How to measure the meanings of words ? Amour in Corneille's work

Cyril Labbé, Dominique Labbé

HOW TO MEASURE THE MEANINGS OF WORDS?

AMOUR IN CORNEILLE'S WORK

CYRIL LABBÉ
Université Grenoble I
Cyril.labbe@imag.fr

DOMINIQUE LABBE
Université Grenoble II
dominique.labbe@iep.upmf-grenoble.fr

Abstract

We present a new method to describe the contextual meaning of a key word in a corpus. The vocabulary of the sentences containing this word is compared to that of the entire corpus in order to highlight the words which are significantly overutilized in the neighbourhood of this key word (they are associated in the author's mind) and the ones which are significantly underutilized (they are mutually exclusive). This method provides an interesting tool for lexicography and literary studies as is shown by applying it to the word *amour* (love) in the work of Pierre Corneille, the most famous French playwright of the 17th century.

Draft of the paper published in :
Langage Resource Evaluation, 2005, n° 39: p. 335-351.

1. Introduction

It is often repeated that, except for "onomatopoeia", words are "empty forms" and that their meanings are not "intrinsic" to them but that these meanings stem from external sources. When one is asked for the meaning of a word in a work or from a certain era— say for example, *amour* (love), which is the most frequent substantive in French literature until the end of the 19th century (Brunet, 1981), especially in 17th century theatre — one must consider three main sources. First, the language — in this case 17th century literary French — gives the "general meaning" of this word. Then this must be compared with the particular uses made by the author under consideration, that is to say: the words he uses when he writes about "love" and the ones he avoids in such circumstances. We propose to name this constellation of associations and repulsions: "specific personal meaning". And third, the researcher must also consider the era, the history of the author's country, his social class and certain more precise questions such as: what were the main events of his life? and how could these events have influenced his work? what company did he keep? what school of thought he could have represented? And so on.

Lexicographers put a stress on the first level, literary critics consider the second one and sociologists or historians, the third one… In any case, the result might be an "impressionistic" patchwork of intuitions aided by quotations more or less arbitrarily selected. To avoid this danger, a set of procedures is proposed below which operates with more objectivity and which may help critics or scholars in their studies. This paper discusses only the two first levels (general meaning in the language of the period, specific meaning in the personal vocabulary of an author), although statistics can also help research on the third level.

Corpus processing is a major trend in computational linguistics, terminology extraction or information retrieval (Grefenstette,1994; Habert & Al, 1997 & 1998). The main concept is the "lexical framework" used by lexicographers when grouping certain words into the same semantic class, assuming that these words are linked together in a structured conceptual location. The corpora make it possible to find meanings of the words through actual contexts and, more precisely, through "concordancers" which indicates the "collocations", ie: the conccurrences, within a limited span, of two or more words (Sinclair, 1995). These methods are very useful for specialised lexicography — especially when applied to scientific lexicons — and for terminologists (Bergenholtz, 1995). But they are less useful in other fields such as literature or less-structured discourses (politics, sports…), because the words are polysemic and very mobile in the sentences (eg. Favre & Al, 1997). Moreover these lists cannot highlight the words which should be present - in regard to their frequencies in the

whole corpus - and are actually absent or not enough associated with the key-word under study.

Pierre Hubert and Dominique Labbé presented some ways to overcome these limits (Hubert & Labbé, 1995). This method ("lexical universe of a word") is fully presented here and, as an example, is applied to the 34 plays of Pierre Corneille (1606-1684), the famous French author of the 17th century (titles and dates of the plays in Appendix 1). The entire corpus length is greater than a half million tokens. In the 19th century, the spellings of words have been standardised in modern French by Charles Marty-Laveaux and published in the collection "Les grands écrivains de la France" (Hachette).

2. First step: replace words in the language system

How can we determine the meaning of a word in 17th literary French? The answer usually lies in consulting a dictionary of that period. By chance, the first French dictionaries appeared at the end of this same century. For example, in appendix 2 one can read the definitions of *amour* given in the well-known *Dictionnaire universel* by Antoine Furetière (1690).

Does Furetière give the exact meanings? The answer depends on which level of the language system is considered. In this document, the main relevant pieces of information are that:

— in the 17th century French, "love" is a masculine or a feminine substantive (in modern French, the feminine is no longer used);

— *amour* is a variable word — ie an "s" is added at the end of the word when plural — and the meaning of this plural is slightly different from the meaning of the singular;

—*amour* has several meanings, as is the case for most of the common words in any language.

Finally, Furetière gives the exact answer at an important level as to which is the part of speech (grammatical category) of the word (substantive). In fact, for nearly four centuries, the parts of speech of modern French language remained relatively stable, with only a few changes of minor importance (like the disappearance of the feminine of *amour*). So it is possible to apply this "nomenclature" (appendix 4) to every electronic version of a work.

Going further into detail, is Furetière a credible witness?

Answering this question requires the preliminary composition of a "representative corpus" containing a large number of excerpts of works from the 17th century — in other words, something built like the British National Corpus for the French language — or, more modestly, limiting the question to a small number of authors for whom some of their works

are available electronically. For the French 17th century, it is the case for the entire works of Pierre Corneille (1606-1684), Jean-Baptiste Poquelin Molière (1622-1673) and Jean Racine (1639-1699), who are undisputedly the most famous playwrights of this century. Then one can examine the vocabulary associated with the word under consideration with the help of concordance lists (for an example, see Appendix 3). But, for French texts, these lists are of little help because, in this language, a large number of words, especially the most frequent ones, have multiple ways of writing them. For example, Table 1 below shows the immediate neighbourhood of "love" in Corneille's entire work.

Table 1. Immediate neighbourhood of *amour* in Corneille's entire work.
(within parenthesis: the absolute frequencies)

Word before :		**Word after** :
l' (the) (686)	***amour*** (love)	*de* (of) (59)
d' (of) (216)	(1,888)	*qui* (which/that) (48)
mon (my) (200)		*est* (is) (40)
votre (your) (110)		*pour* (for) (40)
cet (this) (82)		*a* (has)…

The first letter of *amour* is a vowel, thus the article *le* (the) and the preposition *de* (of) are elided before it, and a "t" is added at the end of the demonstrative *ce* (this)… In addition, it must be noticed that, in French:

— the article *le* (the) has two other inflections: *la*, *les*;

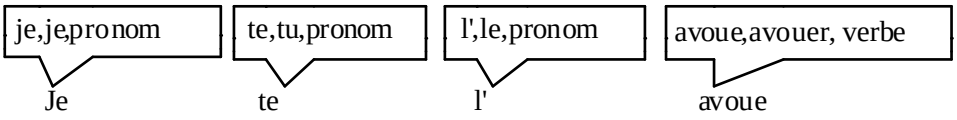
— preposition *de* is written "des" when the noun following is plural (despite the fact that, in French, prepositions are theoretically invariable);

— *mon* (my) has two other inflections : *ma*, *mes*;

— the verb *être* ("to be" is the most used verb in French as in English) can be "conjugated" in more than thirty different ways, etc.

