"O" Stories

André Dugas

The linguistic fund, that is, the actual lexical inventory of a language, is always considerably larger than the sum total of the contents of dictionaries for that language. It corresponds to all the possibilities of derivation and compounding—with the associated word-formation rules. The exploitation of the latter within machine-readable dictionaries should therefore allow a far more accurate coverage of the linguistic fund, by generating thousands of additional entries, some being more or less widely attested in their written form, some representing the set of virtual words generated by a productive rule that have not, for one reason or another, been recognized as existing words. Note that the boundary between the two subsets is not clearcut. The conditions which determine whether a generated form will belong to one or the other have not given rise to extensive studies, neither in linguistics nor in lexicology, in spite of their significance for a better understanding of lexical creativity and word-formation processes. The relevance of such phenomena seems to have been greatly underestimated—as is indicated by the fact that no studies have been fulfilled on the topic of words such as *médico-légal* and *franco-québécois*. We will demonstrate and illustrate for the French language the shortcomings of the simple compiling method and how ignoring them has led to unnecessary complications in the field of electronic lexicography.

Part I

In recent years, there has been a growing interest in the development of machine-readable dictionaries for natural languages.[0] As a consequence, a number of assumptions in the field of lexicography have seen their status change from "apparently settled" to "unsettled and unsettling". It was

[0] A different version of this paper, under the title "Formes simples en o-", was presented at the 11e Colloque sur la grammaire et le lexique comparés des langues romanes, Université Marne-la-Vallée, France, 21-24 septembre 1992. The actual paper was revised and translated by Dominique Bossé to whom we express our gratidude.

thought at first that compiling the entries to be listed simply involved going through existing conventional dictionaries and specialized lexicons. The method soon brought into broad daylight problems that had formerly been dealt with by lexicographers in a leave-well-enough-alone spirit. It was clear that a number of lexicographical issues had either been overlooked totally or had been treated with no great concern for rigor and coherence.

Consider, for example, the following question: What are the conditions under which a compound is listed as a single unit without blanks, or as a complex of units linked by hyphens, or as a sequence of simple units separated by blanks? The situation is further complicated, as recent research has shown,[1] by the fact that wavering usages have to be faithfully reported in dictionaries. We will examine a subset of this question using French data. More specifically, we will turn our attention to forms such as *franco-québécois* and *médico-légal* and to similar forms appearing in dictionaries without the hyphen e.g., *morphosyntaxique* and *stratégico-diplomatique*.[2]

The lexicographical inventory used as the basis for the design of electronic dictionaries is rather limited. There are two main reasons for this state of affairs: 1° the practical impossibility, for lexicographers, to compile more than a portion of the actual inventory of the French language; 2° the rather prosaic issue of size and cost associated with the written medium.[3] The linguistic fund, that is, the actual lexical inventory of a language, is always considerably larger than the sum total of the contents of dictionaries for that language. It corresponds to all the possibilities of derivation and compounding—with the associated word-formation rules. The exploitation of the latter within machine-readable dictionaries should therefore allow a far more accurate coverage of the linguistic fund, by generating thousands of additional entries, some being more or less widely attested in their written form, some representing the set of virtual words

[1] We are referring to works by Nina Catach and Michel Matthieu-Colas as well as to the recent position of the Conseil de la langue française sur la réforme de l'orthographe.

[2] Matthieu-Colas (1991: 51) points out that 《this type of expression is regularly hyphenated but [that] the Grand Robert lists under stratégico: stratégicodiplomatique, stratégicoéconomique, stratégicopolitique…》 (Our translation).

[3] See Dugas (1990a) and (1991), Leeman (to appear), Molinier (to appear).

generated by a productive rule that have not, for one reason or another, been recognized as existing words. Note that the boundary between the two subsets is not clearcut. The conditions which determine whether a generated form will belong to one or the other have not given rise to extensive studies, neither in linguistics nor in lexicology, in spite of their significance for a better understanding of lexical creativity and word-formation processes. The relevance of such phenomena seems to have been greatly underestimated—as is indicated by the fact that no studies have been fulfilled on the topic of words such as *médico-légal* and *franco-québécois*. We will demonstrate and illustrate the shortcomings of the simple compiling method and how ignoring them has led to unnecessary complications in the field of electronic lexicography.

