Corpora and corpus linguistics in France: inventory and applications

Noalig Tanguy
Laboratoire Lattice UMR 8094, Paris, France
Introduction: corpus linguistics in France

Over the last 20 years:

• Corpus linguistics has developed
• So called ‘reference corpora’ have been created

Two contrasting conceptions:

• A unique grammar for each language
• Studies carried out using corpora and based on usage
Outline

1. Inventory of French corpora: written and spoken ones

2. Four applications carried out using corpora:
   - Project ANR Rhapsodie
   - Project ANR Orfeo
   - Projet « Left periphery of Basic Discourse Units »
   - Project PEPS MC4
1. Current situation

The evolution of corpus linguistics in France links it to language sciences and automatic language processing.

French linguists are interested in:
- Corpus linguistics
- Setting up corpora
- Circulating corpora
1. Current situation

- Financial means
- Sharing ressources
- Conferences
- PhDs
- Publications
1. Current situation

Example: *Corpus*

Available at: [http://corpus.revues.org/](http://corpus.revues.org/)

Objectives:

- Provide food for thought on the role of corpora in contemporary linguistic practises
- Provide food for thought on the way the various corpora are constituted
- Provide food for thought on the tools used to make use of corpora
1. Current situation

Review: the resources are not free and not shared

Regarding written French, databases are available for consultation.

**Frantext database**

literary corpora
texts in French, from the 16th to the 21st century, and medieval texts.
Possible requests: word search, lemma search, phrase search. The whole of the digital versions of the copyright-free texts can be found on the CNRTL website.
1. Current situation

**CNRTL** stands for « Centre National de Ressources Textuelles et Lexicales », i.e. National Centre for Textual and Lexical Ressources.

http://www.cnrtl.fr/

Computerized corpora:
- Frantext corpora
- DEDE corpora « DEscriptions Definies », i.e. « Definite Descriptions »
- Journalistic corpus taken from the French daily *L'Est Républicain*
1. Current situation

Spoken French corpora: scattered ressources and heterogenous data

Debaixieux, 2005

« On peut poser qu’il y a sans doute entre quatre ou cinq millions de mots effectivement disponibles mais l’absence de coordination rend l’exploitation de l’ensemble impossible »

= « There must be between four and five million words available but the absence of teamwork / organisation makes it impossible to use them »

Consequence: difficulties in sharing and making an inventory of ressources

Still, in 2005:

Lidil, 31 (2005): « Corpus oraux et diversité des approches »,
Cappeau and Seijido: « Inventaire des corpus oraux en France »
1. Current situation

In Europe (Germany, England, Spain, Portugal, Italy, etc...), constitution of reference written and spoken corpora:

- The British National Corpus (BNC) : 100 million words with morpho-syntactical annotations

- The reference corpus of current Spanish = le Corpus de Référence de l’Espagnol Actuel (CREA) : 100 million words, soon to become twenty-five million

- The German reference corpus= le corpus de référence allemand COSMAS II (Corpus Search, Management and Analysis system : Institut Für Deutsche Sprache à Mannheim)

- The contemporary Portuguese reference Corpus = Corpus de Référence du Portugais Contemporain (CRPC) : 80 million words.
1. Current situation

France has become aware of being late as regards:

(i) the development of resources
(ii) the constitution of a computerized textual data base, especially for spoken French

Since the 2000s: deep reflection on the creation and the hosting of corpora
1. Current situation

Actions have been set up:

- Creation of a « Guide to good pratice » : deontological and legal aspects of the collecting and analysis of spoken data.


- Setting up of the digital national research : the « TGE Adonis », available at  :
  http://www.cnrs.fr/inshs/recherche/TGE-ADONIS.php
1. Current situation

TGE Adonis

Main aim: provide access to and preserve digital data produced by humanities and social science.

Objectives: sharing resources and technologies; preserve data.

Creation of a programme: « Corpus de la parole » = Speech corpus

Available at: http://corpusdelaparole.tge-adonis.fr/
1. Current situation

Creation of the CRDO

Centre de Ressources pour la Description de l’Oral, i.e. ressource centre for the description of spoken language.

Mission : preserve and share spoken corpora.

Has become project COCOON « COollections de COrpus Oraux Numériques » (i.e. collections of digital spoken corpora)
1. Current situation

Creation of computerized spoken database.