This extreme diversity seems to prevent any answer to very simple questions such as: "when he wrote about love, was Corneille overusing or under-using the words "the" or "my", the preposition "of", or the verb "to be"…? The only way to answer, is to replace all these inflections of a single word with a only onr spelling convention: infinitive of verbs, singular masculine of adjectives, etc. In order to "lemmatize" French texts, we have developed various semi-automatic tools. For example, Table 2 gives the first words of Mélite (the first play of Corneille): *Je te l'avoue* ("I confess it to you"). Every word is given a tag by the software (Table 2).

Table 2. Lemmatization of French texts. An example.



The first word *Je* — which can be referred to as "raw token" — has a capital letter (as usual at the beginning of a sentence or at the beginning of a verse). In the tag attached to this word, the first *je*, is the "standardised spelling", the second one is the "dictionary entry" (in French: *lemme*) and *pronom* ("pronoun") is the "part of speech".

In French, many words are ambiguous. In our example, it is the case for *le* which can be an article or a pronoun… In any French text, an average of more than one third of all the words used are "homographs" (one spelling, several dictionary entries). For example *sommes* can be a plural masculine substantive ("snooze"), a plural feminine substantive ("sums of money"), or the verb *être* in the first person plural ("we are")… Thus, standardisation of spelling and word tagging are the first steps for any high level research in quantitative linguistics with French texts (norms and software are described in Labbé, 1990; for a discussion: Pincemin, 2004). All the calculations performed in this paper are made on these "types". In French, the word *vocable* usually refers to this smallest part of the vocabulary, that is to say the association of a dictionary entry with a part of speech (e g : *le* + *pronom*).

Moreover, tagging, through the grouping of large numbers of tokens under the headings of less numerous types, carries many other advantages, especially a great reduction of different units to be counted.

One can compare this operation with the calibration of sensors in any experimental science.

These operations allow one to establish the lexical "universe of a word" and to calculate its "personal" meaning (in our example, the meaning of "love" in the fictional universe created by Pierre Corneille).

3. Second step: determination of the lexical universe of a key word.

The algorithm isolates all the sentences containing the key word under study. The set of these sentences is the *lexical universe of the word*. For example, in Corneille's work, the lexical universe of "love" contains: 1,822 sentences and nearly 60,000 words.

To avoid giving excessive weight in the calculation to some very long sentences, the following procedure is used. First, the software calculates the lengths (in tokens) of all sentences and the standard deviation of these lengths around the mean. Second, it truncates the sentences the lengths of which exceed this mean by two standard deviations (before or after the key word according to its position in this sentence). Other solutions are possible, like choosing a certain span around the key word without regard to punctuation.

Then it is possible to answer certain questions precisely: which words did Corneille associate with "love" and which ones did he exclude as if they were "counter-terms" of this word type?

To answer these questions, let:

— U be the lexical universe (the set of the tokens within sentences containing the key word) and C be the set of all the tokens of the theatrical work of Corneille;

— N_c be the length of the entire corpus (in tokens). The length of the Corneille's theatrical work is: 555,200 tokens;

— N_u represents the length (in tokens) of the lexical universe of the word under consideration. In Corneille's work, the lexical universe of *amour* (the most frequent substantive) occupies 10.8% of N_c. It is the largest lexical universe in this work (if the most frequent articles, adverbs and pronouns are excluded);

— F_{ic} and F_{iu} represent the absolute frequencies of the word type i in the whole corpus (C) and in U.

What is the number of i most likely to occur in a random sample of size N_u tokens drawn out of C? This expected value — or "mathematical expectancy" ($E_{i(u)}$) — can be calculated easily:

$$(1)\quad E_{i(u)} = F_{ic} * \frac{N_u}{N_c}$$

It should be noted that, when a token is drawn randomly out of the vase C, it is not replaced in it, since this is the only way to be sure that the number of a type i, in a random sample of N_u tokens extracted out of C, will be always less than or equal to its frequency in C ($E_{i(u)} \leq F_{ic}$). Considering that, even in very large corpora, the low frequency words are

numerous, this is an important precaution. Thus, this experiment must follow a hypergeometric distribution, not a binomial one.

Then this expected value must be compared with the observed one (F_{iu}). Of course, if they are equal ($E_{i(u)} = F_{ic}$), one may conclude that the word is "neutral" (the same law of distribution operates in the entire corpus and in the specific lexical universe of the word under consideration). But when the two values differ, how is it possible to measure whether the word is *significantly* over- (or under-) used? To answer this question, one must consider the probability of the observed value (F_{iu}) resulting from the combination of two events:

— the number of different possibilities in choosing N_u tokens within N_c ones:

$$C_c^u = \frac{N_c!}{N_u!(N_c - N_u)!} = \begin{bmatrix} N_c \\ N_u \end{bmatrix}$$

— the number of different possibilities in choosing F_{iu} tokens within F_{ic} ones:

$$C_{F_{ic}}^{F_{iu}} = \frac{F_{ic}!}{F_{iu}!\,(F_{ic} - F_{iu})!} = \begin{bmatrix} F_{ic} \\ F_{iu} \end{bmatrix}$$

The joint probability of these two events follows a hypergeometric law, the parameters of which are: F_{ic}, F_{iu}, N_u, N_c:

$$(2)\quad P(X = F_{iu}) = \frac{\begin{bmatrix} F_{ic} \\ F_{iu} \end{bmatrix} \begin{bmatrix} N_c - F_{ic} \\ N_u - F_{iu} \end{bmatrix}}{\begin{bmatrix} N_c \\ N_u \end{bmatrix}}$$

