Multiword expressions in scientific discourse: a corpus-driven database

Agnès Tutin\textsuperscript{1}, Achille Falaise\textsuperscript{2}

\textsuperscript{1}LIDILEM, Université de Grenoble Alpes
Université Stendhal Grenoble III, BP 25
38040 Grenoble cedex 9, France
\textsuperscript{2}LIG-GETALP, Université de Grenoble Alpes
385 avenue de la Bibliothèque,
38400 Saint Martin d’Hères, France

E-mail: agnes.tutin@u-grenoble3.fr; achille.falaise@imag.fr

Abstract

This paper presents an on-going project of corpus-cum-dictionary in the field of academic multiword expressions in French. This project is a lexicographic and pedagogical application of the Scientext corpus and aims to provide writing tools to help academic learners of French as a foreign language to produce high quality texts. Our system is based on a large syntactically annotated corpus and provides an onomasiological and a semasiological access to phraseological entries, including corpus examples stored in dynamic web pages.

Keywords: multiword expressions – collocations – attitudinal multiword expressions – corpora

1. Introduction

As outlined by several scholars, Loffler-Laurian (1995), Gledhill (2000), Pecman (2004), and Granger & Paquot (2010), fixed expressions are prevalent in the genre of academic and scientific writings and need to be fully mastered in order to become a member of the academic community. This paper describes an on-going project of a phraseological corpus-based database for French, where several kinds of multiword expressions can be accessed with several modes, mainly via a semasiological access and an onomasiological access. This project is a lexicographic and pedagogical application of the Scientext corpus, and aims to provide writing tools to help academic learners of French as a foreign language to produce high quality texts.

We will first describe the different types of academic multiword expressions addressed in this project. We then focus on attitudinal multiword expressions and their properties. Finally, the structure of our lexical database is presented as well as selected corpus examples.

2. Different kinds of academic multiword expressions

Our phraseological database includes frequent specific and cross-disciplinary multiword expressions pertaining to scientific activity, scientific reasoning, and scientific discourse (Cf Pecman 2004, Tutin 2007; Tutin forthcoming). The database includes several types of multiword expressions, each type being processed specifically within the database and the corpus. All the multiword expressions (MWE) are frequent and specific expressions were extracted from our corpora with NLP techniques. The different types of MWE are treated differently.

2.1 Referential expressions are widely studied in academic discourse. They are mainly lexical collocations (e.g. faire une hypothèse literal ‘make a hypothesis’) or frozen non-compositional expressions (e.g. tenir compte ‘to take into account’ or point de vue ‘point of view’) related to scientific activity, scientific objects or scientific reasoning. We posit, in line with Explanatory and Combinatorial Lexicology (Mel’čuk, 2011), that these two different kinds of MWEs should be differentiated. As collocations are compositional, each component will receive a semantic tag and be stored within the database entry (e.g. s.v. hypothèse for refuter une hypothèse (‘to disprove a hypothesis’) or hypothèse valide(‘valid hypothesis’). Fully frozen MWE expressions (e.g. mettre en évidence ‘highlight’) will be stored as single units.

2.2 Discursive MWEs aim at relating textual parts and are mainly structural and metatextual markers (to begin with, in other words) and logical connectives such as that is why. Mainly adverbials, but also conjunctions or prepositions, they are rarely specific to academic writing, in constrast to certain referential expressions. Nevertheless, many of them are more frequent in scientific writing such as pour conclure ‘to conclude’ (140 times more common than in literary and newspaper corpora) (Tran, forthcoming). Describing these MWEs requires explaining their discourse function (position, meaning and rhetorical function).

2.3 Attitudinal MWEs are mainly related to stance (e.g. jusqu’à un certain point ‘to a certain extent’) and modality (e.g. deontic modality il est nécessaire de ‘it is essential to’). As previously found with discursive MWEs, the attitudinal MWEs discussed here are not fully specific of our genre, but occur more often in our corpora than in
general French. We selected the most frequent and the most specific attitudinal MWEs present in our academic corpora.

