The collection of data for the Rhapsodie treebank: Corpus design and ethical issues

Anne Lacheret-Dujour, Paola Pietrandrea, Olivier Baude, Nicolas Obin, Anne-Catherine Simon, Atanas Tchobanov

1. Introduction

While text typology is widespread in linguistic studies of written text, this field is currently still in its infancy in speech studies. When present, it is mainly conducted from a comparative approach of written and spoken texts in the English language (Halliday 1989, Biber et al. 2009), focusing on the segmental level (phonological, syntactic or lexical). Rhapsodie follows on from work that aims to extend the scope of textual typology to spoken discourse focusing on intonosyntactic interface, i.e. analyzing correlations between types of text and variations in prosodic and syntactic constructions. Such a program needs a set of data that is as diverse as possible in order to process a sufficient variety of text types and discourse genres.

This chapter addresses the different issues related to achieving this diversity in the Rhapsodie project. After a presentation of the hypotheses that piloted the corpus design and a reminder of the context of speech corpora studies (section 2), we first present the issues relating to the sampling (size, type and number of samples) and the strategy used to achieve the corpus design in order to ensure that a sufficient variety of text types and linguistic structures be represented (section 3). Lastly, we discuss the legal and ethical implications of the distribution of the Rhapsodie resources, as well as the metadata standard used to provide an exhaustive textual characterization of each sample, to provide complete information about source corpora and to precisely describe the annotations of each sample, which are at the heart of the Rhapsodie project (section 4).

2. Purpose, method and context

In this section, first, we discuss the linguistic theoretical and empirical assumptions concerning correlations between situational context, discourse genres, types of text and structural variations, that piloted the data collection in Rhapsodie (section 2.1). Then we present the specificity of this collection with respect to previous experiments in speech linguistics corpora (section 2.2).

2.1. Theoretical approach

We followed, in the collection of our data, Biber et al.'s (1999: 4) approach to corpus design which holds that: "The vocabulary and grammar that we use to communicate are influenced by a number of factors, such as the reason for the communication, the context, the people with whom we are communicating, and whether we are speaking or writing". In other words we assumed that a strict relation exists between textual typologies - i.e. textual patterns defined in intrinsically linguistic terms - and genres - i.e. socio-communicative patterns - in that a particular linguistic structure reflects a particular genre, and vice versa a given genre engenders a limited number of linguistic structures (Bakhtine 1984, Swales 1990, Maingueneau 1996, Adam 1999[1]). As for the specific objectives of our research, we assumed that a correlation exists between discourse types and the distribution of intonosyntactic structures, and hence that a corpus comprising a wide range of discourse types would also comprise a wide range of intonosyntactic structures.

2.2. Previous experiences

[1] See also the notion of 'register' in Biber and Conrad (2009: 16) defined by a bundle of lexico-grammatical features that are frequent and pervasive in texts from a given variety, and that serve important communicative functions.

The majority of existing treebanks are not diversified with regard to discourse genres. The Verbmobil treebanks of English, German and Japanese (Hinrichs et al. 2000), for example, only comprise samples of spontaneous conversations; the Spoken French Ester treebank (Cerisara et al. 2010) is exclusively based on transcripts of French radio news broadcasts; the Venice treebank (Delmonte et al. 2007) includes only (regionally diversified) Italian conversations and the Switchboard corpus (Meteer et al. 1995), only telephone conversations. The design of the CHRISTINE treebank (Sampson 2000) aimed at providing a more diversified and balanced sampling based on extracts from the "demographically-sampled" speech section of the British National Corpus. While this design choice guaranteed a sociolinguistic diversification of the corpus, it did not ensure a text type diversification.

