THE *THESAURUS OCCITAN* : A MULTIMEDIA DATABASE DEDICATED TO OCCITAN DIALECTS *PRESENTATION OF ITS MORPHOSYNTAX MODULE*

Pierre-Aurélien Georges
Laboratoire BCL, CNRS UMR 6039
Université Nice Sophia-Antipolis, MSH de Nice (France)
pageorge AT unice.fr

Abstract

The *Module MorphoSyntaxique* (abbreviated MMS) is a computer tool especially designed for syntactic and morpho-syntactic analysis of Occitan dialects.

It is part of the *Thesaurus Occitan* multimedia database (of which a general presentation can be found in these proceedings in another article by Guylaine Brun-Trigaud).

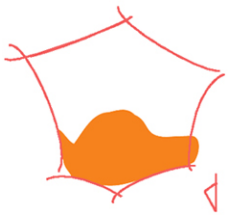
Following the THESOC's general guidelines (i.e. localised and oral data only), this module contains both oral texts (including *ethnotexts*) and single sentences, such as answers to morphosyntactic questionnaires.

The "oral data" criteria can be somewhat flexed: even if this module was originally conceived for oral data processing, its part-of-speech tagger and syntactic parser are still able to process written texts so far as they are written in a familiar or popular style, close to oral register.

The locations where all these texts and sentences have been harvested are stored in the database, thus enabling on the long term a comparison between different dialects on a morphosyntactical or syntactical basis, thus opening new perspectives for dialectology.

Keywords

Comparative syntax ; morphosyntax ; Occitan dialects ; *ethnotexts* corpus ; database



1. General Presentation

The *THESAURUS OCCITAN* (abbreviated THESOC) is a multimedia database which encompasses oral dialectal data from the whole Occitan domain. Aside from its lexical and microtoponymy corpuses, the THESOC also contains a module dedicated to morphosyntax and syntax analysis of Occitan dialects [1]. This *Module MorphoSyntaxique* (MMS)[2] database is aimed at studying oral syntax and morphosyntax in a comparative way.

[1] For a general presentation of all the other aspects of the THESOC, please refer to the article by Guylaine Brun-Trigaud also included in these proceedings, and/or (Georges, not yet published, a).

[2] It was formerly known as « base TEXTES » in (Georges, 2009) because it originally started has a corpus of texts and *ethnotexts*, as explained in (Oliviéri, 2003).

1.1. Conditions

As for the rest of the THESOC, raw data to be included in this MMS module must match the two following criterions:

- *Location condition:* linguistic data must be precisely located. This constitutes an essential condition for diatopic variation studies. In particular, this condition will enable dynamic and automated generation of linguistic maps on demand in the future.
- *Orality condition:* linguistic data should come from oral sources. Indeed, our philosophy here is to integrate linguistic facts collected under oral form (with IPA transcription), which guarantees the reality of these facts under consideration. Moreover, the THESOC allows hearing of the audio tracks recorded during the field works, thus giving to the user the possibility to control or to check the proposed transcription.

1.2. Different types of data

On the one hand, the database contains a collection of single sentences or isolated sentences: answers to morphosyntactic questionnaires such as PAM [3], sentences found in unpublished survey notebooks of the *Atlas linguistiques*[4], or sentences published in some of these atlases, such as the ALMC[5].

On the other hand, the database also contains a collection of texts, such as *ethnotexts* collected on the field, or radio broadcastings.

Similarly to the other parts of the THESOC, multimedia documents such as pictures, sound files, or video files, can be attached to a text or a single sentence.

1.3. Text types and orality condition

Although this module was originally conceived for treating oral data only, its lemmatiser and syntactic parser presented in sections 2.2. and 2.3. are even capable of treating written texts as long as they are written in a familiar or popular style, close to oral register; thus dimming somewhat the orality condition required in section 1.1. This orality condition may then be eventually reformulated as the following: linguistic data must contain oral syntax or popular / close-to-oral syntax.