- ESLO « Enquête sociolinguistique à Orléans » / Université d’Orléans
  Available at: http://eslo.tge-adonis.fr/
- PFC « Phonologie du Français Contemporain » / Université Paris 10 Nanterre
  Available at: http://www.projet-pfc.net/
- CFPP2000 « Corpus de Français Parlé Parisien des années 2000 » / Université Paris 3 Sorbonne Nouvelle
  Available at: http://cfpp2000.univ-paris3.fr/
- CLAPI « Corpus de LAngue Parlée en Interaction » / Lyon
  Available at: http://clapi.univ-lyon2.fr/
1. Current situation

**CFPP2000 (1/2)**

Corpus de Français Parlé Parisien des années 2000
Corpus of spoken Parisian French in the 2000s

Corpus: non-directive interviews about Paris districts and inner suburbs

Specificities: interviews transcribed in spelling and aligned following speech turns. Quite long interviews, around an hour each

February 2013: 550 000 words

Online tools: concordancer and textometrical tools
1. Current situation

Aim: analyze spoken French and study its variations

Diversified linguistic activities: descriptions of some districts, anecdotes, argumentations, etc.

Dialogues vs multilogues

Université Sorbonne nouvelle Paris 3 / direction: Sonia Branca-Rosoff
1. Current situation

Corpus « SMS pour la Science ».  
Université Catholique de Louvain, Belgique / Belgium  
Available at : http://www.smspourlascience.be/

Situation : SMS are an undeniable commercial success. Their development brought about a specific writing style, which entailed much criticism and a great many reactions of linguists, sociologists, psychologists and other communication specialists.

Objective: Carry out broad scale linguistic studies and enable new development in Natural Language Processing.

Size : 30 000 SMS in French

Corpus available on CD-ROM
1. Current situation

IR-Corpus

2011 / CNRS
TGIR : Très Grande Infrastructure de Recherche (= Really Large Research Infrastructure)
Aim : develop search operators cooperation for a use of digital sources in humanities and social science.
June 2013 : fusion of TGE Adonis and of IR-Corpus and création of several consortia

http://corpusecrits.corpus-ir.fr/
http://ircom.corpus-ir.fr/site/accueil.php

Actions :
- Take an inventory of existing corpora and their completion status.
- Bring technical and possibly financial support in order to complete, bring up to standards and disseminate the existing corpora.
- make sure users and trained and informed.
- Foster the integration of French users into international networks.
- Support the creation of work groups.
1. Current situation

Inventories

1. Inventory of the IRCOM consortium corpus oraux
   http://ircom.corpus-ir.fr/site/corpus.php

2. Cappeau et Seijido (DGLFLF), 2005
   http://www.dglf.culture.gouv.fr/recherche/corpus_parole/Presentation_Inventaire.pdf
2. Four applications for corpora
2.1. The ANR Rhapsodie project

ANR
Agence Nationale de la Recherche (= National Research Agency)

2008-2012
Université Paris 10 Nanterre / Direction : Anne Lacheret

Aim: set up a corpus of spoken French, annotated for prosody and syntax.

Corpus now available at: http://projet-rhapsodie.fr/
2.1. The ANR Rhapsodie project

Initial question: Which role do the intonosyntactical elements play in the segmentation of speech into informational and discursive units?

Underlying questions: Which is the congruency degree between syntactical units and prosodic units? In which way do prosody and syntax take part in the segmentation of speech into pragmatic and textual units?

To answer these questions, it is necessary to develop a doubly annotated corpus (prosody and syntax)

Theoretical aim: model the prosody / syntax / spoken French speech interface.
2.1. The ANR Rhapsodie project

Two steps:

1. Develop an annotation pattern in syntax (micro and macro syntax) and in prosody (prominences, disfluencies, melodic contours, intonative periods)

2. Set up a database
   Which questions?
   Which possible requests?
   Aim: an open-access corpus with tools
2.1. The ANR Rhapsodie Project

Corpus /treebank Rhapsodie

57 sound samples of spoken French
Transcription in spelling and phonetical transcription aligned following the sound
syntactical and prosodic annotation

External Sources : 32 samples
    Projets CFPP2000, C-Prom, ESLO, et PFC + PhDs
Collected sources: 25 samples

an average of 5 minutes per sample
A total of 3 hours : 33,000 words

Multi-genre corpora : monologues / dialogues, personal discussions / public speech, face to face interviews, talk-shows, descriptions, argumentative speech, prodecural sequences, etc.
2.1. The ANR Rhapsodie Project

Prosodic annotation

Syllabic annotation
Annotation according to the type of prominence: non prominent / prominent / strongly prominent
Tonal characteristics and presence or not of disfluency

Annotation of silent pauses and non silent pauses

A four-level prosodic structure:
  - Periods
  - Intonation groups
  - Rhythm groups
  - Metrical foot
2.1. The ANR Rhapsodie Project

Syntactic annotation

Three levels of analysis: morphosyntactical / microsyntactical / macrosyntactical.