However, particular attention to including structural variety in the design of the corpus was paid by the creators of C-ORAL Rom (Cresti & Moneglia, 2005) and by those of the British component of the International Corpus of English (Nelson et al. 2002), the Diachronic Corpus of Present-Day Spoken English (Aarts & Wallis 2006) and the CGN Spoken Dutch Corpus (Schuurman et al. 2004). The C-Oral Rom sampling is based on the following sets of variation parameters: (a) Dialogical structure (monologues, dialogues, conversations); (b) Social domain of use (family, private life, public life, media productions); (c) Gender variation; (d) Formal vs. informal distinction; (e) Speaker parameters (Age, Sex, Education, and Occupation). The British component of the International Corpus of English (Nelson et al. 2002), the Diachronic Corpus of Present-Day Spoken English (Aarts & Wallis 2006) and the CGN Spoken Dutch Corpus (Schuurman et al. 2004) are all based on the sampling principles established by the International Corpus of English that recommends including in the corpus equal proportions of monologues and dialogues, public and private speech, scripted and unscripted speech.

While our work was largely grounded on these previous experiences, we proposed some

refinements. We decided to replace the C-Oral Rom distinction between formal and informal speech and between different genres by a less subjective multifactorial analysis of the speech situation. We encountered a difficulty, however. While spoken treebanks were extracted from existing representative corpora, or at least built within the framework of large projects dealing with corpus construction, the creation of the Rhapsodie repository could not rely on any representative corpus of spoken French, since none exists; nor could we, in the scope of this project, count on enough time and money to collect a corpus from scratch. This peculiar situation led us to construct our textually diversified repository of samples partly by extracting samples from a number of relatively small and specific pre-existing spoken corpora of French.

3. Rhapsodie sampling

This section is devoted first to the presentation of the different situational variables selected for the construction and the description of the Rhapsodie repository, and second to the presentation of the different sources of the Rhapsodie collection.

3.1. Rhapsodie corpus design: general principles

It is well known that discourse genres can be described as multifactorial phenomena involving a number of socio-communicative variables that are independent of one another (Biber et al. 1999, Koch and Oesterreicher 2001). A given discourse genre, for example, can be described in terms of the nature of the speech situation (its location, its goals, the degree of formality), in terms of the physical constraints impacting on the speech situation (in particular the type of communication channel), or in terms of the topic, i.e. the semantic content of the exchange.

The design of the Rhapsodie corpus (table 1) was first of all based on the balance between:

(i) monologues (M), i.e., discourse produced by a single speaker addressed to interlocutors who could not freely take a speech turn (whether a large audience or a single addressee);

(ii) dialogues (D), i.e., produced by two or more speakers in a situation of either low or high interactivity.

By including both monologues and dialogues in our corpus, we were able to take into account in the development of our annotations a number of phenomena typically related to interaction such as overlaps, turn-taking or interactional discourse markers.

Secondly, we distinguished between public and private speech[2]:

(iii) Private speech (Rhap-M0 and M1 or Rhap-D0 and D1) is composed of samples extracted either from face-to-face interviews between the linguist researcher and one or more French speakers, or from everyday life interactions. Private speech may cover any topic.

(iv) In Public speech (Rhap-M2 or Rhap-D2), the speaker addresses an audience (conferences, radio or television broadcasts: political speeches or debates, talk shows, scientific press, reportage, forecasts, literary programs, etc.).

Finally, we also took into account:

(v) the degree of planning of the speech (spontaneous or planned);

(vi) the degree of interactivity (interactive, semi-interactive, non interactive);

(vii) the channel of communication (broadcast or face-to-face);

(viii) the type of discourse sequence mostly characterizing the speech (description, argumentation, procedures, etc. – see Adam 1999)[3]

(ix) the type of task (interviews, sermons, sportscasting, movie scene description, travel planning, etc.)[4i]

Table 1 summarizes the types of situational variables that were taken into account in the construction of the Rhapsodie corpus.

[2] See for example the distinction between “communication privée” and “communication publique” in Koch & Oesterreicher (2001: 586).

[3] It is well known that labeling different types of discourse sequences is complex and to a certain extent arbitrary, as a given production can be associated to different kinds of sequences. For example, a narration is often descriptive and a description is rarely neutral, and often contains argumentative features; when this is not the case, it is considered as a procedural sequence (map task for example).

[4] The vocabulary adopted to define these properties is that of the IMDI standard (section 4).

