A Description of the French Nucleus VP Using Co-occurrence Constraints

François Trouilleux

A DESCRIPTION OF THE FRENCH NUCLEUS VP USING CO-OCCURRENCE CONSTRAINTS

FRANÇOIS TROUILLEUX

Abstract

This article presents a fully operational formal grammar of the French nucleus verb phrase. The grammar is implemented in NooJ, with a focus on constraint specification. We take the Properties formalism of (Bès, 1999) as a reference and show how requirement and exclusion properties may be implemented in NooJ, introducing a new type of constraint.

Introduction

In (Bès, 1999), Gabriel G. Bès proposed a new formalism for syntactic description, called "Properties", which he exposed together with a description of the French nucleus verb phrase. Properties are a constraint system; we propose to describe with NooJ (Silberztein, 2003) a set of strings close to that of (Bès, 1999), showing how the different Property types may be coded in NooJ. On this occasion, we introduce a new type of constraint in NooJ: co-occurrence constraints.

This introductory section sets the scene with an overview of the Properties formalism, an informal global definition of the language to be defined and an exposition of our adequacy criteria. We will then refine of our language definition in two steps: (i) by specifying word categories and linearity constraints, as well as optionality and uniqueness, and (ii) by specifying co-occurrence constraints.

Properties

The Properties formalism consists in a set of seven different types of formulas on categories ("properties"), which, interpreted as a conjunction, denote a language[1]. As pointed out in (Trouilleux, 2003), the Properties of

[1] (Trouilleux, 2007) shows that, ignoring *fléchage* ("arrowing") properties, a description in the properties formalism may be interpreted as the intersection (*i.e.* conjunction) of finite-state languages. The *fléchage* property type codes the dependencies between the words within the strings; it does not contribute to the

(Bès, 1999)[2] may be viewed as an extension of the decomposition of information initiated by the ID/LP (immediate dominance/linear precedence) GPSG formalism (Gazdar *et al.*, 1985). LP rules have a direct correspondence in so-called "*linearity* properties", while the information expressed by a set of ID rules will be expressed by five different property types:

- the *alphabet* property specifies the set of categories which may occur in a string of the targeted language (henceforth S_L);
- the *uniqueness* property specifies which categories may only appear once in S_L;
- the *obligation* property specifies which categories are mandatory in S_L, possibly disjunctively;
- *requirement* properties state that if some category appear in S_L, then some other category must also be present; this property type includes agreement constraints;
- *exclusion* properties state that two categories may not co-occur in S_L.

Targeted Language

We define the set of strings to be specified by our grammar in terms of the EASY annotation scheme (Gendner and Vilnat, 2004), so that we will be able to test the grammar against the EASY corpus in future work. The targeted strings are the NV and PV constituents of the EASY scheme, extended to the right by the negation adverb or the past participle(s), and possible intermediate words, e.g. adverbs or pronouns *tout* or *rien*. Here are a few examples (targeted strings are underlined):

(1) Pierre <u>ne le lui a pas donné</u>. / Pierre did not give it to him/her.
(2) <u>Ils ont tous été mangés</u> hier. / They all have been eaten yesterday.
(3) Pierre <u>dit</u> à Marie <u>de ne pas les revoir</u>.
Pierre says to Marie not to see them again..

The combination of infinitives with support verbs or modal auxiliaries (e.g. *il va venir, il la fait travailler*) are not part of our targeted language.

---

definition of languages as a sets of strings, but rather provide an annotation of the defined strings. Even though the NooJ variable system could presumably be used to code such dependencies, we do not address this issue in this paper.

[2] Non French speaking readers may consult (Blache, 2004) for a description in English of the Property formalism.

Adequacy criteria

Our goal is simple descriptive adequacy: precisely specify the set of well-formed sequences, ruling out ill-formed ones. We add two restrictions to the definition of our targeted language: (i) our grammar does not deal with phonological or prosodic matters (e.g. the incorrect *je aime* and *j'y irai* will be specified as well as the correct *j'aime* and *j'irai*), and (ii) it does not account for the government of clitics by verbs: any clitic pronoun will combine with any verb, and no verb will require some specific pronoun. We leave this major issue for future work.

The major challenges in the language we intend to describe are the handling of clitic pronouns and anaphoric quantifiers (*tous*), the choice of the auxiliary verb depending on the past participle and reflexiveness, past participle agreement, and the co-occurrence of items which may be separated by several words (e.g. *ne* and *pas* in *ne me l'a-t-il donc pas donné*).