This possibility has allowed us to extend the corpus with other types of texts: some theater plays, articles from popular press, etc. However, each text record from the database contains a field that informs the user about its type / gender. As shown in Figure 1, *Ethnotexts* are thus easily identifiable by the "genre: Ethnotexte" field content, whereas other types of texts have another tag in this field, such as "Chanson" (song lyrics) or "Presse" (press article) for example. Thanks to this field, it's always possible to filter out written texts and to focus only on *ethnotexts* and/or some other types of oral texts: search queries can be configured to show results from the whole corpus or from only certain types of texts specified by the user. This way, linguists can decide whether to stay on strictly oral data, or to also include some type of written texts in order to get a broader corpus.

---

[3] Parler des Alpes Maritimes, supervised by (Dalbera, 1994).
[4] *Atlas Linguistiques de la France par régions*, éditions du C.N.R.S, as presented in (Séguy, 1973).
[5] Atlas Linguistique du Massif Central, (Nauton, 1957-1963)

Fig. 1: example of an *ethnotext* record.

2. Data processing

2.1. Adding new data to the corpus

Data can be added to the database through its XML import / export functionalities[6] or by typing new records directly within the user interface.

The graphical transcription field of a new record can have two different origins, depending on the nature of this record:

- if it's a text (written in a familiar or popular style, close to oral register), for which a graphical transcription is, by definition, already available, it is directly stored in the database, whatever writing conventions or writing system have been used by its author, since there is no unique official spelling norm (or "*orthographe*") for the Occitan dialects as is the case for French or English, but rather several writing systems that competes[7]. The database also contains an algorithm which tries to automatically detect which writing system has been used within a text[8], for users' information.

- if it's based on a sound track recorded on the field, for which the user doesn't already posses a graphical transcription (as is typically the case with *ethnotexts*), it's possible to easily generate a phonological graphical transcription: the user simply presses on a button (button #2 shown in Figure 2 below) to call an automated transcriber which is available in MMS (as in the rest of the THESOC), that generates a graphical transcription (#3) directly from the IPA transcription (#1), which must therefore be typed or imported in the database first. This automated transcription is based on conversion rules that

[6] Thus, TEI import / export can also be achieved by using an XSLT filter.

[7] The three most common writing systems used are "*graphie alibertine*", as defined by (Alibert, 1966), "*graphie mistralienne*" as in (Mistral, 1979), and in a far smaller proportion, the Eastern part of Occitania sometime uses "*graphie italianisante*", which is inspired from Italian's writing system.

[8] This is based on statistical occurrences of some sequences of letters, and these rules are user-customizable, therefore, any new writing system can be added to this automated recognition feature. Should this algorithm ever fail, on a very short text for example, where the situation is unclear, it is always possible to override it and to manually specify the correct writing system used by the text.

can be modified by the user and adapted to different phonological systems among dialects, so that the transcriber will automatically use different rule sets to transcribe different dialectal areas, depending on the information given by the locality field of the record to be processed. The result is a phonological graphical transcription close to (Mistral, 1979)'s writing system known as "*graphie mistralienne*", which provides users a more readable way to access the content of the text, and allows a first level of abstraction that "smoothes" or "hides" phonetic variation[9].