Event structure	Monologue, dialogue
Social context	Private, Public
Planning	Spontaneous, semi-spontaneous, planned
Interactivity	Interactive, semi-interactive, non-interactive
Channel	Broadcast; face-to-face
Discourse sequence	Argumentation, narration, description, procedural speech, oratory speech
Task	Info-kiosk, sermon, lesson, project description...

Table 1. Situational variables in Rhapsodie

In order to maximize the diversity of speakers, we decided to include in our corpus a large number of short speech samples uttered by 89 Central French adult native speakers (males and females) from the early eighties to nowadays, making a total of 57 short audio samples (5 minutes long on average)[5], amounting to 3 hours of speech and a 33 000-word corpus.

3. 2. Gathering data: external and internal sources

Gathering new data with the aim of building a diversified collection of textual samples would have been too costly and time-consuming for a project whose primary aim was to develop tools for the annotation and analysis of spoken French. As mentioned, we therefore preferred to build the bulk of our corpus (32 samples) by selecting data from 7 corpora of spoken French created in recent years for various scientific projects. Table 2 details the 7 external source corpora that were drawn on for the *Rhapsodie* reservoir: CFPP2000, C-Prom, ESLO, PFC, the Avanzi Corpus, the Lacheret Corpus, and the Mertens Corpus.

Source	Description	Number of samples	Samples
CFPP2000	The CFPP2000 (Le Corpus de Français Parlé Parisien) is made up of interviews about Paris districts and suburbs. It provides data to study Parisian French as it is used in real communication *(Branca-Rosoff et al. 2012)* http://cfpp2000.univ-paris3.fr/	4	D0001, D0002, D0004, D0006
Avanzi Corpus	The Avanzi Corpus was collected for the intonosyntactic study of macrosyntactic phenomena conducted by Mathieu Avanzi for his Ph.D. dissertation at the University of Neuchatel, 2011 *(Avanzi 2012)*	19	M0001, M0003-M0017, D0007,

[5] The samples were randomly extracted from original sources.

			D0008, D0020
Lacheret Corpus	The Lacheret Corpus was compiled by Anne Lacheret for research purposes, and focuses on the continuous and functional modeling of French prosody *(Lacheret 2003)*	2	D2004, D2005
Mertens Corpus	The Mertens Corpus was compiled by Piet Mertens for his doctoral dissertation. Mertens' dissertation was the first approach to the intonosyntactic modeling of French developed with the aim of constructing an automatic system of identification of intonational units *(Mertens 1987)*	2	D2001, D2009
C-Prom	C-PROM is an aligned and annotated corpus developed with the aim of studying syllable prominences in French. It comprises 24 recordings representing 7 different genres produced by French, Belgian and Swiss native speakers of French *(Avanzi et al. 2010)* http://sites.google.com/site/corpusprom/	1	M2001
ESLO	ESLO, L'Enquête Sociolinguistique à Orléans is a 300-hour, 4,500 000 word corpus of spoken French gathered in Orleans, France in 1969-71 with a sociolinguistic aim. It includes 157 interviews and more than 200 recordings of spontaneous private and professional conversations, telephone exchanges, public meetings, and service encounters http://eslo.tge-adonis.fr *(Eshkol-Taravella et al. 2011)*	1	D1001
PFC	The international project Phonologie du français contemporain, directed by Marie-Hélène Côté (University of Ottawa), Jacques Durand (ERSS, University de Toulouse-Le Mirail), Bernard Laks (MoDyCo, Université de Paris Ouest Nanterre la Défense) & Ch. Lyche (Oslo University) aims to obtain an accurate picture of both the similarity and diversity of phonetic varieties of contemporary spoken French. The elements of the PFC database which were used for Rhapsodie sampling are directed conversations between a subject and an interviewer and informal conversations between two persons belonging to a dense social network. http://www.projet-pfc.net/ *(Durand et al. (2009)*	3	D0003, D0005, D0009

Table 2. External source corpora used in Rhapsodie

It should be noted that most of the *Rhapsodie* samples are shorter (by between 1 to 10 minutes) than the original source files. Our aim of maximizing the variety of text types and speakers led us to select only a portion of the source file[6] in order to include in our reservoir a large number of short samples produced by many speakers. We realize that this choice, which allows for fine-grained and complete intonosyntactic analyzes of each sample, is not perfectly suitable for complex semantic textual analyzes. However, as will be shown in section 4, users can retrieve the whole original source files through an identifier that is given in the metadata

[6] We decided to collect samples uttered by a large number of speakers, not in order to achieve good socio-linguistic representativeness (which was not our objective), but to minimize the effects of individual idiosyncrasies.

associated to each sample.

