Travaux du 19ème CIL | 19th ICL papers

Congrès International des Linguistes, Genève 20-27 Juillet 2013
International Congress of Linguists, Geneva 20-27 July 2013

19th International
Congress of Linguists
July 21-27 2013
Geneva - Switzerland

Gabriela BILBIIE

Université Paris Diderot, France
gabriela.bilbiie@linguist.univ-paris-diderot.fr

A quantitative study on Right Node Raising in the Penn Treebank

oral presentation in session: 10 Varia (Stephen Anderson)

Published and distributed by: Département de Linguistique de l'Université de Genève, Rue de Candolle 2, CH-1205 Genève, Switzerland
Editor: Département de Linguistique de l'Université de Genève, Switzerland
ISBN:978-2-8399-1580-9

A quantitative study on Right Node Raising in the Penn Treebank

Gabriela Bîlbîie[1]
[1]**Laboratoire de Linguistique Formelle, Université Paris Diderot-Paris 7 - Labex EFL**

gabriela.bilbiie@linguist.univ-paris-diderot.fr
International Congress of Linguists

21–27 July 2013, Geneva

Plan

Ellipsis – a challenge for grammar

Ellipsis : a form/meaning mismatch (*significatio ex nihilo*)

1. part of the material necessary for the interpretation is missing in the syntactic structure ('incomplete' syntax) ;
2. the missing material is recovered from an antecedent in the context.

- *Descriptive problem :* **A mass of elliptical constructions**, on the basis of several criteria, e.g. syntactic function of the missing material (head or dependent), syntactic context (coordination, subordination ; dialogue), ellipsis directionnality (forward vs. backward ellipsis).
  ⇒ Sometimes, unstable terminology.
- *Theoretical problem :* **A plethora of competitive analyses**, with respect to the level at which reconstruction of the missing material takes place : syntactic reconstruction vs. semantic reconstruction.
  ⇒ Unsolved theoretical problems.

The importance of corpus for ellipsis studies

- Data issues : in the literature on ellipsis, constructed data ; significant variation in acceptability judgments across speakers ; sometimes contradictory data. Use of empirically attested data prevents these problems.
- Need for the contextual dimension (cf. definition of ellipsis above) :
  - investigation of the contextual constraints applying on various elliptical constructions ;
  - observation of preferences between structures with and without ellipsis.
- Quantitative issues : frequency measures of several factors, e.g. which constructions are the most frequent, which constraints are strict or less strict.
- Corpus crucially gives a safer ground to assess the facts and evaluate competing analyses : allows one to choose the best suited analysis based on data.

Corpus investigation is not optional, but a must for ellipsis !

Few corpus studies on ellipsis

English

- American English (Meyer 1995) : *Brown Corpus* (80 000 words ; edited written English) ; *International Corpus of English* (16 000 words ; spoken English)
- British English (Greenbaum and Nelson 1999) : selection of 82 spoken and written texts (176 968 words) drawn from the British component of ICE (*ICE-GB*)

German and Dutch (Harbusch 2011)

- German : *TIGER* (50 474 sentences of written newspaper text) ; *VERBMOBIL* (38 328 spoken sentences)
- Dutch : *ALPINO* (7 153 sentences of written newspaper text) ; *CGN2.0* (130 594 spoken sentences)

Overview of existing corpus studies

Three main elliptical constructions are investigated : Left Peripheral Ellipsis, henceforth LPE (1), Gapping (2) and Right Node Raising, henceforth RNR (3).

(1) The Australian stopped trying to talk a pidgin I could understand, and ___ spoke strange words from deep in his chest.

(2) The top of the sample was nearly flat and the bottom ___ hemispherical.

(3) It is important to consider ___ and experimentally verify this influence...