The treatment of French clitic pronouns has given rise to many articles, so that the constraints on these pronoun sequences are quite well known. A question which has been debated is whether French clitic pronouns should be dealt with *lexically* or *post-lexically* (cf. Heap and Roberge, 2001, §3.3.2). We chose to describe our language with a NooJ *syntactic* grammar. However, this does not mean that we took a strong position on the lexical/post-lexical issue, for two reasons: (i) our grammar is descriptive only, it is not intended to have cognitive adequacy and (ii) arguments in favour of the lexical treatment of French clitic pronouns are typically phonological and we set aside such matters. The choice of a syntactic grammar, however, is supported by the fact that, in compound tenses, the pronoun, while attached to the auxiliary verb, is governed by the past participle (e.g. in (4), *l'* is governed by *lavé*), and as (Abeillé and Godard, 1996) points out, "it is clear that compound tenses concern syntax more than morphology".

(4) Il ne l'a donc pas bien lavé.
He NEG it has thus not well washed. / He thus didn't wash it well.

(Miller and Sag, 1997) proposed a lexical treatment of French clitic pronouns in HPSG. The system produces "cliticized words" from the composition of verbs with clitics, checking and reducing the verb's argument structure as clitics are added. To account for compound tenses, "the tense auxiliaries and their participle complements share arguments"; but they must do so at the syntactic level, so it is most likely that the system either requires multiple auxiliary verb entries (one for each possible type of past participle in terms of argument structure) or generates

all possible clitic-auxiliary combinations. Such a treatment at the lexical level would be possible in NooJ; e.g. in (4), the sequence *il ne l'a* would be analysed as a kind of compound. However, one would need to record all the components of the *il ne l'a* compound using *ad hoc* features in order to correctly combine it with past participles and the negation adverb. We preferred to handle all combinatorial aspects at the same level.

Categories, Linearity, Obligation and Uniqueness

Let us go one step further in the description of our targeted language by specifying the alphabet of categories, linearity constraints, obligation and uniqueness properties. Word order is very much fixed in the sequences we are trying to describe, so that we will introduce categories together with linearity constraints. As for obligation, things are simple: there is only one mandatory category: one non past participle verb form, which we will refer to as VF.

It is customary to account for the rigid order of clitic pronouns to the left of VF using a table as Table 1[3] (see e.g. Bonami and Boyé, 2007).

	NOM	NEG	ACC1r DAT1r ACC1 DAT1	ACC2	DAT2	*y*	*en* ACC3	VF
1	je							dors
2	vous	ne						dormez
3	il		se	l'				achète
4	tu			le	lui			achètes
5	elle			les	leur	y		donne
6	il		vous	les			en	rapporte

Table 1. Clitic slots to the left of the mandatory verb form.

Column headers give the non-terminal symbols used in the grammar for the corresponding items. Subject pronouns come first, followed by the negation particle *ne*, followed by the complement pronouns. Table 1 shows five different slots[4] for these pronouns: (i) the series *me, te, se, nous, vous*, which may either be accusative or dative and be reflexive or not, (ii) third person non reflexive accusative (*le, la, les*), (iii) third person

[3] 1: *I sleep*, 2: *you NEG sleep*, 3: *he himself$_D$ it$_A$ buys*, 4: *you it$_A$ him$_D$ buys*, 5: *she them$_A$ them$_D$ there gives*, 6: *he you$_D$ them$_A$ from-there brings-back.*

[4] In addition to the five pronoun slots of Table 1, our grammar includes an additional slot for an ethical dative pronoun at the beginning of the pronoun sequence.

non reflexive dative (*lui, leur*), (iv) pronoun *y* and (v) pronoun *en*. Word order is fixed, ignoring dialectal variations.

Slots on the right-hand side of the verb are given in Table 2[5]. The first column recalls that these items may combine with items on the left-hand side of the verb. Two types of clitic pronouns in complementary distribution may appear to the right of the verb: complement pronouns with an imperative verb, and subject pronouns with other inflected verb forms. It must be noted that one may have only one subject pronoun, but possibly several complement pronouns, which the table does not indicate. After the pronoun(s) may come a negation adverb (PAS), a subject or object anaphoric plural quantifier(s) (TS) or the pronouns *tout* or *rien*. In final position comes the main past participle, which may be preceded by the past participles of *avoir* (*eu*) or *être* (*été*) in the double-compound tenses ("temps surcomposés") or in the passive voice.

	cf. Tab. 1	VF	NOMi *obj. pro.*	PAS	TS	TS *tout*	EU	ETE	PP
1	n'	aime	-t-il	pas		tout			
2		donne	-le-moi						
3	ils	aiment			tous	tout			
4	ne lui	a	-t-il	pas				été	donné
5	il	est							mangé
6	il	a					eu		mangé
7	elles	ont			toutes			été	aimées
8	il	a					eu	été	aimé

Table 2. Slots to the right of the mandatory verb form.