2.4 More complex semantico-rhetorical routines, fully specific of the scientific genre (see Sandor 2007; Teufel 1998; Tutin 2010; Granger & Paquot 2010; Jacques et al. 2013) are presented. They are associated to rhetorical functions such as comparison with peers or definition of the main research objective. In the Scientext project, which forms the source of our phraseological database, these semantico-rhetorical routines are extracted from the corpus with the help of local grammars by using syntactic relations.

3. Methodology

Our phraseological database is based on several principles:

1. All the phraseological resources are based on a large on-line corpus of scientific writings, Scientext, whose core is freely accessible via the web.

2. The database provides two kinds of access, an onomasiological access via semantic data, and a semasiological access via lexical entries. Simple semantic syntactic and certain discursive information is provided as well as information on frequency.

3. As the database is intended to be a writing aid and can be considered as a dictionary-cum-corpus project (Hartmann 2005; Granger & Paquot 2010), phraseological data are associated to selected on-line examples. A dynamic access to web pages is provided, with the possibility to access larger contexts and statistics, e.g. parts of texts where a multiword expression is most frequent.

Our database is constructed with the help of the Scientext corpus, syntactically annotated with the Syntex parser (Bourigault 2007). Syntactic relations between words are particularly useful to extract complex MWEs in which the components may be non-contiguous. An example of a parsed sentence is given in Figure 1.

The construction of the database follows several stages:

1. **Stage 1: Extraction of frequent and cross disciplinary MWEs** with simple NLP techniques, mainly n-gram techniques and lexical associations. Other techniques using lexicosyntactic patterns (see Corman 2012) will be experimented in this project.

2. **Stage 2: Study and linguistic analysis of MWEs, based on corpus examples.** For this analysis, examples are manually extracted with Scientext queries. Some queries may be quite complex when local grammars in which grammatical relations are involved (Falaise et al. 2012). This is especially the case with semantico-rhetorical patterns (Tutin 2010), as in the following example on “formulating a hypothesis”.

   ```
   //TITRE : Formulation d'une hypothèse
   //INFO: Expressions pour formuler une hypothèse
   (OBJPASSIF,#3,#1) = (SUJ,#1,#2) (AUX,#3,#2)
   $formul=avancer,émeter,faire,formuler,prendre,proposer,retener,soutenir,utiliser
   Main = <lemme=$formul,#1> & & <lemme=hypothèse,#2> :: (OBJ,#1,#2) OR (OBJPASSIF,#1,#2) OR (ADJ,#2,#1)
   ```

   Figure 2: An example of a local grammar associated with the semantico-rhetorical routine “formulating a hypothesis”

Here, the query indicates that several verbs such as *formuler* (‘formulate’) or *émeter* (‘emit’) may have a direct object *hypothèse*. Passive alternation and past participle passive constructions are also included with the help of syntactic relations.

Figure 3 shows concordances associated to the query illustrated in Figure 2. It is interesting to note that syntactic relations enable one to extract word associations that are separated by a large number of words as in example 5, where *proposer* and *hypothèse* are separated by 15 words (*l'hypothèse selon laquelle la multiplication de l'ADN pourrait participer au contrôle de l'expression des gènes est proposée*). Syntactic analysis is essential to reducing noise and silence in queries, i.e. inaccurate examples in corpora or examples which would not be extracted by simpler techniques.
3. Stage 3: Selection of relevant on-line concordances associated with MWEs.

Concordances and authentic examples are invaluable for pedagogical purposes. NLP corpus based projects such as Alfalex are very promising in this perspective (Verlinden 2009). Nevertheless, annotated corpora with the help of NLP techniques cannot be used without human intervention mainly for two reasons. First of all, NLP techniques are not error-free and some extracted examples can be erroneously analyzed, for example an incorrect part-of-speech category for a word. Secondly, some examples may appear inappropriate because they are too complicated to understand, especially for non-native speakers, or they may be atypical. For this reason, we have developed a system to filter on-line concordances, for which larger contexts and statistics remain available. Figure 5 demonstrates how to save the appropriate examples in a file, which will be uploaded and associated to an MWE entry. In this example, we have only selected occurrences for the MWE à l’évidence (‘obviously’, ‘clearly’). Examples including the other MWE se rendre à l’évidence (‘face the facts’) have been discarded.