- LPE – most favorable type of ellipsis in English (86%)
- Gapping and RNR – less favorable (5,5% et 2%, respectively)
  - *[...] given the attention in linguistic theory devoted to the discussion of E-Ellipsis [Gapping], it is quite surprising to see how a truly unproductive process it is.* (Meyer 1995 : 258)
  - *C-Ellipsis [RNR] is a relatively rare form of ellipsis [...]* (Meyer 1995 : 266)

Why an additional corpus study ?

Limits of the previous corpus studies : very biased results

1. The choice of the 'elliptical' constructions under investigation : are treated at the same level constructions unanimously recognized as elliptical (e.g. Gapping and RNR) and constructions lending themselves to a non-elliptical account (e.g. LPE with 'elliptical' subjects, which could receive a non-elliptical treatment in terms of a coordination of verbal phrases instead of clausal coordination with subject ellipsis).
   - All these studies show that the latter type generally has the highest frequency, and hence, a wrong quantitative interpretation of the ellipsis phenomenon in general.
2. The syntactic domain under investigation : only interclausal ellipses are taken into account.
   - Some types of ellipsis are most frequent at the sub-clausal level, e.g. RNR.

Plan

The Penn Treebank (PTB)

A large annotated corpus of American English (4.5 million words).

- Main authors : Mitch Marcus, Ann Taylor (University of Pennsylvania). A three-phases project, started in 1989.
- Sources : 4 sub-corpora of written and spoken English
  - written : *Wall Street Journal* (1989) and *Brown Corpus* (1961)
  - spoken : part of Air Travel Information Services [*ATIS-3*] (1995) and part of *Switchboard* corpus (1991)
- Annotations :
  - morpho-syntactic annotation (POS tags, lemma)
  - constituent annotation (parsing)
  - 4 different functional tags : form/function discrepancies (-ADV, -NOM), grammatical role (-PRD, -SBJ, etc.), adverbials semantics (-BNF, -LOC, -TMP, etc.), miscellaneous (-CLR, -CLF, etc.)
  - dysfluency annotation (for Switchboard corpus)
- Use of the Stanford **Tregex** utility for matching patterns in trees, based on node descriptions or relationships between tree nodes.

RNR annotation in the PTB

- **Right Node Raising** : an elliptical phrase lacking a dependent or the head (in final position) precedes a complete phrase which determines its interpretation.

  (4) [John made ___] and Mary sold a piece of furniture.

- Phenomenon treated in the PTB in terms of null elements.
- For discontinuous constituents, use of a Pseudo-Attach function : a method of showing that non-adjacent constituents are related.