	Prep	NOM	NEG	PAS	TS	TS *tout*	*obj. pro.*	VF
1	**pour**		ne	**pas**			nous les	acheter
2			ne			**rien**	lui	dire
3	**à**				**tous**	**tout**		acheter
4	**de**				**toutes**		se les	acheter

Table 3. Additional options to the left of infinitive VFs.

If one considers infinitive verb forms, Table 1 must be completed by Table 3[6]. Column *Prep* introduces prepositions, column *PAS* introduces a

[5] 1: *doesn't he like everything*, 2: *give it to me*, 3: *they all like everything*, 4: *has it not been given to him*, 5: *he is eaten*, 6: *he has had eaten*, 7: *they have all been loved*, 8: *he has had been loved*.

[6] 1: *in order not to buy them for us*, 2: *say nothing to him*, 3: *to all buy everything*, 4: *to all buy them for themselves*.

possible negation adverb *before* the infinitive verb and the next columns introduce the quantifier *tous* and pronouns *tout* and *rien*, which may combine with complement clitic pronouns. Subject pronouns (*NOM*) may not occur with infinitive or present participle verbs.

Adverbs may appear in between some columns, as (4) shows, and, as (Abeillé and Godard, 1996) points out with (5), long insertions are possible between the auxiliary verb and the past participle.

(5) Il a d'un seul coup, si l'on peut dire, changé de visage.
He has all at once, if one may say so, changed his face.

The grammar we present do not claim to account for such cases, except by allowing one optional adverb in some positions.

The information specified in Tables 1 to 3 may be represented by the NooJ graph in Fig. 1. This graph specifies the categories[7], in appropriate order, as well as optionality, *i.e.* the alphabet, linearity and obligation properties. We view this graph, with its train of categories and its bypasses marking optionality, as typical of a fixed word order language. K. Bogacki and E. Gwiazdecka's article, in this volume, shows how free word order can be dealt with in NooJ.

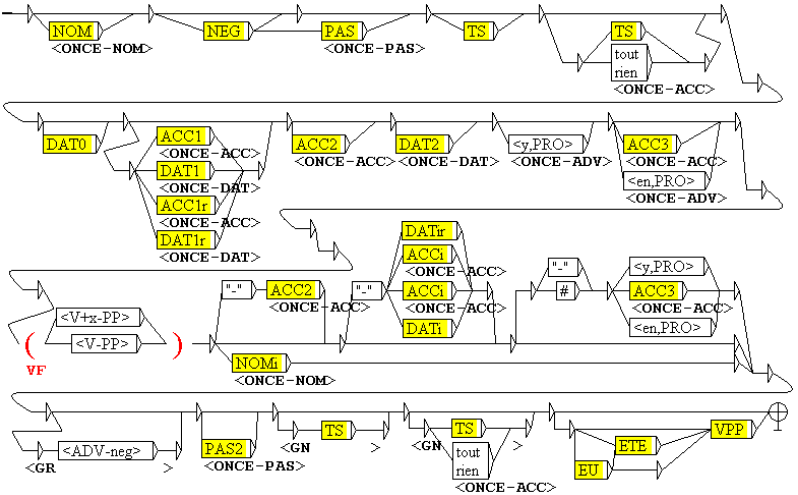

Fig. 1. Alphabet, linearity, obligation and uniqueness graph.

[7] All nodes dominate only one terminal category, except PAS, ETE and VPP, which include optional adverbs. Initial prepositions are omitted for space reasons.

A series of NooJ **<ONCE>** constraints (cf. M. Silberztein's article in this volume), corresponding to uniqueness properties, reduces the set of possible combinations. They state that there is only one subject pronoun, one accusative pronoun, one dative pronoun, one adverbial pronoun and one *pas* negation adverb. Note that they automatically limit the number of clitic pronouns to three and determine their interpretation: for instance, if *y* and *en* co-occur, then *en* is an accusative (cf. *il y en a*, "there are some", it may otherwise have several non accusative functions), if *le* (ACC2) co-occurs with a slot 1 pronoun, then this pronoun is a dative, etc.