Morphosyntactical annotation: breaks between words (lemmas and parts of speech)

Microsyntactical annotation:
- dependency between words: subject/object/addition/etc.
- piles.

Example
{ je | j’ } arrosais { les plantes | ^enfin les fleurs } { le matin | ^et le soir }
2.1. The ANR Rhapsodie Project

Macrosyntactic annotation
   Speech separated into illocutionary units
   Kernel = illocutionary force
   Pre-kernel / post-kernel / parenthesis /discourse markers

Example:
^et { le soir |} <+ par contre <+ { | "ça" le soir } <+ je reviens jamais en métro >+ sauf quand il y a des manifestations //=

Tools availbale on the project website:
http://projet-rhapsodie.fr/
2.2 « Left periphery of Basic Discourse Units » Project

FNRS Project / Belgium
National Scientific Research funds
2011-2014
Centre de recherche Valibel, Catholic University of Leuven, Louvain-la-Neuve
Direction : Liesbeth Degand & Anne Catherine Simon

Research subject : Basic Discourse Units (BDU)
2.2 « Left periphery of Basic Discourse Units » Project

Questions

Can a grammatical annotation be independent from a prosodic annotation?
Which « basic units » can be used to address textual or discursive phenomena?

Left periphery

No syntactic definition
Not only dislocated or segmented constructions
Taking into account the syntactic and prosodic configuration of the elements at the initial position of these discourse units (‘left periphery’)

2.2 « Left periphery of Basic Discourse Units » Project

Corpus

Total of 3h20
44 recorded samples: transcribed, aligned following speech, annotated syntactically and prosodically

45 000 words from different sources:
- Scientific conferences
- Academic discourse
- Religious discourse
- Unguided conversations
- News reports
- Reading
- Political discourse and debate
2.2 « Left periphery of Basic Discourse Units » Project

Syntactic annotation + prosodic annotation
= identification of Basic Discourse Units

Prosodic annotation
A semi-automatic prosodic annotation procedure
Step 1: automatic detection
Step 2: manual validation

Two types of annotation
prosodic boundary
intonation contour
+ hesitations
+ Initial stress
Intermediate and major prosodic boundary

\[ T// = \text{Intermediate prosodic boundary } (///) \text{ (Top)} \]
\[ P/// = \text{Major prosodic boundary } (////) \text{ (Pause)} \]

Intonation contours
Continuation / Finality
2.2 « Left periphery of Basic Discourse Units » Project

A major prosodic boundary (///) is established when one of the following cues is detected on the final syllable of a word (or penultimate when the final syllable is a schwa):
- a subsequent silent pause longer than 250 ms;
- an extra-lengthening (the syllable is three times longer than the syllables in the context);
- a sharp rise of f0 (intra-syllabic f0 rise superior to ten semi-tones), even when the f0 rise does not correlate with the lengthening of the syllable.

We manually exclude a boundary which coincides with a hesitation mark (‘euh’ particle or vowel extra-lengthening with a level contour, and creaky voice), since it has been demonstrated that hesitations are not confounded with prosodic breaks in discourse processing.
An intermediate prosodic boundary (//) arises when the final syllable of a word is lengthened (the syllable is twice longer than the syllables in the immediate surrounding context), bears a sharp rise of f0 (superior to four semi-tones), or is higher than adjacent syllables (higher than five semi-tones).
The second step of the prosodic annotation consists in attributing an intonation contour to each prosodic boundary.

Four alternatives exist:
- Continuation
- Finality
- Focus
- Suspense
Governing Unit
Verbal Governing Unit

Functional sequences Verb
Left Governed Sequence + Verbal Sequence + Séquence Objet (Object)
The starting point of the analysis is thus a verbal micro-syntax in which the verb (or any other governor) and its governed complements are central.

This micro-syntactic analysis results in segmentation into four types of dependency clause:

(i) verbal dependency clauses
(ii) averbal dependency clauses
(iii) elliptical dependency clauses
(iv) interrupted dependency clauses

The syntactic analysis leaves us with a number of un gover ned segments, which belong to the macro-syntax rather than to the micro-syntax. They comprise so-called ‘associés’ (‘adjuncts’) and discourse markers, which are not governed by the main clause, but are semantically or pragmatically linked to the whole dependency clause.
2.2 « Left periphery of Basic Discourse Units » Project

Functional sequences
The second and final step in the syntactic annotation process consists in cutting up each dependency clause into so-called “functional sequences”, i.e. clausal constituents that occupy a main syntactic function like Verb, Subject, Object, etc. (Bilger & Campione 2002: 119).