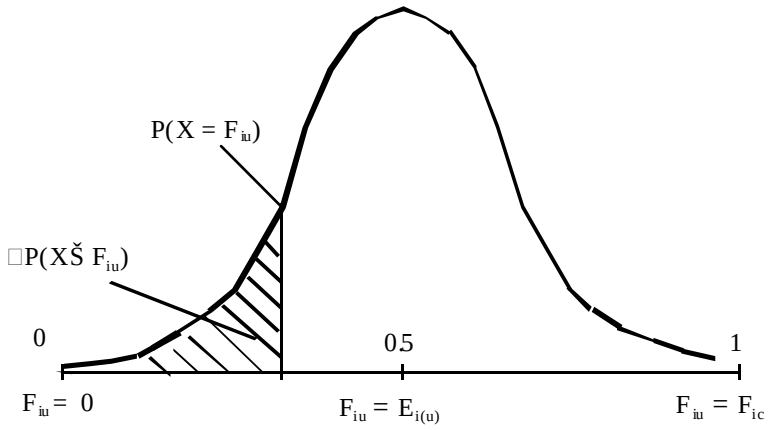
All of the following calculations are made with the use of formula (2). Under conditions where N_c, F_{ic} and N_u are large enough, the values of X are distributed along the interval [0-F_{ic}], following the shape of a curve, the mode of which is reached when $F_{iu} = E_{i(u)}$ (Fig. 1 below).

4. Calculation of the strength of the link between two word types

Here, we propose considering not only a point on the curve but also the surface under it and, more precisely, the section obtained by summing the results of formula (2), X varying,

one by one — as actual absolute frequencies are always integers —, from zero to F_{iu} (Figure 1 below).

Fig. 1. Theoretical distribution of the results of Formula (3)



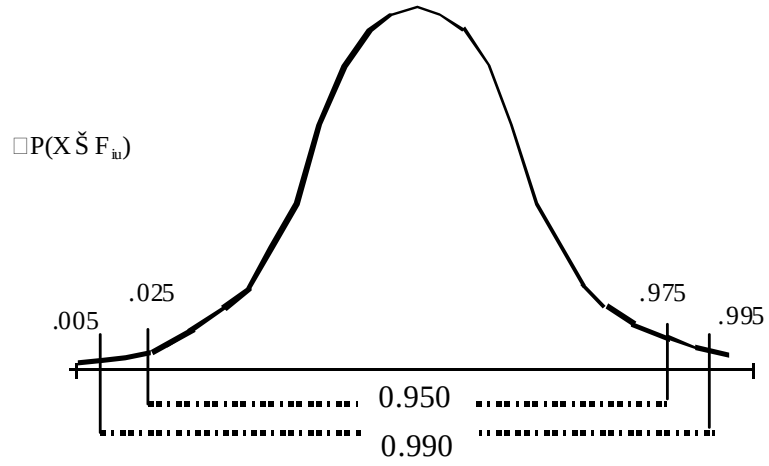
Let L_{iu} be the link between the type i and the lexical universe of the key word under consideration (U):

$$(3)\quad L_{iu} = P(X \leq F_{iu}) = \sum_{j=0}^{j=Fiu} P(X = j)$$

L_{iu} measures the orientation and the strength of the link between the type i and the key word of U. If L_{iu} is very high, it can be assumed that, in the author's vocabulary, the two words have a mutual attraction; and a mutual repulsion when L_{iu} is very low. In the first case, the two words are associated in the author's mind; in the second case, when one is used, the other is rather excluded.

More precisely, one can use two intervals of confidence and two pairs of limits (Figure 2 below).

Figure 2. Two delimited prediction intervals for the results of formula (4).



— If a high confidence level is required, the hypothesis that the same law of distribution governs the occurrences of the type *i* within *C* and within *U* can be rejected with less than a 1% chance of error, with L_{iu} being under .005 or above .995. Under the low boundary, the two words are mutually exclusive (negative link); above the high boundary, they are mutually associate (positive link);

— If a lower confidence level is accepted (5%), the lower limit is: .025 (negative link); and the upper one is: .975 (positive link).

When studying large corpora and very frequent words — like *amour* in Corneille's entire work — it is better to choose the higher level of confidence in order to limit the lists of words and to highlight the most specific ones. It is the approach adopted here (in Appendix 4, words are ranked by decreasing strength of L_{iu}). On the other hand, when the corpus is less extensive or the word under study is relatively rare, using the lower interval gives more consistency to the lexical universe, but the words at the extremes of the lists must be considered with caution.

5. Remarks

This calculation has some drawbacks and its results must be considered occasionally with caution.

First, in light of the fact that the absolute frequency of a type *i* in *U* (the lexical universe under consideration) varies between 0 — no occurrence of this word type *i* in *U* — and F_{ic} (all the occurrences of this word type *i* appear in *U*):

$$(4)\ 0 \leq F_{iu} \leq F_{ic}$$

this interval means that formulae (2) and (3) have two evident restrictions:

— $F_{ic} < N_u$: if the type *i* has, in the entire corpus, an absolute frequency greater or equal to the size of the lexical universe, it is impossible to encounter all its occurrences in the lexical universe (upper boundary of the interval (4)). This first restriction means that, when applied to the lexical universe of a rare word type, the calculation (3) must not be performed on the very frequent word types (the "function words" of which the frequencies are often greater than 1%)…

— $F_{iu} < (N_c - N_u)$: if not, it is impossible not to encounter the word type *i* in the lexical universe. This restriction is easy to understand because, when one uses "probabilistic" reasoning, the corpus (ie: the "vase") must be sufficiently large compared to the lexical universe (ie: the "sample") under consideration. This condition means that the calculation cannot be applied to the very frequent words referred to above because they are used everywhere (common articles, prepositions, pronouns…). This second restriction is also important when the calculation is used to compare two corpus. The reference corpus (*C*) must be much larger than the one under study (*U*) which must not be too short either (at least several thousand tokens).

In other words, large corpora are needed! and formula (3) cannot be applied to the "function words", and these word types must be excluded from the calculation when studying the lexical universe of a rare word.