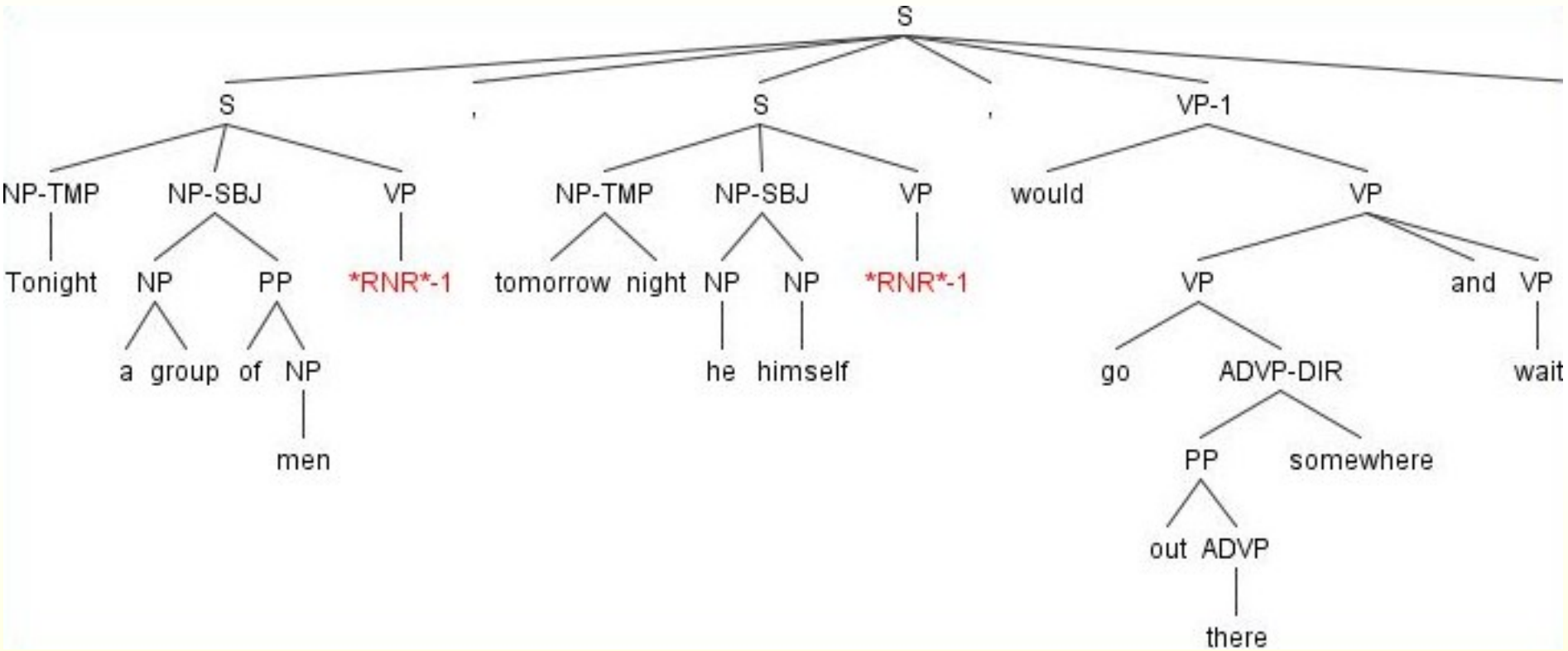
  ⇒ This mechanism is used for several phenomena, such as extraction, extraposition, structural ambiguity and RNR.

Two RNR-types in the PTB

- Among the sub-types of Pseudo-Attach, *RNR*-attach is used for shared constituents (= factorized chunk), i.e. the cases in which a constituent should be interpreted simultaneously in more than one place.
- RNR co-indexing : an index number added to the label of the original constituent is incorporated into the 'null element'.
- Two different annotations for the RNR in the PTB :
  - **Regular RNR** : two (or more) *RNR*-tags co-indexed with the constituent with which the null is associated.
  - **Parenthetical RNR** : a single *RNR*-tag only in the last segment, which has a node labeled PRN.