In order to guarantee the representativeness of all the situational variables listed in section 3, we also collected 27 original samples. Table 3 details the 3 original subcorpora collected for the Rhapsodie reservoir.

Subcorpora	Description	Number of samples	Samples
Movie description Corpus	The Movie description corpus comprises 7 monologues in which 7 different speakers are invited, in an informal setting, to describe a short scene from a Charlie Chaplin movie	7	M0002, M0022, M019, M018, M021, M023
Professional Corpus	The Professional corpus consists of 4 speech samples (monologues and dialogues) in a professional context.	4	M1001, M1003, D1002, D1003
Broadcast corpus	The Broadcast corpus consists of 14 broadcasted monologues, dialogues and conversations downloaded from the Internet for the Rhapsodie project	14	M2004, D2003, M2002, M2006, M2003, D2007, D2008, D2010, D2006, M2005, D2012, D2011, D2013, D2002

Table 3. Internal sources created for the Rhapsodie project

A synopsis of the criteria adopted for collecting and sampling the Rhapsodie corpus is given in figure 1. Table 4 presents all the samples of the treebank according to their number of tokens and their length.

Monologues	Key word	Number of tokens	Duration	Dialogues	Key word	Number of tokens	Duration seconds
M0001	Itinerary: Place Victor-Hugo1	138	100	D0001	Paris by bus	1185	330
M0002	Charlot: silent film	190	67	D0002	Schools in Paris	1107	291
M0003	Itinerary: Notre Dame	353	100	D0003	Childhood memories	470	285
M0004	Itinerary: cinema star1	45	15	D0004	Bookshops in Paris	1281	293
M0005	Itinerary: Albert 1er de Belgique 1	143	48	D0005	Computer programmer	995	240
M0006	Itinerary: Albert 1er de Belgique 2	112	27	D0006	Medical student	1364	367

M0007	Itinerary: Saint Jean de Maurienne1	172	51	D0007	Itinerary: Hubert Dubedout	143	110
M0008	Itinerary: Hermillon	55	18	D0008	Itinerary: cable-car	692	103
M0009	Itinerary: Saint Jean de Maurienne2	205	64	D0009	Crapaud armchair	1343	156
M0010	Itinerary: cinema star2	61	17	D0017	Itinerary: crossing	114	47
M0011	Itinerary: fils de la Sarce	161	53	D0020	Itinerary: rue Lakanal	106	35
M0012	Itinerary: boulevard Gambetta	50	22	D1001	Christian living	1026	434
M0013	Itinerary: Place Victor-Hugo2	176	51	D1002	Art: Yves Klein	516	154
M0014	Itinerary: place Paul Vallier	79	30	D1003	Studies in Portugal	560	136
M0015	Itinerary: avenue Alsace Lorraine	79	27	D2001	Radio: Françoise Giroud	1997	206
M0016	Itinerary: CRDT	244	83	D2002	Radio: Pascal Ferran, Haruki Murakami	936	632
M0018	Charlot: social conflict	262	89	D2003	Match France-Argentina	1398	300
M0019	Charlot: caterer	194	60	D2004	Flying doctors	1023	314
M0022	Charlot: true gentleman	228	103	D2005	Radio: Jacques Attali	1014	306
M0023	Charlot: the girl	333	88	D2006	Mitterrand at the National Assembly	237	170
M0024	Charlot: the bakery	138	45	D2007	Talk show: how to pick somebody up	1055	262
M1001	Clara, 19 years old	386	88	D2008	Radio: influenza	2119	512
M1003	Angelina, 18 years old	828	247	D2009	Radio: Roland Barthes	1040	309
M2001	Sarkozy,	535	217	D2010	Radio:	1126	319

	speech to the army				Marguerite Duras		
M2002	Conference on philosophy	1312	386	D2011	Teleshopping: magic ball	1067	341
M2003	The mass	723	295	D2012	Radio: Szymanowski	1004	306
M2004	Chirac, New Year's Eve broadcast to the French	1263	632	D2013	Press Review	636	182
M2005	Radio: Facebook	355					
M2006	News flash	975					