Items such as *médico-légal* and *franco-québécois* are usually not listed in dictionaries and the few which are to be found are probably simply viewed as set and frequent by traditional lexicographers. It should be noted that the internal structure of these words is usually far from transparent. This is due to the fact that their formation might be the result of derivation, compounding, or syntactic word-formation rules—or a combination of the above, with a good measure of idiosyncracies, as we will see later in a detailed fashion. As to the description of these forms, two immediate difficulties arise regarding the first term of the complex items: 1°, it may be confused with other types of forms ending in *o* that have nothing to do with the processes discussed here, and 2°, it may be mistaken for a prefix—we will show why this view is erroneous. First, we will look into the recognition modes needed for the various types of items ending in *o* and we will then proceed to formulate the local rules that account for the formation of words such as *médico-légal* and *franco-québécois*.

Words Ending in *o* Excluded from Our Study

There is a distinction that needs to be established at this point, particularly for the reader who is not entirely familiar with the French language. The ending *o* is rather atypical in French for words occurring freely in isolation, yet the number and frequency of these words justifies the following qualifications. The lexical items ending in *o* which are involved in the morphological processes we are studying have no

relationship whatsoever to words ending in *o* that can occur in isolation. These words are of various types. A first set consists of borrowings that are usually known to be foreign words. For the most part they originate from other Romance languages, Italian, Spanish, Portuguese, Catalan—cacao, mafioso, tchao (ciao), torero, porto, flamenco, azulejo—, but not exclusively: allô (English), kimono (Japanese), rhô (Ancient Greek), igloo (Inuktitut), gestapo (German), ipso facto (Latin), paréo (Tahitian), taro (Polynesian), etc. Other words are, whatever their origin, perceived as essentially French: lavabo, studio, zoo, stylo, numéro, zéro. Argot (French slang) contains quite a few of these words—clodo, travelo, dirlo—and so does baby talk, a language of communication used by adults when speaking to infants and young children—dodo, bobo, toto, lolo. Incidentally, there are other reduplicated forms which are used between adults in colloquial French: coco, gogo, jojo.

Another series of words to be distinguished from the items that are the object of our study are those resulting from a simple truncation process without vowel change: négo (ciation), labo (ratoire), stéréo (phonie), météo (rologie) ou météo (rologue),[4] condo (minium), frigo (rifique).[5] These items, which are rather set, should be treated formally as the truncated member of a pair: auto/automobile. Due to the fact that the truncated word owes its status as a word to usage, it is not possible to consider generating truncated forms. In addition, it would be formally impossible to retrieve the source of the truncated form in cases such as auto (mobile) among hundreds of words also containing the prefix auto-.[6]

We set in a separate class items such as: dic-o(tionnaire), fach-o(iste), métall-o(urgiste), apér-o(itif), propri-o(étaire). The formation of these items is whimsical. They are not the result of a simple truncation and their *o* ending is not immediately motivated. These items are to be treated as entirely lexicalized entries, just as the previous class of words resulting

[4] Looking at the list of trades and professions, it appears that there is a constraint involving the feature [+hum] which generally blocks the process of truncation. However, there are a number of well-established exceptions such as: radio, dactylo, typo.

[5] If, as might be the case, frigo is formed for the brand name "Frigidaire", then it belongs to the subcategory to be seen next.

[6] See Dugas & Courtois (1988) and Dugas & Molinier (to appear, Dec. 1992) in La Productivité lexicale, no 96, Langue Française, Paris, Larousse.

from the simple truncation process. It may be noted that the process of *o* substitution or commutation is identical to the one we will encounter in the formation of *médico–* in *médico–légal*. But this is as far as the similarities go.

Forms Constructed in *o* Ending and Prefixes Ending in *o*

As we have seen above, items ending in *o* as the result of truncation are autonomous units; this is not the case for items such as médico. Another difference between the two types lies in the fact that substitution of *o* in médico follows a universal syntactic rule that has no equivalent in the case of truncated items.

In the case of *médico–*, the base *médical* is an autonomous lexical item, a fact that distinguishes items of this type from prefixes. We can sum up our observations so far in the following table:

Table 1. Binary Features of Morphemes Ending in *o*

	autonomous	truncation	autonomous base
numéro	+	−	−
météo	+	+	+
pseudo–	−	−	−
médico–	−	−	+

The use of distinctive features for the description of the various types of lexical items ending in *o* is useful in representing the double autonomous status of *météo* (*météorologie* being the autonomous base) and the fact that, while the form *médico–* is not the result of a straightforward truncation, it is nevertheless related to the autonomous base *médical*. Note also how *météo* and *médico–* both share one positive value with the autonomous and untruncated word *numéro*. Prefixes are assigned negative values for all features.