The graph also codes two co-occurrence constraints: (i) PAS before the verb may only occur with NEG (cf. top left corner) and (ii) the auxiliary past participles EU and ETE may only occur with VPP (cf. bottom right corner). Coding co-occurrence constraints between adjacent items as these is well done graphically. However, when there is a co-occurrence constraint between two items which are not necessarily adjacent, coding the constraint graphically will require duplicating the intermediate paths. For instance, to specify graphically that PAS2 (bottom line) after the verb requires NEG (top line), one would have to duplicate everything that goes in between. We will show in the next section that we can save such node duplication using co-occurrence constraints.

Co-occurrence constraints

Tables 1 to 3 and the graph in Fig. 1 ignore co-occurrence constraints which do exist between some items. Our grammar then actually contains a *variant* of the Fig. 1 graph, annotated by a series of co-occurrence constraints. We cannot reproduce this graph here and invite the reader to download it from our web page[8]. We will endeavour to give the reader all the necessary information to interpret the grammar: elements of NooJ syntax, a couple of examples and a complete specification, in natural language, of the implemented constraints.

Elements of NooJ Syntax

Variables. In NooJ, co-occurrence and agreement constraints are specified using variables. Variables are set as labelled parentheses around a node. They record the lexical feature information from the items that

[8] http://lrl.univ-bpclermont.fr/spip.php?rubrique48. The grammar is available as a NooJ project file or as a series of screen captures.

match the nodes and are then used in constraints specified in angle brackets along grammar paths to perform tests on the recorded information.

NooJ makes a distinction between global and local variables. To allow importation in larger grammars, our grammar only makes use of local variables. Access to local variables is limited: a constraint on a local variable in a graph G can only access variable values defined *in* or *below* G. To decide between possibly competing variable values, a breadth-first, left-to-right search procedure is used. Local variables are useful when constraints are limited to syntactic constituents. E.g. in (6), there are two occurrences of the subject-verb agreement constraint: *Il pense* and *tu dors*.

(6) Il pense que tu dors. / He thinks that you sleep.
(7) ($_{P}$ ($_{SN}$ Il) ($_{SV}$ pense ($_{PS}$ que ($_{P}$ ($_{SN}$ tu) ($_{SV}$ dors) *))) *).

Scope of this constraint is limited: *dors* should not agree with *Il*. With a classic constituent structure as in (7), and local variables on the subject pronoun and verb, the agreement constraint should be specified at the P level; the two instances of the constraint (marked by the stars in (7)) will each be evaluated with the appropriate set of values, thanks to the locality constraint for *tu dors* and thanks to the search procedure for *Il pense*.

Co-occurrence constraints. Formally, one may distinguish three types of co-occurrence constraints, summarized in Table 4. The third constraint type is new in NooJ and has been developed by Max Silberztein following our proposition at the NooJ Conference at INALCO, Paris. Co-occurrence constraints have negative counterparts thanks to the negation operator (!).

Syntax	*Semantics*
<$V1$N1=$V2$N2>	The value of attribute **N1** recorded in variable **V1** is equal to the value of attribute **N2** recorded in variable **V2**. This is typically used for agreement constraints, with **N1** and **N2** identical.
<$V1$N1=Value>	The value of attribute **N1** recorded in variable **V1** is equal to the value **Value**. **$N1** may be replaced by _ to denote the *lemma* recorded in **V1**.
<$V1>	The variable **V1** is defined, *i.e.* it records some value(s).

Table 4. Co-occurrence constraint types.

It must be noted that the first two constraint types are considered satisfied if any of the variable referred to is undefined. To avoid unnecessary constraint checking by NooJ and save computational time, agreement

constraints should be set on the path of the *less frequent* of the two variables. In our grammar, pronoun-verb agreement constraints are thus specified on the pronoun paths: they are optional, while the verb is mandatory.

Constraints of the second type are all set on some *optional* item path and refer to a *mandatory* item.

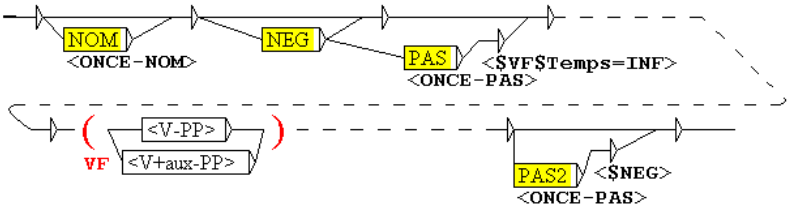

Fig. 2. Two co-occurrence constraint examples.

