DeQue: A Lexicon of Complex Prepositions and Conjunctions in French
Carlos Ramisch, Alexis Nasr, André Valli, José Deulofeu

To cite this version:
Carlos Ramisch, Alexis Nasr, André Valli, José Deulofeu. DeQue: A Lexicon of Complex Prepositions and Conjunctions in French. Language Resources and Evaluation Conference (LREC 2016), May 2016, Portoroz, Slovenia. pp.2293-2298. hal-01464822

HAL Id: hal-01464822
https://hal.archives-ouvertes.fr/hal-01464822
Submitted on 10 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
DeQue: A Lexicon of Complex Prepositions and Conjunctions in French

Carlos Ramisch, Alexis Nasr, André Valli, José Deulofeu
Aix Marseille Université, CNRS, LIF UMR 7279
FirstName.LastName@lif.univ-mrs.fr

Abstract
We introduce DeQue, a lexicon covering French complex prepositions (CPRE) like à partir de (from) and complex conjunctions (CCONJ) like bien que (although). The lexicon includes fine-grained linguistic description based on empirical evidence. We describe the general characteristics of CPRE and CCONJ in French, with special focus on syntactic ambiguity. Then, we list the selection criteria used to build the lexicon and the corpus-based methodology employed to collect entries. Finally, we quantify the ambiguity of each construction by annotating around 100 sentences randomly taken from the FRWaC. In addition to its theoretical value, the resource has many potential practical applications. We intend to employ DeQue for treebank annotation and to train a dependency parser that takes complex constructions into account.

Keywords: Complex prepositions, complex conjunctions, multiword expressions, lexicon, French, dependency parsing.

1. Introduction
Complex prepositions (CPRE) and complex conjunctions (CCONJ) are two types of function words that consist of more than one orthographic word (Piot, 1993). They can be considered as fixed multiword expressions that allow little or no variability. Examples in English include CCONJs even though, as well as and CPREs up to and in front of. Examples in French are shown in Table 1 along with their English (EN) meaningful and literal translations.

CPRE and CCONJ constructions are quite frequent in French. Their linguistic description in the literature is generally limited to building comprehensive lists of such constructions (Sagot, 2010). Most authors assume that these constructions allow no or very little variability (inflection, insertion). Therefore, they would not require a very sophisticated description and representation in machine-readable lexicons and NLP systems, such as the ones required for verbs, for instance (Dubois and Dubois-Charlier, 2004).

An aspect which is often neglected is the segmentation and structural ambiguity that arises when the words composing the complex function word co-occur by pure chance. Consider examples 1 and 2 containing the French CCONJ bien que. It is composed by the words bien (well) and que (that), but when they act as a CCONJ they mean although.

1. Je mange bien que je n’aie pas faim
   (I eat well I am not hungry)
   2. Je pense bien que je n’ai pas faim
   (I think indeed that I am not hungry)

In example 1, bien que is indeed a CCONJ that opposes the main clause (I eat) and the subordinate clause (I am not hungry). In example 2, however, bien que is not a CCONJ and the two words co-occur by chance. The adverb indeed modifies the verb of the main clause think, while the conjunction that introduces the clausal object. Since the word bien is a very common intensifier in French, such accidental co-occurrence cases are likely to occur with all verbs that accept que-clausal complements like think, say and forget.

From an NLP perspective, it is relevant to study these constructions in a parsing pipeline. Most of the time, we would be tempted to simplify the model and treat all of them as multiword tokens or words-with-spaces (Sag et al., 2002). However, accidental co-occurrence, like in example 2, creates ambiguities that are hard to solve at tokenisation time, specially given the simplicity of most automatic tokenisation approaches in French. A simplistic approach such as treating all occurrences of bien que as a single word with spaces inside would introduce an error for sentences like example 2. Conversely, ignoring it in example 1 would mean that both words are treated independently, not capturing the fact that the whole behaves like a conjunction. And what is more, these errors would be propagated to the following processing steps like POS tagging and parsing, certainly generating a wrong analysis.