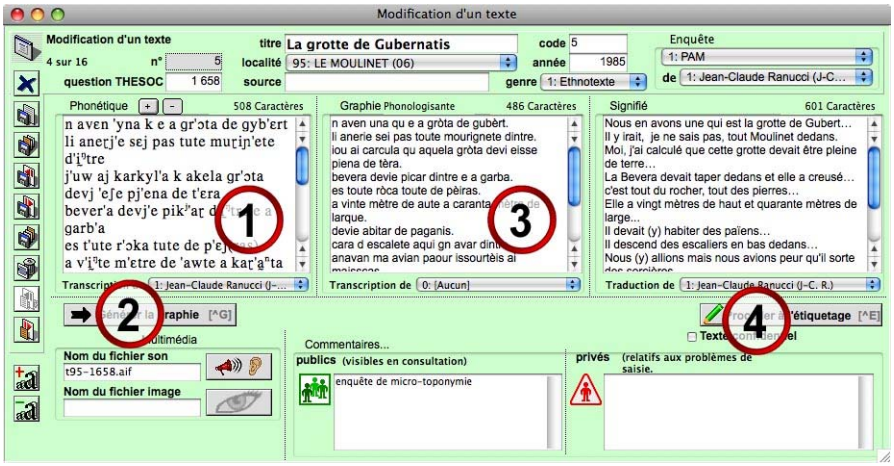


Fig. 2: processing a text record.

In both cases, all further linguistic treatments proposed by the different tools available in MMS are based on this graphical transcription. Thus, while these tools were originally conceived for oral data processing, they are even able to process written texts so far as they are written in a familiar or popular style, close to oral register. This is how it is even technically possible to introduce "oral-style" written texts in the database.

Among these tools, the part-of-speech tagger and the syntactical parser automate a great amount of the annotation work, thus simplifying processing of new data, as shown in following sections.

[9] Since the objective of MMS is to study syntax and morphosyntax, it's not a major issue here to disregard phonetic variations in order to simplify the following linguistic treatments performed.

2.2. Lemmatisation process

The next step after importing or typing new data in the database is the lemmatisation process (button #4). The part-of-speech tagger identifies each individual lexical element of a sentence or a text by using a reference dictionary embedded in the database.

In order to manage efficiently the variation in all its aspects (graphical, dialectal, or inflectional variation), this dictionary is structured into two hierarchical levels: the *variantes,* which are in fact the lexical occurrences found in the corpus, are grouped under *lemmas*. This allows performing searches or linguistic treatments either on a particular form (e.g. such inflection, with such graphical transcription, in such dialect), or on all forms associated to a given lemma. Figure 3 below illustrates this two-levels dictionary structure:

Lemmas		
lo	determiner	"the"
li	pers. pronoun	"him"
fòl	adjective	"crazy"
bèl	adjective	"beautiful"

***Variantes* of a lemma**		
fuala	fem. sing.	Nice
fuale	masc. plur.	Nice
fuali	fem. plur.	Nice
fòlha	fem. sing.	Toulouse

***Variantes* of a lemma**		
bèl	masc. sing.	Nice
beu	masc. sing.	Nice
bela	fem. sing.	Nice
bella	fem. sing.	Nice

Fig. 3: database's dictionary structure.

This schematic illustrates the different types of variation handled:

- phonological variation, between *bèl* (preceding a vowel) et *beu* (preceding a consonant) in the dialect of Nice, called *"Nissart"*
- dialectal variation, between *fòlha* in Toulouse's dialect and *fuala* in Nissart,
- graphical variation, in Nissart, between *bela* and *bella*,
- and inflectional variation, also in Nissart, between *fuala*, *fuali* and *fuale*.

As reference forms, lemmas' graphical forms are based on entries from (Alibert, 1966) when available [10] whereas the *variantes* part of the dictionary is currently populated by entries coming both from the lexical database of the THESOC and from a set of several paper dictionaries, such as (Eynaudi, 1932), that have been digitalized and integrated in the database for this purpose. Moreover, the process of manual lemmatisation of unidentified lexical items sometimes adds new entries to the dictionary, as detailed below in this section, as well as the syntactical-tree annotation of sentences. Thus, dictionary's content is constantly improved by the lemmatisation process of new texts and/or sentences and new entries coming from THESOC's lexical database.

[10] When there is no corresponding entries in (Alibert, 1966), an asterisked form is proposed in Alibert's writing system, known as "*graphie alibertine*".

As illustrated in Figure 4 below, the output of the lemmatisation process (#5) shows the following information about each lexical item identified in the text or sentence processed: its lemma, morphosyntactic category, and inflection. Some items may remain unidentified after the lemmatisation process, either because there is no matching entry in database's dictionary or because there are several potential candidates.