<table>
<thead>
<tr>
<th>SS</th>
<th>séquence sujet</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV</td>
<td>séquence verbe</td>
</tr>
<tr>
<td>SO</td>
<td>séquence objet</td>
</tr>
<tr>
<td>SR</td>
<td>séquence régie</td>
</tr>
<tr>
<td>a</td>
<td>adjoin</td>
</tr>
<tr>
<td>insert</td>
<td>unité de résection qui interrompt une unité en cours</td>
</tr>
</tbody>
</table>
The **congruent BDU**, or BDU-c (one-to-one mapping): a basic discourse unit with congruent mapping between syntax and prosody, in the sense that the syntactic unit (dependency clause) realizes one major prosodic unit.

The **“syntax-bound” BDU**, or BDU-s (one-to-many mapping): the basic discourse unit is uttered in such a way that the speaker pronounces one syntactic unit into successive prosodic units.

The **“intonation-bound” BDU**, or BDU-i (many-to-one mapping): the basic discourse unit is uttered in such a way that the speaker groups two or more syntactic units into one major prosodic unit.

The **“regulatory” BDU**, or BDU-r: This BDU results from the mapping between a major intonation unit and an (isolated) adjunct or discourse marker.
2.2 « Left periphery of Basic Discourse Units » Project

Examples

**BDU-I**  
[(on voit bien) (que nous sommes massivement mobilisés)] <donc>  
[(le mouvement) (ne s'essoufflera pas)] <mais> /// (INT)  
<alors> <(essoufflement)> [(non)] /// (INT)

**BDU-S**  
[(ce parlement /// et la commission européenne) /// (font avancer) /// (en Europe) /// (là même où ont été inventées les plus vieilles démocraties) /// (une machine à broyer la souveraineté populaire /// telle que nous pouvons parler à propos du mécanisme européen dorénavant d'un mécanisme autoritaire)] /// (DIS)

**BDU-R**  
<car> /// ... (DIS)
Left periphery

The left periphery is determined by syntactic (independent from the dependency clause) and prosodic (presence of a major boundary) criteria.

Hypothesis

We suppose that Initial elements and left periphery of BDUs play different parts.
### 2.2 « Left periphery of Basic Discourse Units » Project

<table>
<thead>
<tr>
<th>Name</th>
<th>Syntactically</th>
<th>Prosodically</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Element (IE)</td>
<td>Dependent</td>
<td>Integrated</td>
</tr>
<tr>
<td>Syntactico-Prosodic Left</td>
<td>Independent</td>
<td>Isolated</td>
</tr>
<tr>
<td>Peripheral Element (LPE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syntactic LPE</td>
<td>Independent</td>
<td>Integrated</td>
</tr>
<tr>
<td>Prosodic LPE</td>
<td>Dependent</td>
<td>Isolated</td>
</tr>
</tbody>
</table>

Table I. Summary of initial sequences of BDUs
The categorisation of the discursive function of the LPE is inspired by that of Chafe (1999). It is threefold distinguishing (i) textual, (ii) interactive, and (iii) cognitive functions.

1. **Textual function**: marking a logical relationship, mainly by means of linking devices, or managing topics (introduction, continuation, specification, etc.)

2. **Interactive or interactional function**: vocatives and metadiscursive statements (the way the speakers keep their turn, mark the reception/understanding of the information, etc.)

3. **Cognitive or validating function**: judgements on the validity of the transferred information (markers of subjectivity, enunciative comments, modals: *I’m sure that, of course, it is said that, apparently, ...*).
Hypothesis

We hypothesize that all types of LPEs are of significant importance when it comes to coherence. We suppose that LPEs and IEs play different parts: the former being cohesive expressions, they serve discursive coherence, i.e. they are the result of strategic choices. We furthermore state that their forms and functions vary with discursive genres.

First results

Two different situations: the interview and the political speech

The political speech contains more prosodic LPEs (caused by prosodic over-segmentation and an emphatic and didactic style) and more prosodico-syntactic LPEs, while the interview contains a majority of syntactic LPEs. This difference can be explained by the general tendency in political speech to favour strong prosodic marking.
2.3 The PEPS MC4 project

PEPS = Projet Exploratoire Premier Soutien
MC4 = « Modélisation Contrastive et Computationnelle des Chaînes de Coréférence »
Laboratoire LATTICE, Paris / Frédéric Landragin
2011-2013
2009-2011 : Groupe de travail « COREF »

Research subject: reference and coreference in written texts, in mediaeval and contemporary French.