The creation of DeQue takes place in the context of the development of a statistical dependency parser for French (Nasr et al., 2011). The need to quantify ambiguity has a practical consequence: unambiguous constructions can be included in the lexicon as frozen multiword tokens, while ambiguous ones need to be annotated and dealt with at parsing time.

One way of disambiguating ambiguous multiword units is to keep the tokens as individual lexical units during tokenisation and POS tagging, and then use special syntactic dependencies to indicate the presence of a CPRE or a CCONJ (McDonald et al., 2013; Candito and Constant, 2014; Green et al., 2013). In previous experiments, we demonstrated that this approach is superior to treating all units systematically as words with spaces (Nasr et al., 2015). However, this was only demonstrated for a small set of 8 CCONJs and 4 determiners in French. The present work substantially extends the coverage of the list of potentially ambiguous constructions that can be modelled using that approach.

In the remainder of this paper, we discuss the general properties and syntactic behaviour of prepositions and constructions in French (§ 2.). Then, we present the criteria (§ 3.) and methodology (§ 4.) used to construct the lexicon. Finally, we present the lexicon’s structure and examples (§ 5.). We conclude by listing future extensions planned for this resource (§ 6.).
Table 1: Examples of CPRE and CCONJ in French.

<table>
<thead>
<tr>
<th>Construction</th>
<th>Type</th>
<th>EN meaning</th>
<th>EN literal</th>
</tr>
</thead>
<tbody>
<tr>
<td>à partir de</td>
<td>CPRE</td>
<td>starting from</td>
<td>to leave of</td>
</tr>
<tr>
<td>par rapport à</td>
<td>CPRE</td>
<td>with respect to</td>
<td>for relation to</td>
</tr>
<tr>
<td>bien que</td>
<td>CCONJ</td>
<td>although</td>
<td>well that</td>
</tr>
<tr>
<td>de sorte que</td>
<td>CCONJ</td>
<td>so that</td>
<td>of sort that</td>
</tr>
</tbody>
</table>

2. Prepositions and Conjunctions

Before we can describe the criteria to select CPRE and CCONJ entries for DeQue, we must specify what we consider as simple prepositions (PRE) and conjunctions (CONJ). Indeed, criterion C1.3 below states that CPRE and CCONJ can be replaced by single-word PRE and CONJ. Therefore, we cannot apply it if we do not have a clear definition for these two categories. We distinguish PRE and CONJ according to the criteria below, based on the notion of active and passive valency.

In the framework of dependency syntax, the active valency of a word is defined as its set of acceptable syntactic dependants. For example, nouns can govern determiners, so the active valency of nouns includes determiners. The passive valency is defined as the set of acceptable syntactic governors. For example, adjectives can be governed by nouns, so nouns are in the passive valency of adjectives. Because some complex adverbs behave similarly as complex conjunctions, we also have to define the passive and active valency of adverbs.

Preposition (PRE) Closed-class words (to, for, before) that relate two elements in a sentence, typically introducing verbal or nominal complements as the heads of prepositional phrases.

- **Active valency**: a PRE can govern noun phrases (à la maison, at home), infinitive verbs (sans pleurer, without crying), clauses introduced by conjunctions (pour que je vienne, lit., for that I come), etc. However, they can never govern bare clauses with inflected verbs not introduced by a conjunction (pour que je vienne, *for I come*).
- **Passive valency**: a PRE cannot be the root of a dependency tree, it is necessarily governed by another word. If it is not governed, it is an idiomatic construction: *en avant!* (move forward!), *au secours!* (help!)

Conjunction (CONJ) Closed-class words (that, if, when) that relate two elements in a sentence, typically linking two full clauses.

- **Active valency**: differently from a PRE, a CONJ can govern a bare clause, but it can never govern another phrase introduced by a CONJ.
- **Passive valency**: a CONJ cannot be the root of a dependency tree, it is necessarily governed by another word. If it is not governed, it is an idiomatic construction: *si on allait au cinéma?* (what if we went to the movies?). In other words, conjunctions cannot introduce single clauses, they can only link two clauses.