Fig. 2 gives an example of co-occurrence constraints with the annotation of the two PAS occurrences. The first requires that VF be an infinitive (and requires NEG by graph design), the second requires that there is a negation particle ($NEG is defined in the NEG node), but sets no constraint on the verb. Dashed lines represent all the intermediate nodes it would be necessary to duplicate if the constraints were not available.

Co-occurrence constraints for the French nucleus verb phrase

Tables 5 to 7 list the co-occurrence and agreement constraints specified in the grammar, using an informal natural language formulation which documents the actual downloadable formal grammar. First columns refer back to annotations in the downloadable NooJ graph.

To correctly interpret the formulas, a few definitions are in order:

- A *reflexive pronoun* is one of *me, te, se, nous, vous, toi* which agrees in number and person with the subject. Such pronouns are identified using agreement constraints[9]; a dedicated variable $REF is instantiated when the constraints are satisfied.
- An *auxiliary past participle* is either *eu* or *été* when they are followed by another past participle (e.g. as in *il a été mangé*).
- Sequences ending with a past participle (PP) fall into two categories: *subject oriented* or *object oriented PP phrase*, defined in Table 6.

[9] Additionally, we consider that reflexive pronouns agree in gender with the subject, even though none is overtly marked in gender. We consider that in *elle s'est trompée*, the past participle agrees with the object pronoun *s'*.

The nice thing with properties is that each formula constrains the language on one very specific point, making it possible to illustrate the constraint with specific examples and counter-examples. Tables 5 to 7 give such examples; see more examples in the grammar contract[10].

C1	A subject pronoun forbids that VF be an infinitive, present participle or imperative form.	**il dormir* **il dormant* *#*tu dors*$_{IP}$
C2	A negation adverb before VF requires that VF be an infinitive.	*ne pas dormir* **ne pas dort*
C3	A *tous* quantifier or a non clitic pronoun before VF requires that VF be an infinitive.	*tout manger* *#*tout*$_{OBJ}$ *mange*
C4	A *tous* quantifier either requires a plural subject or a plural accusative pronoun, except *en*.	*il les aime tous* *ils l'aiment tous* **il l'aime tous*
C5	An ethical dative pronoun forbids that VF be a second person form.	*il te lui donne* **tu te lui donnes*
C6	A slot 3 clitic pronoun forbids a slot 1 clitic pronoun.	**il se lui donne* **il me leur donne*
C7	Clitic pronouns before VF either forbid that VF be an imperative, or require that VF is an imperative and there is a negation particle.	*ne le mange pas* *#*le mange*$_{IP}$
C8	An auxiliary VF requires a past participle head verb.	*il a dormi*
C9	A non auxiliary VF forbids a past participle.	**il part dormi*
C10	A slot 2 clitic pronoun to the right of an imperative may end the clitic sequence if it is not a marked unstressed form (*me, te*) and may be followed by a slot 3 pronoun if it is not a marked stressed form (*moi, toi*). (This is the purist's imperative.)	*aime-moi* **aime-me* *donne-m'en* **donne-moi-en*
C11	Complement clitic pronouns after VF require that VF be an imperative and forbid there is a negation particle.	*mange-le* **mangeait-le* **ne mange-le*
C12	A negation adverb requires a negation particle	*il ne mange pas* **il mange pas*
C13	The auxiliary past participle *eu* requires a past participle head verb with feature `Aux=a` (assigned to verbs that require *avoir* as well as pronominal verbs).	*il a eu dormi* *il s'est eu absenté* **il est eu parti*
C14	The auxiliary past participle *été* requires *avoir* as VF.	*il a été mangé* **il est été mangé*
C15	The passive voice forbids an accusative pronoun.	**il l'est mangé*

Table 5. Requirement and exclusion constraints.

[10] Strings preceded by # cannot actually be tested because of lexical ambiguity.

PP1	A *subject oriented PP phrase* is one where either VF is *être*, or *été* is present, and the PP verb either is marked as requiring *être* as an auxiliary or is marked as requiring *avoir*, is transitive and there is no reflexive pronoun (this is the passive voice).	*il est parti* *il a été parti* **il a parti* *il est mangé* **il est dormi*
PP2	An *object oriented PP phrase* is one where the PP verb is marked as requiring *avoir*, and VF is either *avoir* with no reflexive pronoun nor *été*, or *être* with a reflexive pronoun.	*il a dormi* *il l'a mangé* *il s'est mangé* **il s'a mangé*

Table 6. Properties of the past participle phrases.