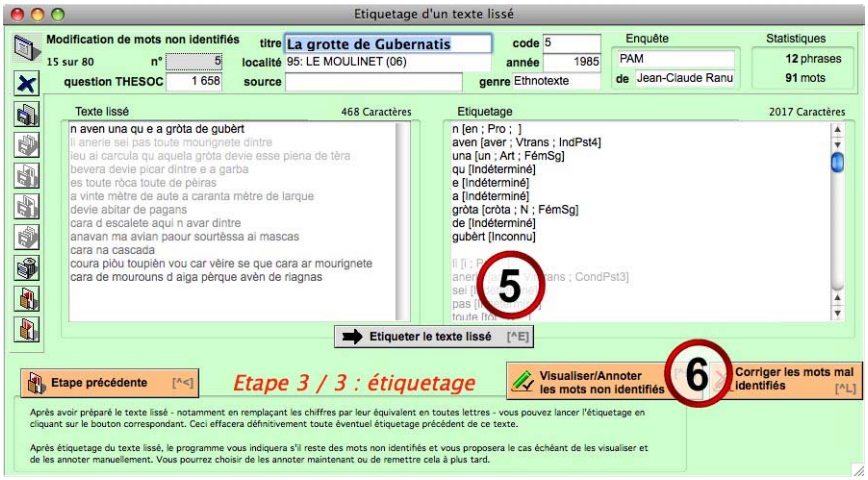


Fig. 4: lemmatisation process.

As can be seen in Figure 4, unknown items are tagged "*Inconnu*" (as is the case of proper noun "*Gubèrt*") whereas ambiguous items are tagged "*Indéterminé*": the graphical form *"e"* can match either with an inflected form of the verb *èstre* "to be" (3rd Pers. Sing.) or with the conjunction coordination "and" ; similarly, *"a"* can match either the definite determiner (Fem. Sing.) or an inflected form or the verb *aver* "to have" (3rd Pers. Sing.).

When a lexical item was not correctly identified or not identified at all by the lemmatisation process, the two buttons labelled #6 in Figure 4 allows to manually identify this lexical item. In the case of ambiguous item, with several potential matching candidates in database's dictionary, the user can choose the right entry within a list of these potential candidates as shown by #7 in Figure 5. If the unidentified lexical item has no corresponding entry in the dictionary, it's also possible to add a new entry in the dictionary (#8) and to identify the lexical item with this new entry (these two points are realized at a glance, as one single operation). This unique user interface also shows the unidentified lexical item in context (bottom of Figure 5) to ease its manual identification by the user.

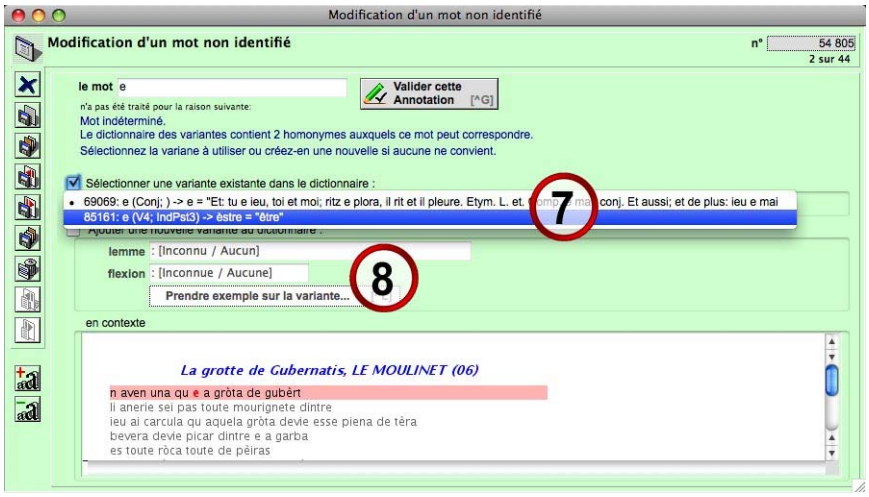


Fig. 5: annotation of ambiguous items.