Adverbs (ADV) Open-class words that generally modify verbs, adjectives or other adverbs.

- **Active/passive valency**: Adverbs induce a special relation between active and passive valency. An ADV cannot govern a CONJ when it is itself governed by another word (*je pense que peut-être qu’il vient* (*I think that perhaps that he will come*). In French, an ADV can govern a CONJ if the ADV is the root of the dependency tree (peut-être qu’elle viendra, lit. *perhaps that she will come*). This distinguishes PRE+ADV constructions (pour que je vienne, *so that I come*) from ADV+PRE constructions (peut-être que, *perhaps that*). When a governed adverb can govern a clause introduced by que (surtout que, alors que, bien que), we consider it as a CCONJ (see examples provided in criterion C1 below).

3. Complex Prepositions and Conjunctions

This paper presents DeQue, a new computational lexicon under development. DeQue lists and models the syntactic behaviour of around 280 CPREs headed by de and CCONJs headed by que in French. The goal of this resource is twofold:

- **Provide a detailed and broad-coverage linguistic description of the possible syntactic analyses of each construction.**
- **Quantify the ambiguity of CPRE and CCONJ constructions based on corpus evidence.**

Constructions in DeQue are CPREs headed by the preposition de (of) and CCONJs headed by the conjunction que (that). These are undoubtedly the most frequent simple prepositions and conjunctions in French. Moreover, they present a very rich co-occurrence pattern, that is, their usages distribution is very heterogeneous.

When used as prepositions and conjunctions, de and que are quite "promiscuous" and combine with many types of modifiers. For instance, the conjunction que can combine with adverbs (bien que, lit. *well that*), prepositional phrases (à condition que, lit. *at condition that*), noun phrases (le temps de, lit. *the time of*), and so on. These modifiers often change or specify the meaning of the relation. For instance, *while que expresses a quite general subordinating relation, bien que expresses opposition, si bien que expresses consequences, and so on.*

One of the challenges in building DeQue was the fact that de and que combine with several complements, including open-class words like nouns, verbs and adverbs. Therefore, it is impossible to guarantee that our lexicon is exhaustive. In addition to that, when we query the corpus for fine POS sequences (see Section 4.), many false positives are returned because of frequent open-class words that accidentally co-occur with de and que.

We define CCONJ and CPRE for inclusion in DeQue based on three criteria. First, they are groups of words that...

---

1The distinction between subordinating and coordinating conjunctions is not relevant for this work.
function as prepositions or conjunctions as a whole. Second, they are potentially ambiguous and contain words that could co-occur by chance. Third, they present some degree of idiomaticity, realised through syntactic and semantic fixedness. Figure 1 summarizes the decision tree used to apply the criteria below in order.

Figure 1: Decision tree corresponding to the application of criteria for lexical entries selection in DeQue.

**C1: Function as PRE/CONJ**

C1.1 A CPRE/CCONJ in DeQue consists of groups of at least two words ending with de/que.

C1.2 A CPRE/CCONJ in DeQue includes at least one open-class (or content) word, that is, one noun, adjective, adverb or verb.

C1.3 A CPRE/CCONJ in DeQue commutes with a similar single-word PRE/CONJ keeping the sentence’s acceptability and similar meaning.

Criterion C1.1 guarantees that the construction is “complex”, meaning that it is composed by more than one token. The last part of the criterion, that is, the fact that the last word is de or que, is only justified because, for the moment, we wanted to limit the scope of DeQue to the most frequent endogenous CPRE and CCONJ. In the future, we intend to extend our lexicon to less frequent function words like CPREs headed by à (to) and CCONJs headed by où (where).

Criterion C1.2 aims at excluding regular syntactic constructions such as simple prepositions followed by que. Most prepositions in French, like pour (for) and après (after), can have their complement introduced by que, which allows using a full clause as the complement of the preposition (see examples 3 and 4). Since this is the case for most prepositions, there is nothing special about the syntactic structure of this construction. Every time it appears, it can be modeled as a preposition that governs a que-clause. Moreover, prepositions always require some postponed complement, and there is no possible accidental cooccurrence here.