Once MWEs are extracted, they are analyzed and encoded in a lexical database with several kinds of information: semantic, syntactic, discursive, or related to frequency. Detail examples of attitudinal MWEs will be discussed in Section 4.
CALL activities. Databases on discursive MWEs and attitudinal MWEs are under development (for discursive MWEs, see Tran forthcoming). Some semantico-rhetorical routines have already been developed (See Tutin 2010, Jacques et al. 2013 for English) but they need to be extended.

4.1 Fields of the database
Our linguistic description includes several fields for MWEs. It is intended both for pedagogical purposes, mainly as a writing aid in French as a Foreign Language, and for NLP and corpus linguistics extractions.

Our lexical entries include several fields, the main fields are presented with examples in Figure 7.

1. Lexical entries and variants of lexical entries. For example, de notre point de vue (‘in our opinion’) has a pronominal variant de mon point de vue (‘in my opinion’).

2. Semantic types are associated to MWEs, especially in the perspective of an onomasiological access. Several semantic types are given such as: ‘certainty/evidence’, ‘opinion’, ‘positive opinion’, ‘negative opinion’, ‘approximation’, ‘intensity’, ‘necessity’, ‘probability’.

3. Part of speech of the MWE. A POS is given for the whole MWE in order to fully grasp the syntactic behavior of the MWE. Most attitudinal MWEs are adverbials but some of them are more complex (e.g. il est possible, il est probable).

4. Part of speech of the components, e.g. for pour ainsi dire (‘so to speak’). This information is useful for NLP applications and corpus extraction.

5. Frequency. This information is highly relevant for pedagogical purposes. We indicate an approximate value (***: frequent, **: quite frequent, *: not so frequent). A precise value is given in the database corresponding to the number of appropriate examples extracted from the corpus.

6. Distribution of adverbials within the sentence is a very complex issue for non-native speakers. A special attention is paid to this issue. For example, à première vue (‘at first sight’) can appear, as can be seen in Figure 6, at the beginning of the sentence (example 1), before the verb (example 6) or more frequently just after the verb (examples 2, 3, 4, 5).

<table>
<thead>
<tr>
<th>expression</th>
<th>type</th>
<th>Cat.</th>
<th>Freq.</th>
<th>position dans la phrase</th>
<th>synonymes</th>
<th>Définition</th>
</tr>
</thead>
<tbody>
<tr>
<td>à première vue</td>
<td>la certitude/évidence</td>
<td>adv</td>
<td>*</td>
<td>préverbale</td>
<td>postverbale</td>
<td>tête de phrase</td>
</tr>
<tr>
<td>bien entendu</td>
<td>la certitude/évidence</td>
<td>adv</td>
<td>**</td>
<td>postverbale</td>
<td>tête de phrase</td>
<td>préadjectivale</td>
</tr>
<tr>
<td>bien sûr</td>
<td>la certitude/évidence</td>
<td>adv</td>
<td>***</td>
<td>postverbale</td>
<td>tête de phrase</td>
<td>bien entendu</td>
</tr>
</tbody>
</table>

Figure 6: Examples of the distribution of à première vue

7. Synonyms are introduced, preferably MWEs.

8. Short definitions are provided.

9. A link to concordances is given.

4.2 On-line Database

The phraseological database is integrated into a simple on-line dictionary-cum-corpus application. Our application is intended to be more of a corpus-cum-dictionary application than its opposite, since we believe that learners need most is a large number of appropriate examples. The dictionary is nevertheless useful to easily access the corpus information.