After lemmatisation process has been performed, the next step is syntactical-tree tagging, eased by MMS' syntactical parser. It is not necessary here to lemmatise each single lexical item of a sentence or a text before using MMS' syntactical parser: if 80 % or 90 % of the lexical items have been correctly lemmatised, this is enough to switch to next step, as will be explained below.

2.3. Syntactical tree tagging

MMS' syntactical parser uses both data generated by the lemmatisation process and the embedded dictionary of the database in order to propose one or several possible syntactic structures for each sentence[11], thus automating in some proportion the syntactical-tree tagging process. The syntactic trees are generated in a generativist theoretical framework, but the syntactic rules on which the parser relies are user-customizable[12].

Figure 6 below gives an example of this process. On the left pane, the results of the lemmatisation process are shown. By clicking on the button #9, the user calls the syntactical parser, which analyses the sentence and output some possible syntactic structures in a list shown just under this button. The user must then browse among these candidates and chose the correct hierarchical syntactic structure(s) that really correspond to this sentence: by clicking on a candidate structure within the list, the hierarchical representation of the selected structure is shown in the right pane[13]. Since the candidates list is sorted by probability (most probable candidates are located on top of the list, whereas least probable ones are at the bottom), users therefore typically only need to look at the two or three first propositions to find the correct one(s).

Thanks to the two buttons under #10, false candidate structures are then eliminated in order to keep only the correct one, or the two or three right ones (when there's an

[11] Wether it's an isolated, single sentence, or a sentence from a text.
[12] The only main constraint is that generated syntactical trees must have binary branching nodes.
[13] Please note here that the syntactical tree structure (in the right pane) is displayed vertically instead of the traditional horizontal presentation. This is due to some technical limitations we are currently working on.

ambiguity in the sentence that can not be re solved, for example if the author of this sentence has deliberately made a play on words). If the appropriate structure(s) is/are not present in the list genera ted by the syntactical parser, the user can choose any proposed candidate and edit th is structure in order to manually build the correc t hierarchical tree(s).