Criterion C1.3 helps excluding constructions that look like CPRE and CCONJ but actually are not. For instance, peut-être que (lit. maybe that) looks like a CCONJ where que is modified by the adverb peut-être. One argument against this interpretation is the fact that it can appear in an isolated clause (example 5). That is, it does not respect the passive valency definition for CONJ described in Section 2. Moreover, here the adverb is the syntactic head, inasmuch as que can be omitted (example 6). Many modal adverbs in French exhibit this behaviour; like certainement (certainly), probablement (probably), sans doute (undoubtedly).

(5) Peut-être que je viendrai ce soir

Maybe I will come this evening

(6) Peut-être je viendrai ce soir

Maybe I will come this evening

**C2: Autonomous Lexical Units**

We require that the individual words composing a CPRE/CCONJ are autonomous lexical units. This means that they have their own distribution, cooccurring with other words in other contexts. Criterion C2 aims at excluding constructions that are surely not ambiguous. For instance, parce que (because) contains the word parce, which does never co-occur with a word other than que. This means that there is no possible accidental co-occurrence, and this sequence of tokens is never ambiguous. Tokenization as a word with spaces suffices to represent it in treebanks and parsers. Expressions that pass the tests for C1 and not C2 are not directly discarded, but listed in a separate lexicon of frozen constructions.

**C3: Fixedness**

We keep in DeQue only those constructions that are somehow fixed. We assume that fixedness is a good proxy for semantic idiomaticity, but offers more formal ways of being tested. The traditional definition of idiomaticity is based on semantic non-compositionality. In other words, the meaning of the parts does not add up to the meaning of the whole. Here, it would be hard (if not impossible) to apply this test since most of the time our entries only contain a single content word.

We cite below some fixedness tests applied depending on the POS of the words preceding de and que. The restrictions below are observed with respect to free combinations of each POS forming the unit. We list below some tests used depending on the POS of the open-class word in the construction.

C3.1 If the unit includes a prepositional phrase, changing the preposition, or using the unit without the preposition, entails a change of meaning of the open-class word. For example, while the meaning of the noun centre is unchanged in the sequences au centre de - vers le centre de (in the centre of - toward the centre of), this does not happen for moins (less) in à moins de - pour moins de (unless - for less than).

C3.2 If the unit includes a determiner, no change of determiner is possible without changing the meaning
of the open-class word. For example, en raison de means roughly because, but en la raison de can only literally mean in the reason of.

C3.3 Restrictions are observed on the range of acceptable including relational nouns like south

Criterion C3, and specially C3.1, helps us excluding some type of intensifier, like

CPRE and CCONJ that include nouns and verbs all distinguish qualifying from intensifying modifiers because


The first step in the creation of DeQue was the selection of our target lexical entries. In order to construct this initial lexicon, we design a methodology that combines linguistic expertise and corpora evidence. This methodology helped us to define precise criteria listed in Section 3. for inclusion of an entry in DeQue. Once the list of entries in the lexicon was stabilized, we model ambiguity using a similar process, combining linguistic expertise and corpora evidence.

The corpus used in our queries is the French web-as-corpus (FRWaC), which contains a web dump of 1.613 billion words of French (Baroni et al., 2009). It was chosen mainly for its size, availability and because it presents a fairly decent balance between formal and informal writing. Additionally, it was automatically tagged with parts of speech (POS) using the TreeTagger.

4. Methodology

4.1. Selection of Lexical Entries

The selection of lexical entries to include in DeQue was performed as follows:

1. We list potential de-CPRE and que-CCONJ based on introspection and existing general-purpose lexical resources like LEFFF (Sagot, 2010). For example, this initial list includes candidate conjunctions like si bien que (so that, lit. so well that) and bien sûr que (sure that).