Part II

We will now focus on the elements of description needed for the automatic recognition of complex items such as *médico–légal* and *franco–québécois*. As we have already seen above, these forms are usually absent from

dictionaries. They result from the application of a general rule of coordinating conjunction reduction: an institute whose activities take place within both the medical *and* legal fields will be identified as a "médical" and "légal" institute or, more simply, as a "médico-légal" institute. In a similar fashion, an agreement taking place between "français" and "québécois" partners will be called a "franco-québécois" agreement. Taking the conjunction of attributes to be the source of these compounds, we can see two successive operations, namely the application of a general rule of coordinating conjunction reduction and that of a substitution rule which turns the suffix or pseudo-suffix ending of the first term into o to match the canonical form of a prefix, thus yielding a pseudo-prefix. The application of the two rules takes place as follows:

Table 2. Rules Yielding Forms Ending in *o*

Source	médical et légal	français et québécois
1. *Reduction rule*	*médical légal	*français québécois
2. *Substitution rule*	médico-légal	franco-québécois

The output of the reduction rule, *médical légal or *français québécois, is ungrammatical because it violates a general syntactic constraint which prevents sequences of lexical items belonging to the same category and bearing the same function:[7]

*Jules Jim viennent voir Jeanne.
(*Jules Jim are coming to see Jeanne.)

*Ce cheval est vieux fourbu.
(*This horse is old exhausted.)

[7] Sequences such as "the nice little French girl" are not to be taken as counterexamples. Rather, the phrase "same function" must be taken in the strictest sense. In the sequence given, "French" qualifies "girl", "little" qualifies "French girl" and "nice" qualifies the whole structure phrase "little French girl". This analysis is borne out by the well-known fact that the terms of such a sequence must respect an order of intrinsicness or inalienability, viz. *the French nice little girl, *the little nice French girl, *the French little nice girl. The same analysis obtains for sequences of determiners in Italian: "la mia famiglia", "un mio amico".

The *o* endings can be substituted to a variety of suffixes or pseudo-suffixes; the specific substitution rules are therefore as numerous as the actual suffix endings to be replaced. There seems to be a preponderance of identical endings in terms to be linked, as for méd*ical* and lég*al*. In terms of automatic analysis, such a suffix identity criterion would be extremely helpful if it were reliable but it is not the case: unidentical endings are also quite common, for example, franç*ais* and québéc*ois*.

Since a coordinating conjunction can only links two words bearing the same grammatical function, it follows that the conjunction reduction rule can only apply when lexical items on either side of the conjunction belong to the same category. For instance, *franco-québécois* can only originates from the conjunction of two adjectives or two nouns, viz.:

français (adj.) québécois (adj.)

or

français (n.) québécois (n.)

but not

français (adj.) québécois (n.)

nor

français (n.) québécois (adj.)

As a matter of fact, most of these items will be analyzed as adjective compounds which can be used nominally, e.g., *euro-asiatique*.[8,9] Other items are formed directly from nouns into nominal compounds: *cyclo-tourisme*, *vibro-masseur*, *cumulo-nimbus*.

The lexical creativity of these compounds is highly constrained as to the

[8] The formation of the compound *euro-asiatique* from the sequence *européen* and *asiatuque* is transparent. Note however that the formation of inflected forms is not entirely clarified. The analysis outlined here seems to suggest that *euro-asiatiques* has *européens* and *asiatiques* as its source but there are independently motivated principles in morphological theory that might require *euro-asiatique* to be inflected after compounding. The issue, for all its significance in general linguistic theory, is not crucial in this study.

[9] The formation of *euro-dollar* would be problematic in our analysis if we were to try to postulate the pair *européen* and *dollar* as its source. However, since the Robert dictionary gives 1965 as the data of the appearance of the word in French, as opposed to 1960 in English according to the Webster's, it is entirely justified to treat it as a simple borrowing—it is even clearer when we think of the economic situation in which the term was introduced.

order of their terms. When the order of the terms is inverted, the degree of acceptibility of the compound drops dramatically, e.g., *?maso-sadique* instead of *sado-masochiste*.[10] However, some compounds involving ethnic or geographic terms seem to be less constrained: *arabo-soviétique* or *soviéto-arabe* – but note *américane – soviétique* and *?soviéto – américain*.