Team: linguistics and IT researchers, reference, coreference, grammaticalisation and corpora analysis specialists.
2.3. The PEPS MC4 project

Linguistic component

Objective
   Modelize the elements in a coreference sequence.

Stages
   Studying referential phrases: proper names, noun phrases, personal pronouns, demonstratives, etc....
   Studying phrases and clues which do not corefer but indicate the existence of a referent: appositions, pronominal constructions, etc.

Three research subjects
   References
   Succession of references in a text
   Coreferentiality in a text
2.3. The PEPS MC4 project

Studying references

Objective: establish a typology of phrases and referential elements.

Steps
- Elaborating a common annotating pattern, including morphosyntactical, syntactical and semantic aspects
- Writing an annotation handbook
- Annotation of several short texts, to test and check the pattern and the handbook
2.3. The PEPS MC4 project

Description of referential and coreferential phrases following several criteria:

- Syntactic category
- Presence of noun expansion: adjectives, noun complementizer, relative clauses, etc.
- Syntactic function
- Syntactic level: main or subordinate
- Actancial roles: agent / patient / other
- Enonciative level: dialogue or reported speech / parenthesis
- Position within the sentence: initial / median / final
2.3. The PEPS MC4 project

Succession of references in a text

To which frequency does one go from one referent to another in a text?

For how long can a referent be alluded to and how?
2.3. The PEPS MC4 project

Coreference chains in a text
- Is there a typology of these reference chains concerning one and only referent?
- How does that contribute to discourse?
- Are there any correlations between the occurrence of a category of expression and informational and textual characteristics?
- How are coreference chains introduced lithine discourse? The project identified several possibilities: through discourse construction, through associative anaphora, through the extraction of a referent, of a group of referents, etc.
- How do we move from one chain to another?
- Is there a link between a certain type of coreference chains and the various text genres?
2.3. The PEPS MC4 project

Tools

Need of IT
Analec software: free access on the lab website.

Reference and access

Victorri B. (2010). « Analec 0.6 : logiciel d’annotation et d’analyse de corpus écrits », downloadable software at:
http://www.lattice.cnrs.fr/-Analec-. 
2.3. The PEPS MC4 project

Diachronic component

Annotation pattern: tested on contemporary texts, texts in Old French and in Middle French

Modifications to handle all the language facts

Example: the pro-drop pattern (unexpressed pronoun)
2.3. The PEPS MC4 project

**Fullfilled objectives**
- Constitution of a short corpus
- Annotation of the corpus
- Study of the reference chains
- Comparison between text genres and linguistic phenomena
- Quick availability of the annotated corpus

**Forthcoming, writing of a special number of a periodical**
- Typology of reference chains in a corpus of medieval narrative texts
- Reference chains in medieval Anglo Norman texts
- Reference chains in non narrative texts (public reports and legal texts)
- Coreference in spoken French : links with syntactic and prosodic segmentation
- Reference and coreference of the indefinite pronoun "on"
- Computer assisted analysis of coreference chains
2.4 The ANR ORFEO Project

ORFEO = Outils et Recherches sur le Français Ecrit et Oral
Tools and Researches on Written and Spoken French

Laboratoires : Lattice, Modyco, Atif, Lif, Clle-Erss, Loria, Icar
2013-2016
No website so far

Pursuance of the Rhapsodie Project
Call for projects ANR « Corpus »

Objective : constitution of a Corpus for the Study of Contemporary French, i.e. Corpus d'Etude pour le Français Contemporain (CEFC)
2.4 The ANR ORFEO Project

6 actions

1. Gather open access corpora in accord with the initial developer.

2. Collect data for the uncollected types, so as to gather a corpus of various types, a total of 15 million words.
   - 1/5 oral
   - 4/5 written (fiction / press / academic discourse / unplanned written texts)
   = Multi type and multi-usage corpora.

3. Set up a platform to access data and meta-data with search and analysis tools.
2.4 The ANR ORFEO Project

4. Secure the perennial conservation of documents, filing annotated corpora in digital resource centres: CNRTL, SLDR, or the future Equipex (devised by their laboratories and the Paris Ouest University and Orléans University).

5. Automatically annotate the whole of the corpus: syntax / prosody / speech.
   - prosodic and syntactic annotations: Rhapsodie Project
   - Discursive Annotation: Annodis Project

6. Develop pilot studies using all the tools of corpus linguistics in various fields
   A few linguistic phenomena:
   - lists
   - attitudes markers
   - clause combining
Corpora and corpus linguistics in France: inventory and applications

Noalig Tanguy
Laboratoire Lattice UMR 8094, Paris, France