2. For each candidate in this list, we manually annotate the fine POS sequence and global chunk tag of the elements that co-occur with de and que. For instance, si bien que has the fine POS sequence ADV-ADV-que, and the chunk tag GADV-que.\(^3\)

3. We query the FRWaC, retrieving all n-grams that have the fine POS sequences annotated in the previous step, and that occur more than 20 times. For instance, the search for ADV-ADV-que returned new entries like alors même que and si peu que.

4. We select, in this list, additional CPRE and CCONJ entries that we consider relevant according to the criteria described above. Some of the entries that were initially selected in step 1 were removed because they do not respect the inclusion criteria. For instance, bien sûr que was discarded because it does not behave as a conjunction and cannot be replaced by a single-word CONJ, not meeting criterion C1.3.

Some constructions selected as initial candidates turned out to be quite infrequent in the corpus (e.g. au moment que). We decided to keep them in the lexicon because this is due to the nature and quite informal register of the FRWaC. The final list of selected constructions contains 228 CPRE and 49 CCONJ.

4.2. Ambiguity Assessment

For each target construction, we would like to estimate whether it is ambiguous. In that case, we would also like to know what proportion of uses correspond to CPRE and CCONJ readings with respect to accidental cooccurrence. Therefore, we also employ a heterogeneous methodology mixing linguistic expertise and corpus linguistics.

1. We build artificial sentences that exemplify the usage of each lexical entry. We number the examples, 1 for a use as a CPRE/CCONJ and 2 for other uses. For instance, examples 1 and 2 discussed in Section 1. are the sentences that exemplify the usages of the lexical entry bien que.

2. We select sentences in the FRWaC containing the word sequence of the lexical entry. as follows:

(a) We select any sentence in the FRWaC that contains exactly one occurrence of the target construction, including contractions like du (de+le) and qu’ (que+vowel).

(b) We keep only sentences that have more than 10 words (enough context is provided) and less than 20 words (annotation is faster).

\(^3\)For fine POS sequences, we use the POS tagset of the FRWaC corpus. Chunk tags are: adverbial phrase (GADV), prepositional phrase (GPRE), noun phrase (GNOM), subordinate clause phrase (GCSU) and verb phrase (GVRB), suffixed by de or que.
(c) We shuffle the order of sentences to favour variability.
(d) We highlight the target construction to facilitate subsequent annotation.

3. For each sentence, we annotate it as 1 (CPRE/CCONJ) or 2 (other uses). Sentences that have too many orthography and/or grammar errors are discarded. Sentences that are ambiguous and require extra context (previous/next sentence) are discarded as well. We annotate around 100 sentences per construction (or less, according to their frequency in the corpus). Each sentence was annotated by at least 2 experts, and conflicts were resolved during meetings.

4. Based on the insights from annotation, we describe the full syntactic structure of the construction by annotating the full dependency tree of the example sentences of each usage case. This includes comments about the most natural internal structure. For instance, it is reasonable to argue in favor of que acting as a syntactic head and bien being its dependent in bien que (Nasr et al., 2015).

The first step models the ambiguity of each construction in theory. Therefore, it is possible to know whether a construction is potentially ambiguous and requires some special treatment. Steps 2 and 3 quantify this ambiguity in practice, through empirical evidence. For example, in the case of bien que, we have annotated 99 sentences, from which 37.4% are CCONJ uses and 62.6% are other uses. The result of the last step details the syntactic ambiguity and suggests a representation for the target construction in a parser and/or treebank.

5. Resource Structure

The methodology outlined in the previous section results in a modular resource, composed of 4 parts, shown in Figure 2.