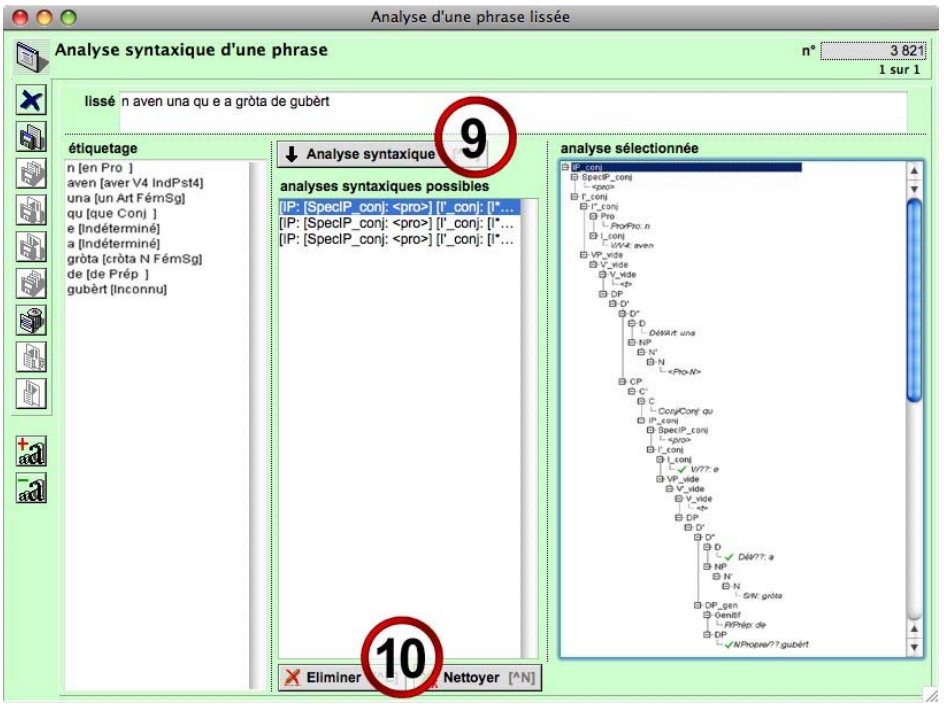


Fig. 6: syntactical structure annotation.

The use of MMS' syntactical parser thus simplifies and speeds-up the syntactical-tree tagging task by providing the user a semi-automated way of generating hierarchical structures. Moreover, it also simplifies and speeds-up the lemmatisation process: Figure 6 shows for exam ple that the unknown proper noun " *Gubèrt*", that had been n either correctly identified by the a utomated lemmatiser nor manually tagged by the user, has nevertheless been correctly identified as a proper noun by the syntactical parser. Similarl y, the am biguous elements "*e*" and " *a*" have been successfully identified as respectively an inflected verb and a determ iner. Therefore, the syntactical parser can autom atically "guess" some lexical item s that remained unidentified after the lemmatisation pro cess, and disam biguate some ambiguous lexical items, thus also sim plifying the lemmatisation task, since it's no longer mandatory to tag 100 % of all the lexical items of a sentence.

After these 10 steps have been perfor med, annotation process is now com plete: each lexical item of each sentence has been lemmatized, and each sentence is tagged with one or several hierarchical syntactic structure(s). Data are then ready to be exploited through different work features offered by the database.

3. Work features

3.1. Search engine

Once the sentences and texts from the database have been annotated with the help of these tools mentioned in section 2., MMS provides several search options to select data with different criterions. For example, it's possible to search all occurrences:

- of a given *variante*
- of all *variantes* associated to a given lemma
- of all *variantes* with one or several given morphosyntactic categories
- of a given sequence of part-of-speech categories[14]
- of a given location[15]
- having a graphical similarity with a given lemma or *variante*.

It's also possible to perform searches based on syntactical tree tagging previously processed: one can search for all sentences containing a particular syntactic structure or syntactical tree fragment, such as all sentences containing a DP within another DP for example. When an ambiguous sentence has several syntactic structures associated to it, such a search query will show this sentence in the search results if it matches any of its associated structures.

As illustrated in Figure 7, the results are shown in context: the list of matching occurrences is presented in column, and each matching occurrence is displayed within the full original sentence where it is coming from.

3.2. Automated work corpus generation

Whichever search query is formulated, a work corpus can then be automatically generated from these search results[16]. As presented in Figure 7, it is possible, for example, to search for all occurrences with morphosyntactic category « personal pronoun », then to select some particularly interesting ones in the results list displayed[17], and to automatically generate a work corpus in which these selected occurrences are highlighted in their original context: text title[18], location, full sentence or full text where the term occurs. In this example, we chose to generate a "concise" work corpus, i.e. only matching sentences are being extracted and displayed gathered, but it is also possible to generate a "full" work corpus, which gathers both matching isolated sentences and the full texts that contains at least one sentence containing one occurrence matching the search query.

---

[14] Some wildcard options are available, such as "Beginning of sentence only" / "End of sentence only".

[15] This is, all occurrences of all texts associated to a particular given location

[16] Except if the search query concerns all occurrences of a given location, because this would generate a full list of all texts and isolated sentences coming from this location, and thus would have merely no interest here since it is already possible to get this information from elsewhere in the database.

[17] Of course, it's also possible to select all occurrences from the results list in order to generate an exhaustive work corpus from the database.

[18] If the sentence comes from a text.