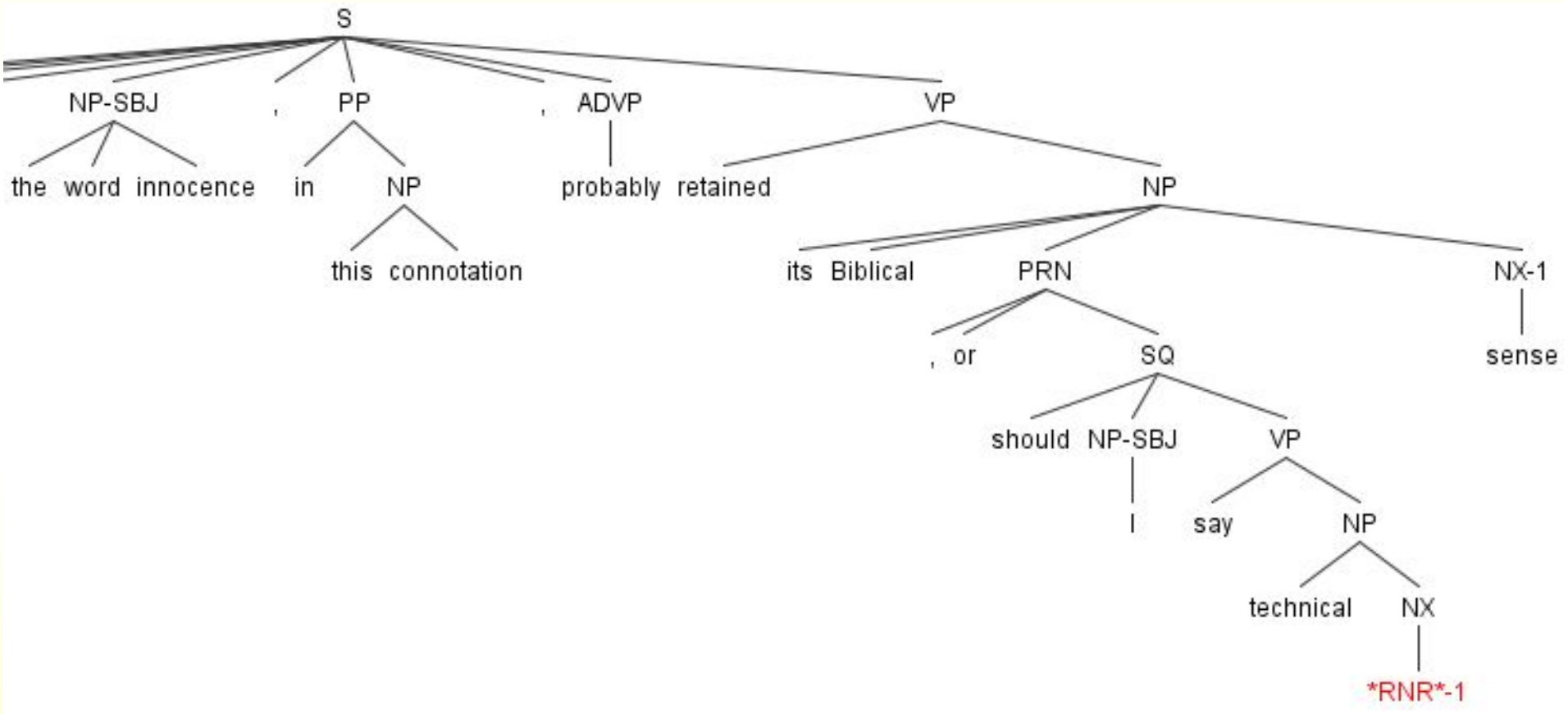
Sample with regular RNR in the PTB

(5) Tonight a group of men *RNR*-1, tomorrow night he himself *RNR*-1, would go out there somewhere and wait. (brwn-12426)



Sample with parenthetical RNR in the PTB

(6) ...the word innocence, in this connotation, probably retained its Biblical, or should I say technical *RNR*-1 sense... (brwn-16291)



RNR occurrences in the PTB

- Extraction of 474 occurrences (in 469 sentences), matching the pattern /*RNR*-[0-9]/.
  - Discrepancy between speech and writing : RNR seems to be more frequent in writing (314 occurrences) than in spoken language.

	WSJ	**Brwn**	**Swbd**	**Nb.occurrences**
Reg-RNR	210	104	66	380
PRN-RNR	0	9	85	94
Nb.occurrences	210	113	151	474

- Annotation and analysis problems with PRN-RNR : Among the 94 PRN-RNR occurrences, only 7 seem to come close to RNR phenomena.

(7) ... because they used, especially Disney used, a lot of classical music... (swbd-124010)

Reanalysis of PRN-RNR

- 87 among 94 occurrences of PRN-RNR should be treated in terms of *syntactic amalgams* (Lakoff 1974), involving *weak verbs* (Blanche-Benveniste and Willems 2007) : frequent use at the first person ; epistemic, evaluative or evidential meaning.