Lexicon

The main lexicon contains information about the CPRE or CCONJ entries. Some examples of lexical entries are shown in Tables 2 and 3. In addition to the entry’s canonical form, it provides fine and chunk POS tags. The fine POS sequence corresponds to the POS tags of the individual words used in corpus searches. The tagset comes from the POS tags in the FRWaC corpus, which was automatically tagged by the TreeTagger. The third column shows the number of entries in DeQue that follow each fine POS pattern. We observe that prepositional phrases are the most productive complements of de-CPREs while adverbs seem to be the most common types of complements in que-CCONJs. The chunk POS is useful to group similar patterns and observe paradigms on a coarser scale. In addition to the fields shown in the tables, we also provide corpus-related information such as the number of sentences of the FRWaC that contain the lexical entry’s tokens. This is a raw number, and represents all uses regardless of whether the entry was really used as a CPRE/CCONJ or if it was an accidental co-occurrence.

Ambiguity Examples

We represent the ambiguity of lexical entries in two different ways, which account for the possibility and the likelihood of an entry to be ambiguous. Ambiguity examples are artificial sentences that we build up in order to illustrate all possible uses of an entry. They correspond to prototypical uses of the construction that help annotators understanding the ambiguity. Most of the time, they are adapted from annotated sentences described below.

![Figure 2: Structure of DeQue lexicon and example entry.](image-url)
**Ambiguity Annotation** Some constructions are likely to be employed in different uses, but others have a skewed distribution that makes one of the uses very rare. As explained above, in order to quantify this ambiguity, we selected a set of around 100 sentences per entry which were annotated using a simple distinction: 1 for CPRE/CCONJ and 2 for other uses. The examples below show some real sentences annotated for bien que. We note that “other uses” merges different phenomena. For example, the second sentence below contains the noun bien (goods), which is different from the first sentence where bien has its more usual role as adverb. Both are annotated as 2 because we consider that POS ambiguities will arise and be solved by an exterior process, not using information in DeQue.

1. Bien qu’elle l’aït inspiré, la religion vedique est très différente de l’hindouisme d’aujourd’hui.
2. Il semble bien que la profession préfère temporiser pour rafiler une part des recettes provenant d’internet.
3. Je veux toutefois être encore de vos amis ; mais ne demandez plus un bien que j’ai promis.
4. La pièce n’a besoin de rien, bien qu’il n’y ait rien là.
5. Je sais bien que votre cœur ne se détachera pas de lui-même.

**Syntactic Description** We propose a full dependency tree for the ambiguity examples. This provides a way to distinguish CPRE and CCONJ constructions from accidental cooccurrence using special relation MORPH between the constituting elements (Nasr et al., 2015). When interpreted as a CPRE or CCONJ, the whole MWE acts as a single preposition or conjunction. Therefore, we argue that de and que should be the syntactic heads. Modifiers like noun phrases and adverbs often have regular syntactic structure, and their heads are governed by the preposition or conjunction. For instance, the syntactic structure of expressions bien que (although) and à condition de (conditioned to) in reading 1 is shown below:

![Diagram of the syntactic structure of bien que and à condition de]

6. **Future Developments**

In the future, we would like to extend this lexicon to other CPRE and CCONJ constructions. This includes, for instance, CCONJ headed by si (if), quand (when) and où (where), CPRE headed by à (to) and en (in) and also complex determiners and pronouns.

This lexicon can be very useful for parser development and adaptation to a given domain. For instance, if we want to build a very robust parser for literary texts, we would need to model all theoretically ambiguous constructions using MORPH links. On the other hand, a fast parser for speech transcriptions could safely ignore constructions that rarely co-occur by accident, setting a threshold on the proportion. For instance, all combinations that occur 90% of the times as complex function words will be simply concatenated as a single token. This helps the parser designers to make more informed decisions about the best moment to deal with these complex function words in the analysis pipeline.

We would like to build more fine-grained syntactic-semantic clusters for each construction type, depending on the distribution and fixedness of internal elements. For instance, locative relational nouns like north, centre, etc can build relational CPREs like au nord de (north of). They accept similar modifications and are very different from causative CPREs like en raison de and à cause de (both mean roughly because). We would like to obtain this information by studying the link between the corpus distribution and cooccurrence pattern of the internal elements and of the whole expression. This can be related either to the linguistic context of the construction itself, but also to the usage context, that is, genre and domain of the text.

7. **Bibliographical References**