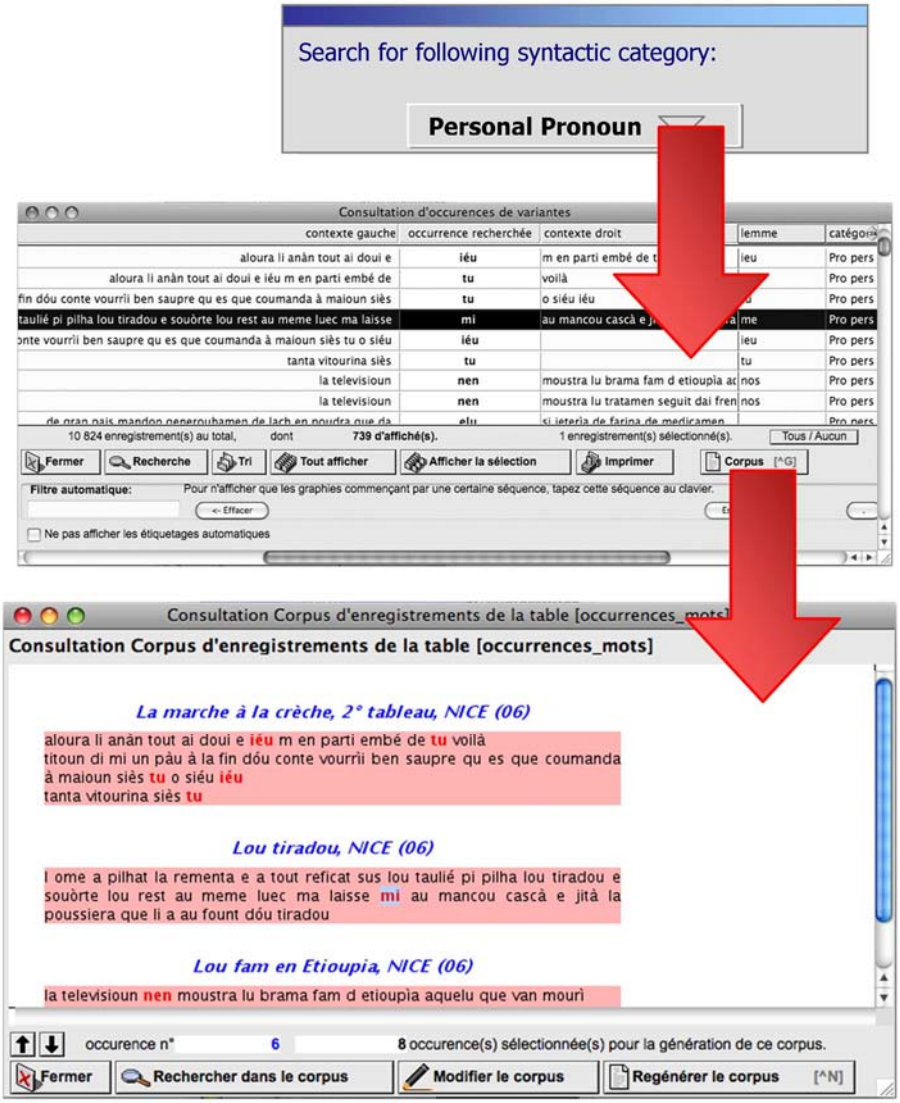


Fig. 7: search by syntactic category and work corpus generation features.

In the same way, one can generate a work corpus from search results based on syntactical tree tagging previously processed. These work corpuses can be exported and saved as RTF or Microsoft Word format for further exploitation.

3.3 User-customizable taggings

In 2009, a new feature has been added to MMS: users now have the ability to attribute some user-defined "tags" to some sentences of the database [19]. Then, it's possible to perform cross searches based on these tags within the database. This allows, among other things, to search for presence or absence of a correlation between two or several linguistic parameters, confirming or invalidating user's hypotheses[20].