(8) My ex-husband's grandmother had been in a nursing home now for, oh, it must be seven, eight years. (swbd-50999)

Pattern	Nb.occurr.	Samples
'say'	27	*let's say, I would say, I want to say*, etc.
'I think...'	18	*I think {it is/it was/they call them}*, etc.
'It is...'	15	*it's, is it, it must be, it was like*, etc.
'I guess...'	13	*I guess {it's/it was/they're called}*, etc.
'call'	7	*what I call, they call it*, etc.
'I believe...'	4	*I believe it was*, etc.
'I don't know...'	2	
'I sensed'	1	

⇒ Our study so far takes into account only the Reg-RNR.

Plan

Challenging common beliefs on RNR

Our investigations both reassess the theoretical literature on RNR and the results of the previous corpus studies, and corroborate some undeservedly neglected criticisms (Abbott 1976, Chaves et Sag 2007, Chaves 2008, Abeillé et Mouret 2011).

- *Method :* Spreadsheet (Excel file) where each occurrence is classified on the basis of criteria below :
  - connector type
  - syntactic domain in which ellipsis operates
  - parent-category of the factorized chunk
  - category and syntactic function of the factorized chunk
  - semantic contrast between elements preceding the factorized chunk

RNR and coordination

- *Idée reçue :* RNR is often defined as an ellipsis construction peculiar to coordination (like Gapping).
- *Facts :* RNR occurs most frequently in coordinate constructions (in particular with the conjunctions *and* and *or*), but it is not ruled out in comparatives or other subordinate constructions.

Type	Total	Connectors
Coord	331	*and*(213), *or*(104), *but*(10), *as well as*(3), *nor*(1)
Paratax	23	comma, dash or nothing
Compar	8	*rather than*(5), *much less*(2), *than*(1)
Subord	6	*if not*(4), *though*(2)
Other	12	*from...to*(8), *to...from*(1), *instead of*(1), *versus*(1), *out of*(1)

⇒ RNR may occur in any syntactic context, and not only in coordinate constructions as usually assumed.

(9) Nearly half of them argue that Hong Kong's uneasy relationship with China will constrain – though not inhibit – long-term economic growth. (wsj-30583)

RNR and clausality

- *Idée reçue :* RNR is often defined as a relation between two (or more) clauses, the syntactic pattern mostly discussed in the literature being :

  [Subject + Vb + COORD + Subject + Vb + shared Complement]

  (10) John made and Mary sold a piece of furniture.

  ⇒ On this basis, previous corpus studies consider RNR as a rare, or even marginal phenomenon.

- *Facts :* RNR is not restricted to the clausal domain, it may occur in other phrases too (VP, NP, PP, ADJP).

RNR frequency across syntactic domains

- Overrepresented categories : VP and NP.

	WSJ	Brwn	Swbd	Total
S-type	13	13	15	41
VP	97	48	14	159
NP-type	71	25	28	124
PP	20	10	5	35
ADJP	6	5	0	11
UCP	3	3	4	10

(11) Motorola either denied or would not comment on the various charges. (wsj-28924)

(12) a. ...this was a formal or a informal dinner party. (swbd-132959)
b. Stephen N. Wertheimer was named managing director and group head of investment banking in Asia... (wsj-5968)

- 'Clausal' RNR is underrepresented.

⇒ Any corpus study which takes into account only 'clausal' RNR passes over most instances of RNR.

RNR and the syntax of the shared element

- *Idée reçue :* In the classical definition of RNR, the factorized chunk is an NP/PP dependent (in general, a complement of some verb).
- *Facts :* There is no restriction on the syntactic category, nor on the grammatical function of the factorized chunk.

Category	Total
NP	164
PP	82
NX	57
SBAR	31
S	22
VP	20
ADJP	4

Function	Total
Complement	258
Head	60
Adjunct	50
Comp/adj	12

(13) Tonight a group of men, tomorrow night he himself, would go out there somewhere and wait. (brwn-12426)

RNR and constituency I

- *Idée reçue :* The factorized chunk forms a constituent (unique and complete).
  - ⇒ RNR – used as a constituency test in generative grammars (a.o. Bresnan 1974, Postal 1974, Hankamer 1971)
  - ⇒ marginal reassessments (Abbott 1976), always on the basis of constructed data.