The effect of absolute frequencies on Formula (3) poses a second problem. Theoretically, results of (3) are independent of the absolute frequency of the type *i*. In practice, this is not the case: the most frequent types appear unusually numerous in the lists like the one in Appendix 4 (Labbé & Labbé, 1997). In fact, formula (2) postulates that any word can be associated in a sentence with all the others. Concerning low frequencies, the theoretical number of combinations (calculated by means of the hypergeometric law) is not very different from the number of actual combinations possible within the usual grammatical and syntactic rules. This is not the case for the high frequency words: the theoretical number of combinations into which they can enter with the other words of the lexical universe is nearly infinite, but most of these combinations are impossible because of the grammar and the syntax. In other words, the weight of the constraints stemming from the use of "natural" languages, combined with the author's project, bears more heavily on word types of high frequency, especially if they are

"function words" (Labbé & Labbé, 1997). As a result, it can be assumed that the more two words are frequent in a work, the more they are likely to be associated in a lexical universe. Therefore, it is necessary to consider frequencies of word types (not absolute frequencies but relative ones): when the same index is obtained from two word types, the less frequent of the two can be considered as the more "specific".

For example, in Corneille's work, *cœur* (heart) and *haine* (hatred) are the two types most closely linked to *amour*. In the lexical universe of "love", the first type occurs 262 times (F_{iu}) — although it is expected only 185 times (E_{iu}) — and the second type 91 times (against 45 expected). Although the association of "heart" and "love" is more frequent, the link between "hatred" and "love" is clearly more significant (see below).

This last precaution is especially important when certain particular parts of speech are considered.

6. The influence of the parts of speech.

The grammatical category (part of speech) of the type under consideration may influence the content of the lexical universe. For example, one has more chance of meeting verbs when studying a personal pronoun than in the proximity of a substantive, or of meeting more adjectives when studying a substantive than surrounding a personal pronoun… If formula (2) and (3) consider only the frequencies within the whole corpus, the lexical universe of a pronoun will likely contain an excess of verbs (for the positive links) and a deficit of substantives (for the negative association), etc. This tendency is confirmed even in the case of very large lexical universes.

Let *A* be the part of speech of the type *i* ; F_{ac} and F_{au} being the sum of the frequencies of all the types belonging to this part of speech in *C* (the entire corpus) and in *U* (the universe of the key word *i* under study). The "densities" of *A* in *C* and in *U* are:

$$Dens_{ac} = \frac{F_{ac}}{N_c} \; ; \; Dens_{au} = \frac{F_{au}}{N_u}$$

As an example, Table 3 below shows a comparison between the densities of the main parts of speech within Corneille's entire work and within the sentences containing *amour* in this work. The first row in this table indicates that, in the whole corpus, there is an average of 193 verbs out of every 1,000 tokens and that, in the lexical universe of "love", this density is only

180 verbs out of 1,000 tokens — a deficit of 6.6% — that is to say that, when Corneille wrote about "love", he used significantly fewer verbs…

Table 3. Densities of parts of speech in Corneille's entire work and within the lexical universe of *amour* in this work.

Parts of speech	C $Dens_{ac}$ (‰)	U $Dens_{au}$ (‰)	(U-C)/C (%)
Verbs	193.1	180.3	-6.6
Inflected forms	*128.5*	*120.7*	*-6.1*
Past participles	*16.5*	*14.2*	*-14.0*
Present participles	*2.8*	*3.4*	*+25.0*
Infinitives	*45.2*	*41.9*	*-7.3*
Proper nouns	13.3	8.9	-33.1
Substantives	158.0	172.1	+8.9
Adjectives	44.7	47.0	+5.0
Past participle adj.	*6.2*	*6.8*	*+9.7*
Pronouns	185.3	165.1	-10.9
Personal pronouns	*115.6*	*102.8*	*-11.1*
Determinants	140.9	155.3	+10.2
Articles	*69.4*	*79.7*	*+14.7*
Numbers	*3.7*	*3.2*	*-13.5*
Possessive adj.	*44.3*	*49.4*	*+11.5*
Demonstrative adj.	*12.2*	*12.0*	*-1.7*
Indefinite adj.	*11.3*	*11.0*	*-2.7*
Adverbs	76.3	74.2	-2.8
Prepositions	116.4	122.4	+5.2
Conjunctions	70.8	74.1	+4.7
Foreign words	1.2	0.6	-47.4

Except for adverbs, all deviations mentioned in Table 3, are significant: with less than a 1% chance of error, one can consider that the over or under-uses, cited in the last column of the table 3, cannot occur by chance. The verbal group (verbs + personal pronouns + adverbs) is significantly under-utilised; the nominal group (substantives + adjectives + determinants + prepositions) is over-utilised. Linguistically, some details are very interesting. For example, it is well known that, in French, present participles of the verbs (in English: -ing) share many characteristics of the nominal group: in the lexical universe of "love", dominated by the

nominal group, the density of present participles is greater than expected. The behaviour of personal pronouns is also interesting: they not only follow the pattern of verbs but they always amplify it. On the other hand, in Table 3, some effects do not stem from language but from Corneille's mental universe. For example, the huge deficit of proper nouns. It would be expected that "love" is associated with the names of beloved persons (or the hated ones for the negative links); on the contrary, "love" is relatively reticent. In Corneille's plays, when someone is talking about "his love (or loving)" he rarely names the person she (or he) loves, and he directly addresses this beloved person very rarely…

Thus, in order to give each token the same fair chance of being present in *U*, calculations must take into account the specific weight of the different parts of speech in this lexical universe. Let α be the relation between the two densities defined above:

$$\alpha = \frac{Dens_{ac}}{Dens_{au}}$$

Formula (3) becomes:

$$(3)\quad L_{iu} = P(X \leq \alpha F_{iu}) = \sum_{j=0}^{j=\alpha F iu} P(X = j)$$

As an example, let us consider in detail the results of this calculation applied to the word "love" in the work of Pierre Corneille.

7. *Amour* (love) in a major French dramatic work of the 17th century

The complete results can be read in Appendix 4. In this appendix, the word types associated with (or rejected by) *amour* are classified according to their parts of speech and ranked by decreasing values of L_{iu}. For example, the substantives which have the strongest positive links with "love" are: *haine* ("hatred"), *cœur* ("heart"), and *feu* ("fire"); the verbs are: *céder* ("to give up") and *éteindre* ("to die", "to extinguish") ; adjectives show that in Corneille's mind, love is mainly *conjugal*, *parfait* ("perfect") and *paternel* ("fatherly")… These are the major associated meanings of "love" in Corneille's mind.

As a preliminary remark, one must note that some of these links are not only semantic. For example, though *cour* ("court") or *retour* ("coming back") are associated with *amour* for

obvious semantic reasons, it is not the case with other types like: *tour* (turn), *jour* (day) or *séjour* (stay). If the computation shows that these three words are strongly linked to *amour*, it is because, in French, there are few words ending in "our" that provide a rhyme with *amour* when this word is placed at the end of a verse. Of course, versification constraints are not taken into consideration in the calculation!

