quand* and s_2 of *avait*.

On the final picture, these special long distance axioms are shown as dashed (red) lines. Such lines can cross other axioms. The cut simple types are also erased from the picture. This interpretation is placed in the Net simplification step of PPQ parser.



Grammar construction

Other packages concern the construction of XML pregroup grammars : xslt programs have been developed for this task, including a specific mode for the French Paris7 Treebank. Another set of programs (XML2CTX, LIS2XML) provides an interface with Camelis/Glis (http://www.irisa.fr/LIS/ferre/camelis/index.html) an implementation of Logical Information Systems (LIS [10]), allowing navigation. A user can define a lexicon with Glis, then save it as a LIS context, where objects are words ; this context is then transfered to the pregroup XML format (using LIS2XML). Conversely, a pregroup grammar in XML format can be transfered to a LIS context (using XML2CTX). Camelis/Glis has been used for several prototype languages (English, French, Breton, Bambara), either for the definition or updates of parts of the grammar or for its visualisation and control (see next schema, except its upper-left dedicated to Lefff) . For a large coverage of French, Lefff [19] has been used to make a link between a French lexicon and pregroup types, through "macro-types" regrouping pregroup types in classes (see next schema).



Conclusion

The pregroup parser PPQ implements majority partial composition. This program, that can be used inline through a PHP webserver or as a command line program, uses XML files for describing a pregroup grammar. An optional indexed database can speed up the lookup in the lexicon. The result is a HTML or text page with pregroup nets as syntactical analysis that is convenient for human reading. The command line program can also produce a XML output if the result must be used by another program. This model also enables a form of semantical interpretation limited to the reduction of annotated “cuts”.

This program which is rather a test platform than a finished software has at present a large cover of the French language (however with a rough pregroup type system) and several toy lexicons for English, Breton (a Celtic language) and Bambara (an African language).

Acknowledgements

This work has benefited from useful discussions and colloborations with several collegues, a special thank to Daniela Bargelli ; we also thank A. Dikovsky, S. Ferré and E. Garel.

References

1. Bargelli D. and Lambek J. : An algebraic approach to french sentence structure. In Philippe de Groote, Glyn Morill, and Christian Retoré, editors, Logical aspects of computational linguistics: 4th International Conference, LACL 2001, Le Croisic, France, June 2001, volume 2099. Springer-Verlag, 2001.
2. Bargelli D. and Lambek J. : An algebraic approach to Turkish syntax and Morphology. Linguistic Analysis 34(1-2)
3. Buszkowski, W. : Cut elimination for the lambek calculus of adjoints. In Abrusci, V., Casadio, C., eds.: New Perspectives in Logic and Formal Linguisitics, Proceedings Vth ROMA Workshop, Bulzoni Editore (2001).
4. Béchet, D. : Parsing pregroup grammars and Lambek calculus using partial composition. Studia logica 87(2/3) (2007).
5. Béchet, D., Dikovsky, A., Foret, A., Garel, E. : Optional and iterated types for pregroup grammars. In: Proceedings of the 2nd International Conference on Language and Automata Theory and Applications (LATA 2008), March 2008, Tarragona, Spain. Lecture Notes in Computer Science (LNCS), Springer (2008) 88–100.
6. Cardinal. K. : An algebraic study of Japanese grammar. Master’s thesis, McGill University, Montreal, 2002.
7. Casadio C. and Lambek J. : An algebraic analysis of clitic pronouns in italian. In Philippe de Groote, Glyn Morill, and Christian Retoré, editors, Logical aspects of computational linguistics: 4th International Conference, LACL 2001, Le Croisic, France, June 2001, volume 2099. Springer-Verlag, 2001.
8. Degeilh, S., Preller, A. : Efficiency of pregroup and the french noun phrase. Journal of Language, Logic and Information 14(4) (2005) 423–444.
9. Dosen, K. : Cut Elimination in Categories. Kluwer Academic publishers, Dordrecht, Boston, London (1999).
10. Ferré S. and Ridoux O. : An Introduction to Logical Information Systems. Information

Processing & Management 3(40): 383–419, 2004.

11. Foret, A. : Pregroup calculus as a logical functor. In: Proceedings of WOLLIC 2007. Volume LNCS 4576., Springer (2007)
12. Lambek, J. : The mathematics of sentence structure. American Mathematical Monthly 65 (1958) 154–170.
13. Lambek, J. : Type grammars revisited. In Lecomte, A., Lamarche, F., Perrier, G., eds.: Logical aspects of computational linguistics: Second International Conference, LACL ’97, Nancy, France, September 22–24, 1997; selected papers. Volume 1582., Springer-Verlag (1999).
14. Lambek. J. : Type grammar meets german word order. Theoretical Linguistics, 26:19–30, 2000.
15. Lambek J. : From word to sentence, a computational algebraic approach to grammar. Polimetrica, Milan, 2008.
16. Lambek J. and Preller A. : An algebraic approach to the german noun phrase. Linguistic Analysis, 31:3–4, 2003.
17. Oehrle, R. : A parsing algorithm for pregroup grammars. In Moortgat, M., Prince, V., eds.: Proc. of Intern. Conf. on Categorial Grammars, Montpellier (2004).
18. Sadrzadeh. M. : Pregroup Analysis of Persian Sentences, in C. Casadio and J. Lambek (Eds.), Recent Computational Algebraic Approaches to Morphology and Syntax, Polimetrica, Milan, 2008.
19. Sagot, B., Clément, L., de la Clergerie, E., Boullier, P. : The lefff2 syntactic lexicon for french: architecture, acquisition. In: LREC’06. (2006).