The phraseological entries can be accessed through an onomasiological access, with the help of the semantic types presented above. We are currently incorporating data for attitudinal MWEs and discourse MWEs as well as semantico-rhetorical routines, but our final application will also include referential MWEs, be they collocations or frozen expressions. Figure 8 provides four MWE expressions, sorted by descending frequency, of onomasiological access with the help of the semantic type “approximation”. The user has chosen one of these expressions au sens large (‘in the wide sense’) and a brief lexical entry is provided for the word including the POS, a definition and synonyms (with links in the database when available). Selected examples are also given on concordances, which can be dynamically extended to a larger context (Figure 9). Statistics (on genres – research articles, communications or theses –, textual sections and disciplines) are also available.

1 These MWEs are far more numerous than those previously mentioned and collocations, being compositional, are complicated to implement.
Expressions indiquant le point de vue et l'attitude dans les écrits scientifiques

### Onomasiological Access

**Expressions**

- en général
- en moyenne
- pour aller droit
- pour arriver
- pour être vrai
- pour être vrai
- pour être vrai
- pour être vrai

**Exemples**

1. L'expression « en général » est souvent utilisée en sciences naturelles pour désigner la moyenne des résultats obtenus.
2. « En moyenne » est souvent utilisé pour décrire les performances d'une équipe ou d'un groupe.
3. « Pour aller droit » est une expression qui implique de toujours avancer d'une façon ferme et directe.
4. « Pour arriver » signifie souvent que l'on réussit à atteindre un objectif.
5. « Pour être vrai » peut signifier que quelque chose est correct ou conforme aux règles.
6. « Pour être vrai » est une expression qui est généralement utilisée pour montrer que quelque chose est vrai.
7. « Pour être vrai » est une expression qui est généralement utilisée pour montrer que quelque chose est vrai.
8. « Pour être vrai » est une expression qui est généralement utilisée pour montrer que quelque chose est vrai.

Figure 8: An example of onomasiological access: MWEs of approximation

### Semasiological Access

**Expressions**

- à l'évidence
- à la fois
- à proprement parler
- à la dernière minute
- à bon escient

**Exemples**

1. L'expression « à l’évidence » est souvent utilisée pour décrire quelque chose qui est évident ou manifeste.
2. « À la fois » est une expression qui signifie que deux choses sont vraies à la fois.
3. « À proprement parler » est une expression qui est souvent utilisée pour dire que quelque chose est vrai par définition ou par définition.
4. « À la dernière minute » est une expression qui est souvent utilisée pour décrire quelque chose qui est fait à la dernière minute.
5. « À bon escient » est une expression qui est souvent utilisée pour décrire quelque chose qui est fait d'une manière prudente ou avisée.

Figure 10: An example of semasiological access: the example of à l’évidence

5. **Perspective and conclusion**

Our project can be considered more as a corpus-cum-dictionary than a dictionary-cum-corpus project. It aims at showing and teaching academic phraseology to non-native speakers of French with the help of authentic selected examples. The lexical database can be considered essentially as an onomasiological access to MWE occurrences drawn from the corpus. The originality of our database is that it will largely be built on a syntactically annotated corpus both for the extraction of MWEs and for the selection of relevant examples. Dynamically selected on-line concordances enable the learner to explore phraseological units in authentic scientific texts.

Several extensions are being considered. As in the LEAD Project or the BLF project (Verlinde 2009), we plan to associate CALL activities with this database by using annotated corpora including phraseological and semantic information on cross-disciplinary phraseological lexicon. First, we plan to display different kinds of MWE annotations, e.g. structural MWEs and attitudinal MWEs on the same text. Freely available research articles of the Scientext corpus could be used for this purpose. Second, we plan to build on-line exercises on specific aspects of MWEs, following several studies conducted by our research team in French (Cavalla & Loiseau, forthcoming) and in English (Hartwell & Jacques, 2012).
6. Acknowledgements
Special thanks to Laura Hartwell and Patricia Andreoli-Jones who very kindly read and commented a previous version of this paper.

7. References