In Appendix 4, the list of positive associations confirms many points of Furetière's definition. For example, it is not surprising to find *Vénus* or *Psyché* as the proper names most associated with *amour*. They are well-known clichés. The complete list in the appendix seems to suggest a conformist vision of love. The usual links revolve around certain metaphors which are also in Furetière's definition. For example, associations of the word types "heart" or "fire" with love are stereotypes in the 17th century theatre (also in Racine and Molière). In Corneille's work, "fire" is particularly rich in associations (Table 4).

Table 4 Corneille's lexical universe of love is organised around a main theme: *feu* ("fire").

verbs: *éteindre* ("to die out", "to extinguish"), *allumer* ("to light"), *éclater* ("to break out"), *étouffer* ("to suffocate" or "to put out"), *rallumer* ("to relight"), *brûler* ("to burn")…

substantives: *feu* ("fire"), *amorce* ("fuse"), *ardeur* ("burning"), *froideur* ("coldness")…

adjectives : *éteint* ("died away"), *puissant* ("powerful")…

Some minor metaphors are also very conventional: "knots" (*nœuds*) and "links" (*liens*) associated with *hymen*; "tenderness" (*tendre, tendresse*), "gentleness" (*douceur*) of love, etc. The computation also demonstrates that Corneille's lexicon is very consistent. For example, one "always" loves (*toujours*, positive link in the list of the adverbs) but "tomorrow" is not considered (*demain*, abverb, negative link). As is shown in the lists of determinative words, "love" is always *le premier* ("the first one") and never the *second*, etc. All these clichés are predictable with the help of traditional tools like concordances or lists of collocations. The computation only gives such answers in a surer and quicker way.

It will be noticed that this list also presents many contradictions to common sense, and that these contradictions may not be apparent when using concordance or lists of collocations. For example, the extreme overuse of the third person (she, he, it) and the lack of the second person: *tu* and *vous* ("you") which are two of the major underused word types in the lexical universe of "love", even if the contrary might be expected. In fact, in Corneille's drama, the

word *amour* is rarely used face-to-face between concerned characters. It usually occurs in dialogues between one of two supposed lovers with a third person (generally the "confidant"). Of course, it suggests a pessimistic vision: for Corneille, love is rarely mutual.

In detail, the lists of Appendix 4 show that the deficits (or repulsions) are also very indicative, and that these under-utilizations are not apparent with the help of dictionaries, concordances or lists of collocations (how does one compute collocations which *must* be present and are actually absent or significantly under-used?) For example, in the list of adjectives, because of the presence, near the head of the list, of *conjugal* and "fatherly", one also expects to find: *maternel* ("maternal") and *filial*. It is not the case. Husbands love their wives and their children, but the opposite is not true: the lexical universe of "wife" and "children" shows that they *respect* their husbands and/or fathers, but they rarely "love" them!

The lists of negative links indicate what are, in Corneille's mind, the obstacles in the way of love: "god(s)" and "heaven" (*dieu, ciel*) and powerful people like "king(s)" (*roi*), "prince(s)" and more widely, *les gens* ("people") and "fate": *mort* ("death"), *combat* ("struggle"), *sort* ("fate"), *malheur* ("adversity"), *guerre* ("war")…

In any case, the major conclusion is the strong association, in Corneille's mind, of two word types: "hate" (*haine*) and "love". Hundreds of theses, memoirs, books and articles have been written on Corneille's work and nearly all this research indicates the importance of *amour* (eg. Nadal, 1948), but no one has previously underlined this peculiarity. The latter is not very surprising if one thinks that the word type "hate" is quite rare and that, in some sentences containing "love", it is sometimes separated from it by a relatively large span of words…

Of course, lexicographers also need typical sentences, like the quotations given in a dictionary in order to illustrate the different meanings of a word. Thus, when the specific vocabulary is computed, the algorithm re-reads all the work and evaluates each sentence. When this sentence contains a word type which is overused with *amour*, its score is incremented by one, and, when an underused word type is read, the score is decreased by one. Table 5 below presents the sentence which contains the highest number of types attracted by "love" and the fewest number of types repulsed by it. One can see that within a few verses, the main themes: *cœur* ("heart"), *flamme* ("fire"), opposition of *passion* to *devoir* ("duty") or to "honour", and, overall, the opposition of "love" to "hate" (*ressentiment, colère, haïr*). The algorithm has "rediscovered" the most famous passage ("Chimène's stanza") of the most famous of Corneille's plays (*le Cid*)… As can be seen, statistics sometimes lead to an agreement with literary common sense!

Table 5. Corneille's most characteristic sentence dealing with "love" :

CHIMENE.
C'est peu de dire aimer, Elvire : je l'adore ;
Ma passion s'oppose à mon ressentiment ;
Dedans mon ennemi je trouve mon amant ;
Et je sens qu'en dépit de toute ma colère,
Rodrigue dans mon coeur combat encor mon père :
Il l'attaque, il le presse, il cède, il se défend,
Tantôt fort, tantôt foible, et tantôt triomphant ;
Mais en ce dur combat de colère et de flamme,
Il déchire mon coeur sans partager mon âme ;
Et quoi que mon amour ait sur moi de pouvoir,
Je ne consulte point pour suivre mon devoir :
Je cours sans balancer où mon honneur m'oblige.
Rodrigue m'est bien cher, son intérêt m'afflige ;
Mon coeur prend son parti ; mais malgré son effort,
Je sais ce que je suis, et que mon père est mort.
(*Le Cid*, Act III, scene 3, verses 810-824)

One can find an echo of this dilemma in the following scene during which the two lovers briefly encounter each other:

DON RODRIGUE.
Ton malheureux amant aura bien moins de peine
A mourir par ta main qu'à vivre avec ta haine.
CHIMENE.
Va, je ne te hais point.
DON RODRIGUE.
Tu le dois.
CHIMENE.
Je ne puis.
(*Le Cid*, Act III, scene 4, verses 961-963)

8. Conclusion

One must bear in mind that, since it is impossible to measure a phenomenon whose observations are made without precision, all these calculations require careful standardisation of word spelling and, for French, tagging ("lemmatisation") of each token in the texts.