(14) a. *John offered, and Harry gave, Sally a Cadillac. (Hankamer 1971)

b. I borrowed, and my sisters stole, large sums of money from the Chase Manhattan Bank. (Abbott 1976)

RNR and constituency II

- *Facts :* The factorized chunk is not necessarily a constituent.
  - Sometimes, the factorized chunk contains more than one immediate constituent (multiple RNR), cf. (15-a).
  - In some cases, the factorized chunk is syntactically unsaturated, cf. (15-b) and (15-c).

(15) a. Combustion Engineering Inc., Stamford, Conn., said it sold and agreed to sell several investments and nonstrategic businesses for about $ 100 million *U*... (wsj-48463)

b. Now was this a one parent or two parent family ? (swbd-88331)

c. I have a five and a half and a three and a half year old that play with them. (swbd-26003)

RNR and syntactic parallelism

- *Idée reçue :* The conjuncts in RNR constructions must exhibit an identical syntactic structure (Hartmann 2000).
- *Facts :* Syntactic asymmetries are possible.
  - Different levels of complexity (± embedding), cf. (16-a).
  - Different selectional requirements (PP/VP – 13 occ. ; VP/PP – 23 occ.), cf. (16-b).
  - Voice mismatches (16-c).

(16) a. ...David O. Maxwell, who visits Tokyo at least once a year to explain and drum up investor interest in mortgage securities. (wsj-44398)

b. Mr. Baker interviewed or wrote to hundreds of catfish farmers... (wsj-515)

c. Whereas persons of eighth grade education or less were more apt to avoid or be shocked by nudity, those educated beyond the eighth grade increasily welcomed and approved nudity in sexual relations. (brwn-23827)

RNR and the Right Periphery Condition

- *Idée reçue :* RNR affects the entire right edge of the conjuncts (Féry and Hartmann 2005).
- *Facts :* The shared constituent may be followed by a constituent which is appropriate for the second conjunct, but not for the first ('wrapping').

(17)
a. But those dollars at risk pale in comparison to the investment required to make and ship spring goods to Campeau stores. (wsj-7615)
b. Pencil pushers chew and put the plastic models behind their ears... (wsj-47559)
c. ...its job is usually to help the lawyers identify and remove such people from the jury. (wsj-42428)
d. We discovered how much work it is just to organize and get it together. (swbd-9069)

RNR and semantic contrast

- *Idée reçue :* It is generally assumed (cf. Hartmann 2000 a.o.) that in RNR constructions, there must be a semantic contrast between strings across which RNR applies (like in Gapping constructions) ⇒ contrastive pairs, based on :
  - similarity : elements of a contrastive pair must belong to the same domain (or alternative set), and at the same time
  - dissimilarity : there must be a semantic opposition between them.
- In order to test this hypothesis, we have carried out a semantic encoding on the subset containing a verbal node (S or VP level) : 200 occurrences.

⇒ 3 main classes :

1. Contrast (66 occurrences)
2. Scalarity (79 occurrences)
3. Scenario (51 occurrences)

- *Facts :* The semantic contrast is lessened from one class to the other, the third one doesn't imply any contrast.

Class 1 : Contrast

66 occurrences : true contrasts alongside dimensions such as lexical or contextual antonymy (18-a), polarity (18-b), tense and aspect (18-c) or modality (18-d).

(18)
a. She learns how to relax them to accept – instead of contracting them to repel – the entering object. (brwn-23785)
b. Did you or did you not say what I said you said... ? (brwn-4498)
c. But the South is, and has been for the past century, engaged in a wide-sweeping urbanization... (brwn-16897)
d. Who is and who should be making the criminal law here ? (wsj-6370)

Class 2 : Scalarity

79 occurrences : one of the 'contrastive' elements (in general, the last one) is higher on a lexical or contextual scale.