A1	A subject pronoun agrees in person and number with VF.	**tu dort* **il dorment*
A2	A *tous* quantifier agrees in gender with the subject or the direct object pronoun (see also C4, Table 5).	*ils sont tous partis* **elles sont tous partis*
A3	In a subject oriented PP phrase, the past participle agrees in number with the subject[11] or it may be singular if the subject is second person plural.	*ils sont partis* **ils sont parti* *vous êtes parti*
A4	In a subject oriented PP phrase, the past participle agrees in gender with the subject.	*elles sont parties* **elles sont partis*
A5	In an object oriented PP phrase, if there is an accusative pronoun, the past participle agrees in number and gender with the accusative pronoun, otherwise it is masculine singular[12].	*ils ont mangé* **ils ont mangés* *il les a mangés* **il les a mangé*

Table 7. Agreement constraints.

Conclusion

.Our goal in this paper was twofold: demonstrating the coding of co-occurrence constraints in NooJ and specifying a fully operational grammar. Looking at our *large* graph, the reader might wonder what is the point in this style of coding. The point is *modularity*. The graph is large because it is made of an accumulation of observations, but most of these observations are fairly simple and it is easy to add or remove constraints.

The grammar should be primarily evaluated against its "contract", *i.e.* a set of strings marked as grammatical and ungrammatical, designed, as seen in Tables 5 to 7, to illustrate each constraint to be satisfied. In that sense, the grammar is not only a formal description of a set of strings, but also a test suite for the French nucleus verb phrase. Prior to evaluating the

[11] As the subject agrees in number and person with the verb, we code number or person agreement with the subject as an agreement with the verb.
[12] The grammar does not deal with other cases of agreement with the direct object.

grammar against running text, bear in mind that the specified language is a purist's version of the standard modern French nucleus verb phrase—not appropriate for all uses. Tests on corpus showed some strings were not identified because they were ungrammatical (e.g. **je vous ait dit*). This is good for error detection, bad for information extraction. We also found a few strings where word order is not the one our grammar allows: *pour n'y plus revenir*, *sans lui rien apprendre*. However, these are from XIXth century literature and sound outdated; modern word order would be *pour ne plus y revenir*, *sans rien lui apprendre*, which our grammar does recognize. The problem is an observation problem: what is the good set of strings to specify? (Bès, 1999) showed that properties could easily accommodate variations in the observations; NooJ can now have this quality, with our transposition of properties into the NooJ formalism.

References

Abeillé, A. and D. Godard, 1996, « La complémentation des auxiliaires français ». *Langages*, n° 122, 32-61, Paris : Larousse

Bès, G. G. 1999, « La phrase verbale noyau en français », *Recherches sur le français parlé*, 15:273-358. Université de Provence.

Blache, P. 2004, "Property Grammars: a Fully Constraint-Based Theory", *Constraint Solving and Language Processing*, 1-16, Springer

Bonami, O. and G. Boyé, 2007, "French pronominal clitics and the design of Paradigm Function Morphology". *In* Booij, G., B. Fradin, A. Ralli, and S. Scalise (eds), *Online Proceedings of the Fifth Mediterranean Morphology Meeting*.

Gazdar, G., E. H. Klein, G. K. Pullum, and I. A.Sag, 1985, *Generalized Phrase Structure Grammar*. Oxford: Blackwell.

Gendner, V. and A. Vilnat, 2004, Les annotations de référence PEAS. perso.limsi.fr/Individu/anne/Guide/PEAS_reference_annotations_v2.2.html

Heap, D. and Y. Roberge, 2001, « Cliticisation et théorie syntaxique, 1971-2001 ». *Revue québécoise de linguistique*, 30(1):63-90.

Miller, P. and I. A. Sag, 1997, "French clitic movement without clitics or movement." *Natural Language and Linguistic Theory*, 15:573-639.

Silberztein, M. 2003, *NooJ Manual.* http://www.nooj4nlp.net (220 pages, updated regularly).

Trouilleux, F. 2003, « Note de lecture sur *Philippe Blache, Les Grammaires de propriétés* », *TAL*, 44(2):256–259. Hermès.

Trouilleux F. 2007, "Specifying Properties of a Language with Regular Expressions", *Proceedings of RANLP*, 609–613, Bulgaria