These calculations enable quick and simple explorations of large series of texts, such as literary corpora, and they can provide much more information in a comparison of two authors. For example, what is the meaning of *France* in General de Gaulle's speeches and in those of François Mitterrand? (Labbé, 1998). In this case, formulae (2) and (3) must be modified to take into account the fact that the sizes of the two corpora to be compared are very similar.

The same computations can be applied to another interesting problem: comparison between different parts of a corpus. For example, which is the specific vocabulary of one of Corneille's plays compared to his entire work or to a specific part of it? Or what are the distinguishing peculiarities of Corneille compared to the other French 17th century writers? In this case, the size of the corpora to be compared must not be too small…

The same reasoning can also be applied to the relations of synonymy, hyponymy and antonymy between different sets of words, leading to a semiautomatic lexicography (Leselbaum, Labbé, 2002).

At the very least, with the help of these methods, it seems possible to answer some of the interesting questions posed by H. Craig concerning authorship attribution studies: " If you can tell authors apart, have you learned anything about them?" (Craig, 1999).

ACKNOWLEDGMENTS

The authors are grateful to Pierre Hubert (Ecole des Mines de Paris) — who helped to carry out the first experiments in the 1990s — to Charles Bernet (Institut National de la Langue Française) who provided the texts of Corneille and to Tom Merriam for his accurate reading of our first translation and for his most helpful comments and advice.

BIBLIOGRAPHY

Electronic versions of the entire dramatic works of Pierre Corneille, Jean-Baptiste Poquelin Molière and Jean Racine are available on the web site of the Bibliothèque Nationale de France (gallica.fr). We have sub-edited, standardised and tagged all these plays (This version is available on: http://ota.ahds.ac.uk/2466).

Bergenholtz Sven Tarp (ed) (1995). Manuel of Specialised Lexicography. Amsterdam-Philadelphia: John Benjamin.

Brunet Etienne (1981). Le vocabulaire français de 1789 à nos jours. Paris-Genève: Slatkine-Champion.

Craig Hugh (1999). "Authorial Attribution and Computational Stylistics: If You Can Tell Authors apart, Have You Learned anything about them?". Literary and Linguistic Computing. 14-1: 103-113.

Fabre Cécile, Habert Benoît & Labbé Dominique (1997). La polysémie dans la langue générale et les discours spécialisés. Sémiotiques. 13: décembre 1997, p 15-30.

Furetière Antoine (1690). Dictionnaire universel. Rotterdam: Augsgaden Den Haag.

Greffenstette (1994). Explorations in Automatic Thesaurus Discovery. Dordrecht: Kluwer.

Hubert Pierre, Labbé Dominique (1995). "La structure du vocabulaire du général de Gaulle" in Bolasco Sergio, Lebart Ludovic & Salem André (eds), III Giornate internazionali di Analisi Statistica dei Dati Testuali. Rome: Centro d'Informazione e stampa Universitaria. Vol II: 165-176.

Labbé Cyril, Labbé Dominique (1997). Que mesure la spécificité du vocabulaire ?. Grenoble: CERAT. December 1994 & june 1997. Published in Lexicometrica. 3-2001. On line:

Labbé Dominique (1990). Normes de saisie et de dépouillement des textes politiques. Grenoble: Cahier du CERAT. On line:

Labbé Dominique (1998). "La France chez de Gaulle et Mitterrand" in Fiala Pierre & Lafon Pierre (dir). Des mots en liberté. Mélanges Maurice Tournier. Fontenay-aux-Roses: ENS Editions. 183-193.

Leselbaum Jean, Labbé Dominique (2002). "Lexicographie assistée par ordinateur. Signification de “Banque” dans le vocabulaire économique" in Morin Annie & Sébillot Pascale (eds). VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002). Rennes: IRISA-INRIA.Vol. 2: 447-456.

Nadal Octave (1948). Le sentiment de l'amour dans l'œuvre de Pierre Corneille. Paris: Gallimard.

Pincemin Bénédicte (2004). Lexicométrie sur corpus étiquetés. In Purnelle Gérald, Fairon Cédrick & Dister Anne (Eds). Le poids des mots. Louvain: Presses Universitaires de Louvain, p 865-873.

Sinclair John M. (1995). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Appendix 1. Corneille's dramatic work

(All Corneille's plays are in verse)

		Year of first presentation	Genre	Length (in tokens)
1	Mélite	1630 ?	Comédie	16 690
2	Clitandre	1631	Tragi-comédie	14 402
3	La Veuve	1631	Comédie	17 661
4	La Galerie du Palais	1632	Comédie	16 140
5	La Suivante	1633	Comédie	15 160
6	Comédie des Tuileries (one act)	1634	Comédie	3 627
7	Médée	1635	Tragédie	14 269
8	La Place Royale	1634	Comédie	13 801
9	L'illusion comique	1636	Comédie	15 428
10	Le Cid	1636	Tragi-comédie	16 677
11	Cinna	1641	Tragédie	16 126
12	Horace	1640	Tragédie	16 482
13	Polyeucte	1641	Tragédie	16 472
14	Pompée	1642	Tragédie	16 492
15	Le Menteur 1	1642	Comédie	16 653
16	Le Menteur 2	1643	Comédie	17 675
17	Rodogune	1644	Tragédie	16 842
18	Théodore	1645	Tragédie	17 121
19	Héraclius	1647	Tragédie	17 433
20	Andromède	1650	Tragédie	15 514
21	Don Sanche	1650	Comédie héroïque	16 947
22	Nicomède	1651	Tragédie	16 923
23	Pertharite	1651	Tragédie	17 121
24	Oedipe	1659	Tragédie	18 618
25	Toison d'Or	1661	Tragédie	20 343
26	Sertorius	1662	Tragédie	17 675
27	Sophonisbe	1663	Tragédie	16 858
28	Othon	1664	Tragédie	16 971
29	Agésilas	1666	Tragédie	18 227
30	Atilla	1667	Tragédie	16 788
31	Tite et Bérénice	1670	Comédie héroïque	16 697
34	Psyché (2/3 verses)	1671	Comédie	10 067
32	Pulchérie	1672	Tragédie	16 630
33	Suréna	1674	Tragédie	16 545

This corpus contains 34 plays; 32 of them are complete. It is 553,190 tokens long and its vocabulary is: 15,535 standardised spelling forms and 6,258 different word types.