Some compounds involving ethnic or geographic terms are quite old and diachronic change has blurred the transparency of their formation. The number of such items is quite limited and, therefore, the principled necessity of listing them extensively in the electronic dictionary of French poses no practical problems. Here are some examples:

chamito-sémitique
finno-ougrien
sino-russe
nippo-américain
gallo-romain
hispano-africain
britto-normand
luso-espagnol
nilo-saharien
judéo-chrétien
malayo-polynésien

Part III

We will now give an overview of the main cases of substitution which play a role in the formation of compounds such as *médico-légal* and *franco-québécois*. At the outset, it is important to establish a few facts about the *o* found in the first term. The *o* of these items is unique in French in that, unlike what prevails for French affixes in general, this phonetic segment has no semantic value and its sole role is that of a marker to indicate that

[10] The case of the pair *sado-masochiste* and *?maso-sadique* is particularly interesting because of the existence of *maso* as an autonomous truncated form having *masochiste* as its autonomous base. It goes to show one thing: there is no principle of economy that would require an existing canonical form, viz. *maso* to be used to avoid the creation of a new one, *sado*. On the contrary, there might be a constraint on the use of items of that type in a prefix position.

the compound is formed from the reduction of a coordinated sequence of nouns or adjectives. It is interesting to note that these forms do not admit the plural marker, neither from a phonological point of view nor in the corresponding graphic form: *[frankozamerikin] and *Francos-Américains.

There are basically two types of *o* substitution: in the first type, *o* is simply added to a monomorphemic word ending in a consonant; in the second type, a suffix or suffix-like element is truncated and replaced by *o*. In either case, independently motivated morpho-phonological-or graphical-adjustment rules apply when required by the context.

Table 3. Types of Substitution

First type: *o* added to a monomorphemic noun or adjective

… : picardo-normand

Second type: *o* replaces a suffix or suffix-like element

ain : cubano-américain
aire : musculo-vasculaire
aire : égalitaro-subjectif, militaro-politique[11]
ais : anglo-saxon
al : spatio-temporel
ateur: vibro-masseur
e[12] : serbo-croate
icain : afro-asiatique
ie[12] : syro-égyptien
ien : aéro-naval, arméno-libanais
ieur : antéro-postérieur
ite : broncho-pneumonie
ique : germano-tchèque
isme : sado-masochisme
ois : dano-suédois
logue: socio-linguiste
logie : socio-linguistique
us : cumulo-nimbus

[11] The source of *militaro-* in this compound is ambiguous: is it *militaire* or *militariste*?

[12] -e and -ie are not, of course, suffixes but suffix-like endings.

Adjustment Rules

Morpho-phonological adjustment rules apply when the consonant which precedes the *o* is immediately preceded by the anterior vowel /è/or by the nasal vowel /ain/. Here follows an overview of the modifications to the radical of words to which *o* is attached by substitution or by simple addition.

Vowel Modifications

In the context of a consonant followed by *o* in word-final position, there are:

- two cases of denasalization

/ain/$_1$ changes to a-n, as in cubain/cubano

/ain/$_2$ changes to i-n, as in latin/latino

- one case of mid vowel raising, the mid-low /è/ raising to the mid-high /é/, as in grec/gréco
- one case of /è/ lowering to /a/, as in militaire/militaro

Consonant Modifications

When the consonant is /f/, it changes to /v/, as in collectif/collectivo-. In the case of français/franco-, the sound, /s/ changes to /k/, with the corresponding loss of the cedilla in the written form.

Strictly Graphical Adjustment Rules

Graphical adjustment rules apply to the following segments when they occur before the added o:

- qu becomes c, as in tragique/tragico
- e becomes é, as in Alger/algéro

Part Ⅳ

A further study, still in progress, is aimed at determining other properties of compounds whose first term ends in *o*. Of particular interest are the

productivity constraints of derivational processes. While suffixation and back-formation seem commonplace (e.g., judéo-chrétien/judéo-christianisme, socio-linguiste/socio-linguistique), preliminary results reveal a basic incompatibility between this type of compounding and prefixes. It is not yet clear whether the presence of a prefix prevents the truncation process which has to take place before the addition of *o*, or if all prefixation itself becomes impossible after compounding. The fact is that prefixed radicals are not truncated and that truncated radicals are not prefixed, viz. soviéto-américain but *post-soviéto-américain.[13] On the one hand, the compounds under study line up quite nicely with other French compounds for which prefixation is forbidden, but on the other hand, the crucial data, namely prefixed words which would have undergone truncation, are totally lacking at this point.[14]

In the same line of thought, we are also focusing on the differences in productivity between certain *o* series such as if, iste (collectif, collectiviste), *e*, iste (intime, intimiste), aire, iste (militaire, militariste). There might be a systematic relationship between the nature of the suffixes and the degree of productivity of *o* compounds. We are still in the process of checking whether compounding with *o* would be restricted to nouns and adjectives bearing specific suffixes—for example, *-ais* or *-ois* in the case of ethnic or geographical compounds or *isme, logue, logie, ieur* for other semantic types.