Table 4. Rhapsodie samples: length and number of words of each sample

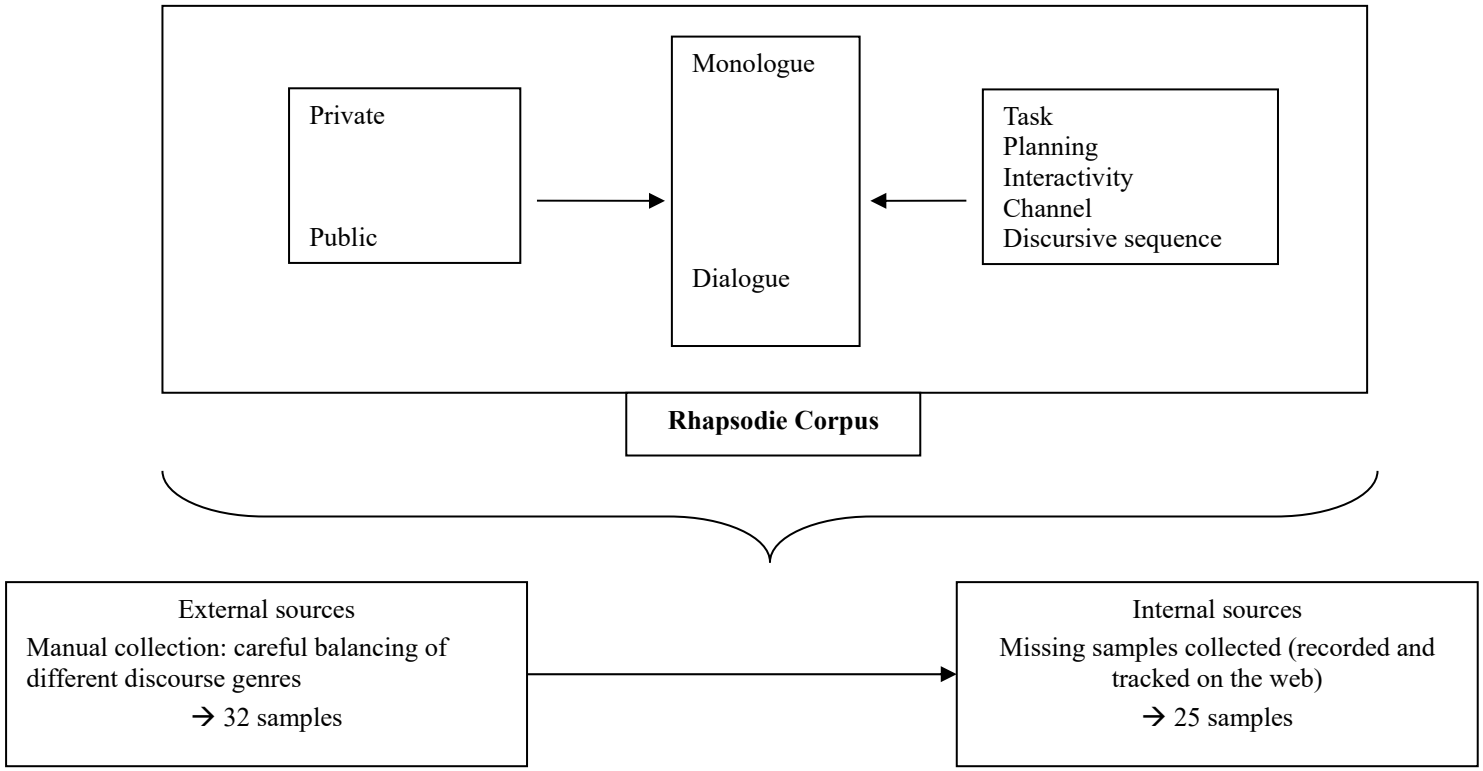

Figure 1.