Appendix 2. *Amour* in Furetière's dictionary (1690)

AMOUR. Subst. m. et f. Passion de l'âme qui nous fait aimer quelque personne ou quelque chose. L'*amour* divin est le seul qui nous doit enflammer. Les Romains se font sacrifier pour l'*amour* de la patrie. Il faut donner l'aumône pour l'*amour* de dieu. L'*amour* paternel, l'*amour* conjugal sont les *amours* les plus violentes. L'*amour* des richesses est la cause de tous les vices, l'*amour* de la gloire est la cause de toutes les belles actions. On dit aussi, il aime d'*amour*, pour dire d'une amitié violente. Ce prince est l'*amour* des peuples.

Se dit principalement de cette violente passion que la nature inflige aux jeunes gens de divers sexes pour se joindre, afin de perpétuer l'espèce. On dit qu'un jeune homme fait l'*amour* à une fille, quand il la recherche en mariage. On le dit aussi odieusement, quand il tâche de la suborner. Il s'est marié par *amour*, c'est-à-dire désavantageusement et par l'emportement d'une passion aveugle. On dit qu'une femme fait l'*amour* quand elle se laisse aller à quelque galanterie illicite. Il y a aussi des *amours* brutaux, monstrueux et contre nature.

On dit aussi des animaux qui sont en chaleur, qu'ils entrent en *amour*, lorsqu'ils recherchent leurs femelles.

AMOURS se dit aussi au pluriel. Les livres, les tableaux sont ses *amours*, il nourrit de folles *amours*, c'était ses jeunes *amours*, ses tendres *amours*. Il signifie aussi l'objet aimé. Mon cœur, mes *amours*, m'aimerez-vous toujours?

AMOUR. Subst m. Se prend encore pour la divinité fabuleuse des païens, qu'ils s'imaginaient présider à l'*amour*. Cupidon est le dieu d'*amour*. L'*amour* est tout nu, les flambeaux de l'*amour*, les flèches de l'*amour*, le bandeau de l'*amour*, l'*amour* est aveugle.

Il signifie aussi en ce sens, tous les petits agréments qui naissent de la beauté. Les jeux, les ris, les *amours* et les grâces. Vénus est la mère des *amours*.

AMOUR, se dit proverbialement en ces phrases : il n'est point de belle prison, ni de laides *amours*. On dit encore : tout par *amour* et rien par force. On dit encore qu'une femme laide est un remède d'*amour*. On dit aussi : à battre faut l'*amour*.

(Antoine Furetière, 1690)

Appendix 3 Concordancer of "Amour" in Corneille's work (first page)

déplaisirs de mon coeur irrité, et soutient mon **amour** contre sa cruauté ; mais ce flatteur espoir qu'il
aque jour et ne lui pas tenir quelques propos d' **amour** ; mais d'un vain compliment ta passion bornée
s meilleurs partis... Trêve de ces raisons ; mon **amour** s'en offense, et tiendroit pour supplice : de rec
donner l'honneur d'accompagner ses yeux ; et l' **amour** , qui ne put entrer dans son courage, voulut ob
me il faut aimer. L'abondance des biens pour l' **amour** conjugal a de puissants liens : la beauté, les att
la cuisine ; et l'hymen qui succède à ces folles **amours** , après quelques douceurs, a bien de mauvais j
is, madame, apaisez la querelle. Un esclave d' **amour** le défend d'un rebelle, si toutefois un coeur qui
l'assurer plutôt qu'il trouveroit en ce mépris d' **amour** qui le seconderoit. Si le coeur ne dédit ce que l
dédit ce que la bouche exprime, et ne fait de l' **amour** une plus haute estime, je plains les malheureux
ans cause avec raison m'étonne : je ne reçois d' **amour** et n'en donne à personne. Les moyens de donn
e si frêles sujets ne sauroient exprimer ce que l' **amour** aux coeurs peut lui seul imprimer, et quand vo
s de glace à qui brûle pour vous. Un ennemi d' **amour** me tenir ce langage ! Accordez votre bouche a
demeure à soi. Mon coeur, jusqu'à présent à l' **amour** invincible, ne se maintient qu'à force aux term
que l'honneur t'en fera souvenir. En matière d' **amour** rien n'oblige à tenir, et les meilleurs amis, lors
tour. Sache donc qu'il ne vient sinon de trop d' **amour** . J'eusse osé le gager qu'ainsi par quelque ruse
crainte et l'espoir en balance car s'il faut que l' **amour** naisse de ressemblance, mes imperfections no
: ma présence importune te laisse à la merci d' **amour** et de la brune. Continuez les jeux que vous ave
ntiments cachés : ils ont des rendez-vous où l' **amour** les assemble ; encore hier sur le soir je les surp
. Il a lieu de s'y plaire avec quelque justice : l' **amour** ainsi qu'à lui me paroît un supplice ; et sa froid
e son teint, se rend incomparable, et je suis en **amour** ce qu'elle est en beauté. Quoi que puisse à mes
envoyant au jour, donnèrent pour nous deux d' **amour** et de mérite, elle a tout le mérite, et moi j'ai to
e mérite, elle a tout le mérite, et moi j'ai tout l' **amour** ." tu l'as fait pour Eraste ? Oui, j'ai dépeint sa f
sage jamais ne m'auroit arrêté, s'il falloit que l' **amour** fût tout de mon côté. Ma rime seulement est u
à me représenter qu'une vieille amitié de mon **amour** s'irrite, qu'Eraste s'en offense et s'oppose à Mél
ai honte de me voir insensible ou perfide : si l' **amour** m'enhardit, l'amitié m'intimide. Entre ces mou
u bout du conte, que c'est contre ton gré que l' **amour** te surmonte. Tu présumes par là me le persuad
e pour cette belle ? Il en meurt. Ce courage à l' **amour** si rebelle ? Lui-même. Si ton coeur ne tient plu
O le honteux motif à changer de maîtresse ! En **amour** . Cloris m'aime, et si je m'y connois, rien ne pe
légion de rivaux de sa sorte ne divertiroit pas l' **amour** que je vous porte, qui ne craindra jamais les hu
ispensent mon devoir de ces formalités. Que d' **amour** et de joie un tel aveu me donne ! C'est peut-êtr
coup, qui sous le faux appas des preuves d'un **amour** qui ne les touchoit pas, prenoient du passe-tem
si ton espérance à la fin n'est déçue, ces deux **amours** auront une pareille issue. Si cela n'arrivoit, je
beaux discours, un rival inconnu possède ses **amours** , et la dissimulée, au mépris de ta flamme, par
e une récompense extraordinaire d'un excès d' **amour** , dont elle tâche de suppléer au défaut des grâc
insi que sans honte à mes yeux tu subornes un **amour** qui pour moi devoit être sans bornes ? Suis mo
e ; j'en croyois ses regards, qui tous remplis d' **amour** , étoient de la partie en un si lâche tour. O ciel
x mots mon malheur et le tien. En nos chastes **amours** de tous deux on se moque : Philandre ... Ah ! l
d'être désabusé ? Apprends qu'il te faut être en **amour** plus rusé ; apprends que les discours des filles
t jours ; Philandre est aujourd'hui l'objet de ses **amours** , et peut-être déjà (tant elle aime le change !) q
larcin qu'il m'en fait me vole peu de chose, et l' **amour** qui pour lui m'éprit si follement m'avoit fait bo
éponde, et sans embarrasser son coeur de leurs **amours** , leur faire bonne mine, et souffrir leurs discou
table jour que votre bon accueil lui donna de l' **amour** , dedans ce désespoir a chez moi rendu l'âme, e
u'à ce jour, que vous relevassiez de l'empire d' **amour** ; j'ignorois qu'aussitôt qu'il assemble deux âme
que ma supercherie tournât si lâchement tant d' **amour** en furie ? Inutiles regrets, repentirs superflus,
passe ou l'égale. C'est en vain que vers moi ton **amour** se ravale ; fais lui, si tu m'en crois, agréer tes a
ous n'avons plus besoin de votre confidence : l' **amour** en liberté peut dire ce qu'il pense, et dédaigne
z ce blasphème, la bouche est impuissante où l' **amour** est extrême : quand l'espoir est permis, elle a d
mon heur, il est vrai, si tes désirs secondent cet **amour** qui paroît et brille dans tes yeux, je n'ai rien dé
on en a vu l'effet, lorsqu'à force de pleurs mon **amour** et mes soins, aidés de mes douleurs, ont fléchi
ouchant votre retour la tient encore en peine. L' **amour** a fait au sang un peu de trahison ; mais Philan
ce fut pour moi qu'il osa s'en dédire. Et pour l' **amour** de vous je n'en ferai que rire. Et pour l'amour d
l'amour de vous je n'en ferai que rire. Et pour l' **amour** de moi vous lui pardonnerez. Et pour l'amour d
l'amour de moi vous lui pardonnerez. Et pour l' **amour** de moi vous m'en dispenserez. Que vous êtes
e, inventée à dessein de nous nuire, avance nos **amours** au lieu de les détruire ; de son fâcheux succès,
our nous nous sommes rendu tant de preuves d' **amour** , et de ce que l'excès de ma douleur sincère. A
uvé de contraire à ses voeux ; outre qu'en fait d' **amour** la fraude est légitime ; mais puisque vous voul
perdant l'espérance ; encore avez vous vu mon **amour** irrité mettre tout en usage en cette extrémité ; e
u jour vous vous rendrez sensible à ce naissant **amour** . Vous prodiguez en vain vos foibles artifices ;