It's possible, for example, to search within the database for texts that contain at least one sentence with a given tag but do not contain any sentences with another given tag (evaluating a hypothetical positive correlation between two parameters within the same speaker, therefore the same idiolect). Another use would be to search within the database for locations for which there is at least one sentence with a given parameter (no lexically realised subject pronoun, for example) and also at least one sentence with another given parameter in the same location (evaluating an hypothetical negative correlation between two parameters within a dialect).

In a generative theoretical framework, one could for example use this feature to try to check if (Rizzi, 1982)'s correlation between *pro drop* and *free inversion* of subject and verb can be verified (or invalidated) on the field by dialectal data, such as Occitan dialects[21].

Conclusion

The MMS module of the THESOC contains flexible tools and features that can be used of several maners and under different theoretical frameworks. Therefore it can fully play its role of a computer tool aiding syntactical and morphosyntactical research on Occitan dialects on a comparative basis.

Currently, we are working on adding cartographic functionalities to MMS, that would fosters the use of MMS for diatopic variation study, giving the users the opportunity to visualize variation of some syntactic or morphosyntactic features among the occitan domain and the possibility to generate maps on demand to illustrate and to support their hypothesis.

Aside from the diatopic perspective, the data entries in the database also contain a "date" field to indicate for each text and each single sentence, at which date they have been collected on the field. So, if enough data is available, it would be theoretically possible to study variation also from a diachronic point of view.

---

[19] Whether isolated sentences or sentences from a text.

[20] For some interesting applications of this feature and concrete cases, cf. (Georges, not yet published, b)

[21] Dialects from the northern part of Occitania show overt pronoun subjects, whereas the majority of the Occitan dialects are *pro drop*. Comparing these Occitan dialects, which are all closely related, within a microvariational approach, could therefore shed new light on this hypothetic correlation.

References

Alibert, Louis, 1966, *Dictionnaire Occitan-Français d'après les parlers languedociens*, Toulouse: Institut d'Etudes Occitanes.

Dalbera, Jean-Philippe, 1994, *Les parlers des Alpes-Maritimes, étude comparative, essai de reconstruction*, Association Internationale d'Etudes Occitanes.

Eynaudi, Jules, 1932, *Dictionnaire de la Langue Niçoise*, Imprimerie de "l'Eclaireur de Nice", Nice.

Georges, Pierre-Aurélien, 2009, *Présentation de la base Textes associée au THESOC*, actes du colloque La dialectologie hier et aujourd'hui (1906-2006), Lyon, 2006, Brigitte Horiot (Ed.). pp. 81-94.

Georges, Pierre-Aurélien, not yet published, a, *Le THESOC: bases de données et outils d'analyse consacrés à l'étude des dialectes occitans*, actes du colloque Bases de données, Méthodes, Modèle s de description: de nouve lles perspectives pour la recherche sur les langues régionales et minoritaires? Tübingen, 2008, Stauffenburg Verlag (DeLingulis).

Georges, Pierre-Aurélien, not yet published, b, "Les ch aînes de c litiques: l'outil informatique au service de l'analyse comparative", in: *actes du colloque Mémoires du terrain: enquêtes, matériaux, traitement des données* (Lyon, 2009).

Mistral, Frédéric, 1979, *Lou Tresor Dóu Felibrige ou Dictionnaire Provençal-Français*, Raphèle-lès-Arles: Culture Provençale et Méridionale.

Nauton, Pierre, 1957-1963, *Atlas Linguistique et ethnographique du Massif Central*, Editions du CNRS, Paris.

Oliviéri, Michèle, 2003, *Constitution d'une base de textes occitans*, 36ème colloque de la Societas Linguistica Eur opaea : Linguistiqu e et Corpus, Lyon, 4-7 septembre 2003.

Rizzi, Luigi, 1982, *Issues in Italian Syntax*, Dordrecht, Foris.

Séguy, Jean, 1973, "Les atlas linguistiques de la France par régions", in: *Langue Française* Volume 18, Numéro 1, pp. 65-90.