Corpus design and data collection in the *Rhapsodie* Project

4. Questions to be answered

This section first discusses legal and ethical issues, in particular with respect to intellectual property (section 4.1.) and presents the procedure to acknowledge the intellectual property of the creators of the source corpora, as well as strategies to refer to source corpora and to ensure the possibility of retrieving the original samples. Then it describes the tool and the method used to process the metadata available on Rhapsodie home page (http://www.projet-rhapsodie.fr) for each sample (section 4.2).

4.1. Legal and Ethical issues

Digital formats along with the development of the Internet have greatly facilitated the dissemination of linguistic resources in the academic and scientific world. The development of open access projects has however raised some crucial ethical and legal issues concerning the people recorded and authors' rights (Baude 2006). The collection of the Rhapsodie corpus raised three important ethical issues concerning (i) respect of the intellectual property of the creators of source corpora: How can the right of the scientist to have free access to sources be balanced against the protection of authorship? How should the scientific enrichment of existing sources be valued? What deontological measures should be taken when citing a second-hand sample in a publication? (ii) respect of the privacy of participants. What measures need to be taken in order to avoid the open publication of recordings causing harm to the recorded speakers? (iii) open access to data. How can a copyright on data, and at the same time free enrichment of the data, be guaranteed? In the absence of shared good practices, we adopted specific strategies to achieve a balance between these three competing priorities.

In order to guarantee respect of intellectual property, we endowed our corpus with CMDI metadata. The CMDI format allows for a full description of the source corpora and their

related publications.

We also adopted a well-defined citation format for the examples drawn from the Rhapsodie reservoir and based on external sources: these sources are cited by specifying the name of the Rhapsodie sample, preceded by the prefix 'Rhap-' and followed by the name of the source corpus, e.g.

(1) *j'accorde une puissance énorme à l'acte d'écrire* [Rhap-D2009, corpus Mertens]
(2) *c'était ils préféraient rigoler que de travailler* [Rhap-D0002, CFPP2000]
(3) *je suis heureux de me retrouver ce soir parmi vous* [Rhap-M2001, C-PROM]
(4) *et puis finalement bah on a choisi de rester* [Rhap-D0003, PFC]

Furthermore, we gave precise instructions for citing the publications related to source corpora: see column 2, table 2, the references in italics.

In order to guarantee the privacy of speakers we mostly selected recordings for which informed consent had been obtained at the outset and we anonymized all proper nouns, included toponyms. It should also be noted that the short-sampling strategy used in the constitution of the corpus limits the amount of information provided about speakers: as they are not easily identifiable, they are better protected.

In order to freely distribute our treebank and at the same time to protect our copyright, as well as the copyright of the source corpora, we adopted the Attribution, No Commercial, Share Alike Common Free License (http://creativecommons.org/licenses/by-nc-sa/3.0/fr/deed.en).

4.2. Metadata

We chose to encode our metadata in the IMDI-CMDI format developed at the Max Planck Institute for Psycholinguistics in Nijmegen (CMDI, http://www.clarin.eu/cmdi) (Broeder et al. 2011, 2012). This format is flexible enough to be adapted to the metadata encoding needs of different projects. In Rhapsodie, we did not need a fine-grained description of speaker sociolinguistic characteristics, nor did we need to detail the modalities of data collection, since this had been basically conducted in the framework of the source projects (section 3).

Rather we needed: (i) to explicitly mention and describe the source corpora; (ii) to guarantee access to the source files; (iii) to describe in detail the speech situation; (iv) to provide for each sample a thorough description of the annotations; (v) to acknowledge the intellectual property of annotators.