Appendix 4. Lexical universe of *amour* in Pierre Corneille's dramatic work
(classified by parts of speech and by decreasing weight links ; confidence interval limits: less than 1%)

1° over-utilized types

Proper names: Vénus, Psyché, Léon, Placide, Créuse, Amarante, Phinée, Daphnis

Verbs: céder, éteindre, opposer, allumer, naître, éclater, trahir, paraître, intéresser, aimer, surmonter, succéder, croître, vaincre, étouffer, pardonner, inspirer, tourner, déférer, rallumer, gémir, éprouver, couronner, mériter, brûler, souffrir, presser, faire, fléchir, produire, combattre, seoir, changer, traiter, flatter, vouloir, préférer, devoir, renaître, unir, animer

Substantives: haine, coeur, tour, jour, retour, excès, amitié, amorce, cour, noeud, estime, tendresse, objet, beauté, séjour, douceur, ardeur, soin, force, pitié, feu, espérance, respect, désir, devoir, loi, idolâtrie, aile, espoir, dépit, mère, excuse, prix, caresse, cause, impatience, faveur, discours, partage, jeunesse, balance, violence, patrie, divorce, feinte, conduite, lien, mérite, raison, transport, froideur, amant, effort, hymen, ambition, maîtresse, passion

Adjectives: conjugal, parfait, paternel, véritable, fort, extrême, tendre, aimé, doux, éteint, chaste, puissant, fou, mutuel, forcé, vertueux, éternel, solide, aveugle, feint, simple, léger, indigne, aimable, beau, ferme

Pronouns: dont, se, qui, il, lui

Adverbs : peu, toujours, plus, d'autant, aussi, ensemble, tant, quelquefois, si, auprès

Articles and other determinative words: mon, premier, tel

Prepositions and conjunctions: que, malgré, ni, soit, pour, vers, dans, quand, contre, car

2° under-utilised word types

Proper names: Romain, Rome, César, Pompée

Verbs: être, dire, aller, laisser, venir, prendre, arriver, attendre, connaître, penser, pouvoir, sortir, revoir, sembler, amener, craindre, hâter, choisir, trancher, couler, falloir, recevoir, prétendre, défaire, secourir, tomber, plaindre, rougir, marcher, pousser, suivre, fuir, punir, ouvrir, chercher, éviter, perdre, garantir, vanter, mentir, achever, pleurer

Substantives: seigneur, dieu, ciel, roi, adieu, madame, mot, prince, gens, terre, pied, monsieur, humeur, ordre, heure, ami, homme, comte, mort, lieu, temps, sort, tête, loisir, malheur, mort, avis, coup, combat, soeur, traître, destin, frère, ouvrage, bonheur, guerre, fer, zèle, sang, foudre, bataille, ombre, assassin, main, mal, monstre, père, événement, réponse, fois, oeil, victime, chef, vie, nombre, bourreau, soldat, avenir, affection, fils, pas, châtiment, place, porte, conseil, peuple, épée, parole, effroi, sujet, fortune, état, point, autel, comble, artifice, encens, gendre, assurance, vérité

Adjectives: funeste, prêt, autre, bon, faux, las

Pronouns: tu, vous, ils, en, quoi, nous, y, cela, autre, leur, vous-même

Adverbs : là, demain, bien, vrai, pas, déjà, bas, trop, encore, oui, ici, pourtant, mieux, tout

Articles and other determinative words: quel, second, ce, ton, tout, trois

Prepositions and conjunctions : donc, après, voici, jusque, avec, mais, si, sur