⇒ Elements contrast with respect to their position on the scale (they share the same scale on different degrees).

(19) a. ...a substantial number of Rockefeller's faculty were upset over or even opposed to Dr. Baltimore's impending appointment. (wsj-37959)

b. The police said, all the people said, that's fine. (swbd-104656)

Class 3 : Scenario

- **51 occurrences** : a heterogeneous set of non-contrastive relations, such as sequentiality, causality, or more general appropriate actions.
  - ⇒ Occurrences of this kind are to be labeled as *scenarios* or *frames*, i.e. clusters of typical actions directed to the referent of the shared constituent (elements that are naturally related in the world).

(20) a. You first peel and then cook the potatoes.
b. The teacher carefully reads and harshly grades exams.

(21) a. There is not space to hold or force to guard any increased number of prisoners. (brwn-142)
b. ...the Energy Department tests and conducts research on nuclear weapons. (wsj-40057)
c. The waves of a 1923 tsunami in Sagami Bay brought to the surface and battered to death huge numbers of fishes that normally live in a depth of 3,000 feet. (brwn-22363)

Challenge for the contrast requirement assumed to operate with RNR : this class obviously contains elements which are not contrastive.

RNR and the contrast position

- *Idée reçue :* A RNR construction always contains a contrastive focus on the element immediately preceding the factorized chunk (Hartmann 2000).

(22) *Bill likes, and Mary likes, the TV show. (Ha 2008)

- *Facts :* The elements immediately preceding the targets in the conjuncts may not contrast with each other.

(23)
a. ...she knew and we knew that it was cowardice that had made one more radish that night just too impossible a strain. (brwn-5115)
b. The Mexican government is trying and a lot of the larger Mexican businesses are trying to, oh, make themselves Americanized, I guess,... (swbd-19338)
c. ...people get very frightened when they see the Japanese moving in and the Russians moving in certain areas of technology, you know, that we use to dominate. (swbd-121900)
d. The Fed is responsive to, and can not help being responsive to, the more overtly political part of the government. (wsj-48903)

General conclusion on RNR

- RNR is a much more frequent phenomenon and is much less constrained than what is usually assumed.
  - Our investigations invalidate several hypotheses put forward in the literature on ellipsis.

  $\Rightarrow$ Corpus data should be taken into account more seriously.

Thank you !

Selective references

Bies, A., **Ferguson**, M., **Katz**, K. & R. **MacIntyre** 1995. *Bracketing Guidelines for Treebank II Style Penn Treebank Project*. University of Pennsylvania.

Blanche-Benveniste, C. & D. **Willems** 2007. Un nouveau regard sur les verbes faibles. *Bulletin de la Société Linguistique de Paris 102(1)*. 217-254.

Chaves, R. 2008. Linearization-based Word-part ellipsis. *Linguistics and Philosophy 31(3)*. 261-307.

Greenbaum, S. & G. **Nelson** 1999. Elliptical clauses in spoken and written English. In P. Collins & D. Lee (eds.), *The clause in English*. Amsterdam : John Benjamins.

Harbusch, K. 2011. Incremental sentence production and clausal coordinate ellipsis : A treebank study comparing spoken and written language in Dutch and German. *Dialogue and Discourse 5*. 313-332.

Hartmann, K. 2000. *Right Node Raising and Gapping*. Amsterdam : John Benjamins.

Lakoff, G. 1974. Syntactic Amalgams. In M. Galy, R. Fox & A. Bruck (eds.), *Papers from the 10th meeting of the CLS*. 321-344.

Meyer, Ch. 1995. Coordination Ellipsis in Spoken and Written American English. *Language Sciences 17(3)*. 241-269.

Mouret, F. & A. **Abeillé** 2011. On the Rule of Right Node Raising in French. Talk given at the international workshop *Topics in the Typology of elliptical constructions - part I*. Paris 7 University.