Using CMDI we defined the profile of the metadata we needed. Generally speaking, we chose to provide information on the sample, speakers, discourse situation, corpus sources, and written and media resources associated to each sample. The use of these components easily allowed us to meet our needs. In particular, we used the IMDI "source" component to give a detailed description of the source corpora. In the "session" component we assigned each sample a unique identifier that allows for rapid retrieval of the source sample and its metadata. With only a few modifications to the IMDI "discourse" component, we were able to provide a complete description of the situational variables characterizing each sample. In the "written source" component we both provided detailed information concerning the annotation of each sample and we acknowledged the intellectual property of annotators.

In order to manipulate the metadata we used the Arbil tool associated to the CMDI format (http://tla.mpi.nl/tools/tla-tools/arbil/description/), which allowed us to easily edit XMLs for our metadata and to convert them into HTML.

A tool available on the site of the project, www.projet-rhapsodie.fr, can be used to browse the metadata and select samples on the basis of specific characteristics: the sex of speakers, their age, the type of genre (discourse or religious/ritual speech), the type of sub-genre (argumentative, descriptive, narrative, oratory, procedural, or conversational), the task (advertising, lessons, life-stories, etc.), the type of interactivity between speakers (interactive, non interactive, semi-interactive), the social context (private or public), the event structure (monologue, dialogue or conversation), the planning type (planned, semi-spontaneous, spontaneous), and the source corpus (Figures 2 and 3).