Our present data do not enable us yet to discern any significant regularities as to compounds containing more than one term ending in *o*, of the type examplified in the following sentence:

…les récents accords canado-américano-mexicains sur le libre échange…

The formula "les accords Canada-États-Unis-Mexique", using the full substantives instead of a compounded adjective, is clearly preferred, at least in the electronic press. It is as if the very length of the compound acted as a major constraint, rendering it increasingly opaque with each successive addition of an element ending in *o*: consider, for example, *bio-*

[13] But note *pseudo-logico-philosophique.*

[14] We are thinking of something like *intello* for *intellectuel* but in which the *in-* would be a negative prefix.

logico-physico-mathématique or *franco-italo-anglo-belge* (Mathieu-Colas, 1991). In actual speech, a rather marked pause is necessary between each element of the compound to ensure intelligibility.[15]

Conclusion

In this study we have attained two types of results. On the one hand, we have established a number of facts concerning compounds whose left member ends in *o* and we have provided a set of formal criteria to be used in the automatic analysis of these compounds. On the other hand, this study offers a contribution to better understanding of lexical productivity by raising, in very concrete terms, crucial questions relevant both to general morphological theory and to the automatic treatment of natural languages.

References

Catach, Nina (1980) *Orthographe et Lexicographie-II Les mots composés*, Paris, Nathan.

Catach, Nina (1989) *Les Délires de l'orthographe*, Paris, Plon.

Conseil de la Langue française (1990) *Les Rectifications de l'orthographe*, Journal officiel de la République française, édition des documents administratifs, n° 100.

Corbin, Danièle (1987) *Morphologie dérivationnelle et structuration du lexique*, Tubingen. Max Niemeyer Verlag, 2 vol [XVII+957 p.].

Courtois, Blandine (1985) *Dictionnaire alphabétique inverse du français*, Rapport technique n° 10. LADL, Université Paris VII.

Dugas, André (1990) 'La Création lexicale et les dictionnaires électroniques,' in Bl. Courtois et M. Silberztein, ed., *Dictionnaires électroniques du français*, *Langue française*, n° 87, Paris, Larousse, pp. 23-30.

Dugas, André and Christian Molinier, ed. (1992) *La Productivité lexicale*,

[15] In fact such super-compounding appears to be so undesirable that formulas such as "l'entente Canada-États-Unis-Mexique" or "la coalition France-Angleterre-États-Unis" become an absolutely acceptable alternative to the form compounded with *o* even though they involve a violation of the principle of "no coordination without coordinators" which we described earlier when we explained the formation of *franco-québécois* from **français québécois*.

Langue française, n° 96, Paris, Larousse.

Dugas, André (1991) 'Automatic Generation of New Lexical Items in an Electronic Dictionary,' in *Proceedings of the International Conference on Computational Lexicography*, Balatonfüred, Hungary, 8–11 sept. 1990. Research Institute for Linguistics, Hungarian Academy of Sciences, pp. 51–73.

Dugas, André and Blandine Courtois, 'Les verbes préfixés en auto–,' *VIIe Colloque International sur le lexique et la grammaire des langues romanes*, Lalonde–les–Maures, France, 20–24 sept. 1988. Inédit.

Laporte, Éric (1992) 'Adjectifs en –ant dérivés de verbes,' In Dugas and Molinier, 1992.

Leeman, Danielle (1992) 'Adjectifs et formes verbales en –ble,' In Dugas and Molinier, 1992.

Matthieu–Colas, Michel (1991) 'Composés à trait d'union: la productivité des thèmes en –o,' *Rapport L2/91, Programme de Recherches coordonnées –Informatique linguistique*, LADL, Université Paris VII, pp. 47–58.

Molinier, Christian (1992) 'De la productivité adverbiale des adjectifs,' In Dugas and Molinier, 1992.

Muller, Claude (1990) 'Contraintes de perception sur la productivité de la préfixation verbale en dé–négative,' *Travaux de Linguistique et de Philologie* XXVIII, pp. 171–192.

Nam, Jee–Sun (1990) 'Sur une construction N_0 N_1 –ita en coréen,' *Linguisticae Investigationes* 15.2, Amsterdam/Philadelphia, Benjamins, pp. 301–341.

Silberztein, Max (1989) 'The lexical analysis of French,' in *Proceedings of the 11th Spring School in Theoretical Computer Sciences*, Berlin, Heidelberg: Springer Verlag.

Thimonnier, René (1967) *Le système graphique du français*, Paris, Plon.

Zribi–Hertz, Anne (1972) *Remarques sur quelques préfixes du français*, Doctoral Thesis 3e cycle, Université Paris VIII.

Université du Québec à Montréal
Case postale 8888, succursale A
Montréal (Québec)
Canada H3C 3P8