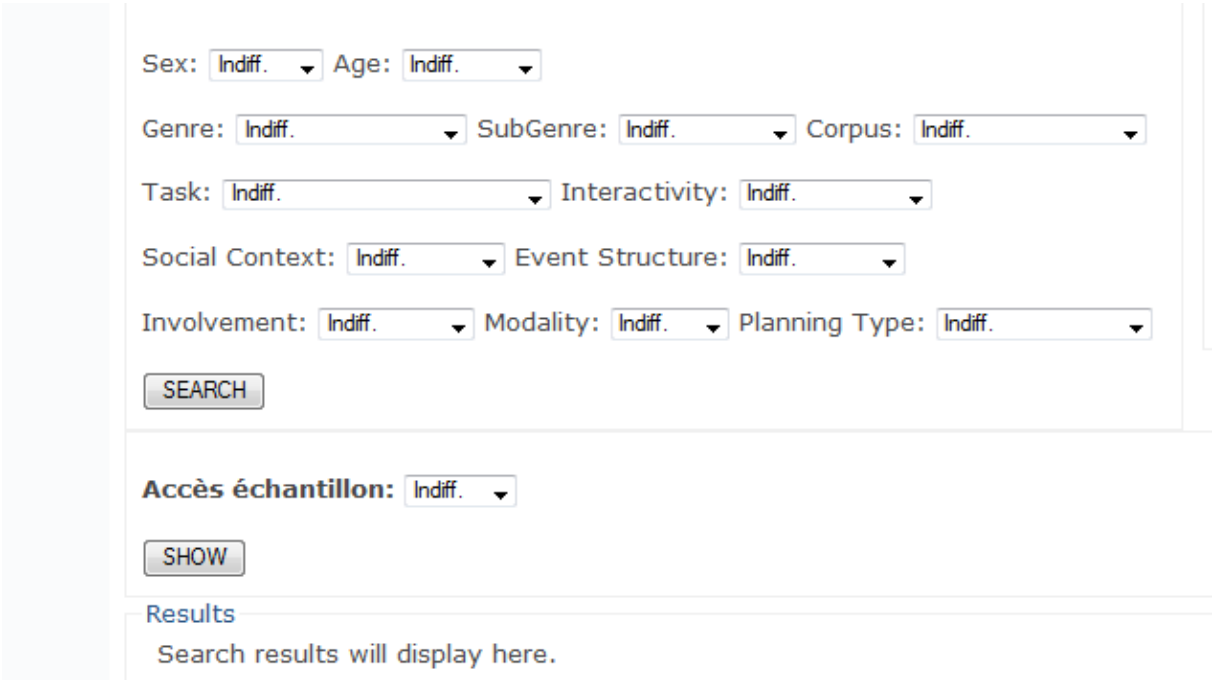

Figure 2. The Rhapsodie metadata browser

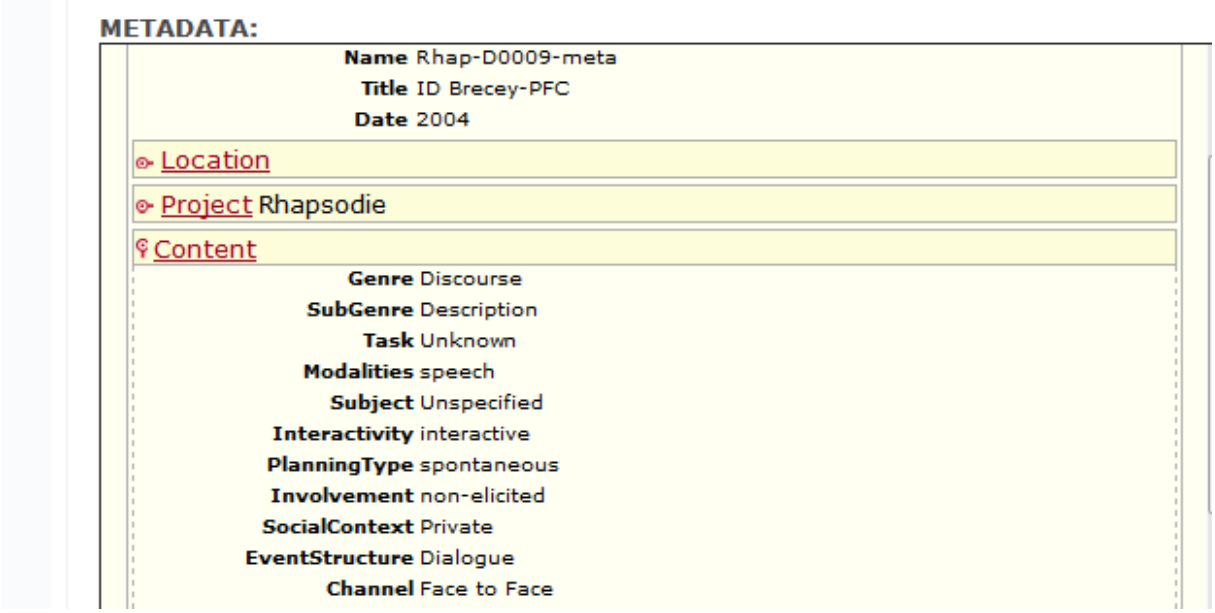

Figure 3. The description of genres based on the IMDI vocabulary for the discourse component

5. Conclusions

In designing the Rhapsodie corpus we had a clear objective: to collect a sufficient variety of text types to test and improve the flexibility of our annotation schemata. We also had a significant constraint: we could not rely on a pre-existing representative corpus of spoken French from which to extract our reservoir.

This limitation obliged us to select and extract excerpts from a number of different corpora and to collect new data wherever a given text type was not represented. This *bouquet de corpus* approach had never been adopted before, at least in France, which raised a number of challenging new theoretical, ethical, and legal questions.

From a theoretical point of view, we could not count on a unified model of textual diversity. We decided therefore to presuppose a correlation between the heterogeneity of speech situations and the variety of text types. We hypothesized in other words that each particular speech situation engenders a number of specific linguistic constructions. We collected data along a number of axes of situational variation: monologues vs. dialogues, private vs. public speech, interactive vs. non interactive speech, face-to-face vs. broadcast samples, more or less formal register. As we will see in the following chapters, the completeness of the *Rhapsodie* annotation schemata is precisely due to the variety of texts that have been annotated.

From an ethical and legal point of view, a number of questions arose due to the fact that we actually made public use of public resources. This apparently trivial task led us to define Good Practice guidelines with respect to the acknowledgment of the intellectual property of both authors and annotators as well as of the privacy of speakers. This led us to propose a short sampling strategy so as to optimize the anonymization, to define a standard for the citation of second-hand data, and to choose a flexible and detailed format of metadata such as CMDI to guarantee a fine-grained and complete description of source corpora, annotations, and speech situations.