3-14-1989

The Reaccentuation of French Text

Deryle W. Lonsdale
THE REACCENTUATION OF FRENCH TEXT

Deryle W. Lonsdale
ALPNET

OVERVIEW

Accented letters serve an important role in French orthography. This paper begins with a brief description of the function of accents in French.

Although the learner of French, native or otherwise, usually learns the correct use of accents through many dictation drills, the accentuation of French text is not always a clear-cut issue in practice. Samples of different French text types reveal inconsistency in the accentuation of upper-case letters, disagreement on the alphabetical order of accented letters, and varying degrees of disregard for accents. An examination of these and other uncertainties of French accent usage will be made.

While the absence of accents in French text does not usually inhibit the human's understanding, there are cases when accents can play a major role in a wide range of natural language processing applications. The motivation for providing correct accent information to such applications is presented.

Finally, this paper investigates the development of on-line "reaccenting" methods capable of re-inserting accents into French text where they may be absent. Different methods for accomplishing this task are presented, and conclusions about their relative computational and practical merits are drawn.

THE FRENCH ALPHABET AND ACCENTS

The French language uses the Roman alphabet with the same 26 letters as English has, as well as the following 13 accented letters:

"â", "ê", "û" with the grave accent
"é" with the acute accent
all vowels with the circumflex: "â", "ê", "î", "ô", "û"
"ç" with the cedilla
"ë", "ï", "ü" with the dieresis

These letters constitute the standard 39-letter alphabet accepted by most modern French lexicographers.
Some other accented letters do occur in French, on words such as "cañon", "sertão", and "Ångström", borrowed from other languages. We will not consider these exceptional accents; nor will we regard the dots over "i" and "j" as accents, though some experts do.

FUNCTION OF ACCENTS

French accents have been classified into three functional types: phonographic, morphographic, and logographic. A phonographic accent phonetically alters its associated letter. A morphographic accent is used to distinguish similar morphological inflections of the same base form. Logographic accents, too, serve to differentiate otherwise indistinguishable words, but in this case unrelated ones.

The cedilla plays a purely phonographic role, as do the dieresis and the acute accent. They all have an effect on the pronunciation of the letter. An interesting function of the acute accent occurs in some subject-verb inversions; when "je pense" is inverted, for example, the structure is written "pensé-je".

The circumflex has all three roles:

1) phonographic, lengthening the vowel sound, as in:

"pâle, frèle, maître, ôter, vôtre"

2) morphographic, especially to distinguish between subjunctive and indicative verb forms:

"fut, fût"
"eut, eût"
"tint, tint"
"dites, dîtes"
"prit, prît"

but also possessive adjectives and pronouns:

"votre, vôtre"
"notre, nôtre"

3) logographic as in:

"du, dû"
"forêt, forêt"
"matin, mâtin".

The grave accent is, for the most part, logographic, playing an important role in the discrimination of high-frequency words which would otherwise be homographs, such as:
"a, à"
"la, là"
"ça, çà"
"ou, où"
"les, lès"

When over "e", the grave accent usually induces a more open articulation, and is therefore phonographic in cases such as:

"trèfle, dès, père"

ACCENT USAGE

French orthography is difficult, and the Académie Française has been invaluable in defining proper orthographic use, accents included. On the other hand, there are still several problem areas in French accent use which are seldom, if ever, addressed in the literature.

For example, the Académie has not published its authoritative dictionary since the eighth edition of 1932-35. Only a few lexicographers or dictionary publishers have dared further the official position by proposing and implementing their own up-to-date set of guidelines for accent usage.

The correct accentuation of borrowed words from other languages has always been problematic. For example, the Académie accepts "à priori", but almost nobody uses the accent. "Speculum" can often be seen with an acute accent over the "e", but not always. While foreign words continue to enter into French, foreign word accentuation patterns remain ambiguous, and writers often follow their own preferences and inclinations.

The alphabetization of accented French is another problem. The dictionary industry has apparently not yet reached a consensus on the alphabetic order of accented letters. For example, Robert and Littré list these words in order: "cote, côte, coté, côté" while Larousse orders them "cote, coté, côte, côté". Similarly, Robert and Littré show "gène" followed by "gène" while Larousse orders them in the opposite fashion. There seems to be no published account of these decisions, much less a standard in general for French alphabetization.

While several French language reference books discuss accent use in general terms, very few endorse or proscribe the use of accents with uppercase letters. In France and Francophone Europe in general, the tendency is not to accent capital letters. Nevertheless, there is little documentary support for such an approach. In only one French orthography manual
have I found such guidance. It mentions that the cedilla must be used on the uppercase "C" even when it is handwritten, but that otherwise one should avoid using accented capital letters, except perhaps when the results could be confusing.

On the other hand, the French Canadians tend to accent their capital letters more often, again with little documentary justification. Jean Darbelnet states that "Å" is used by some printers, the circumflex is often left off "O", and "É" is justified in all cases. In another reference work, a Canadian typesetter states that one must use all accents on capital letters in French.

There are abundant examples illustrating the inconsistent use of French accents on capital letters. One need only consult daily, weekly, or monthly newspapers, newsletters, and magazines.

Another related, and curious, phenomenon occurs in the formation of French acronyms. Even in instances where an accent is used on a capital letter in an expanded name, the corresponding acronym will rarely have the accent present. Why, for example, does the Canadian federal government use "É" in "Ministère de l'Énergie, des Mines et des Ressources" but not in the corresponding acronym "EMR"?

Technology plays an integral role in the use of accented capitals and the alphabetization of accented words. The ability (or inability) of typewriters and typesetting machines to generate these letters has probably set today's accentuation trends by default. Similarly, today's computer hardware manufacturers, and developers of word processors, desktop publishing systems, and communications protocols, unknowingly influence current and future practices as they define their product offerings. To the extent that they provide easy processing for accented forms, they will encourage rather than discourage the wider use of accented letters.

For example, Canadian phone books seldom show the accent on the first letter of a person's last name; this is probably due to the fact that alphabetical sorting procedures, which operate on the name, would have to take these accented letters into account. Yet in the same phone books, accented word-initial capital letters are used when no alphabetic order is required, such as in the middle of a yellow-pages advertisement. Consider also the dictionary which had entries like "A BON MARCHÉ" in its 1952 edition--was "Å" avoided for technical reasons? This seems likely since a newer version of the same dictionary published in 1987 shows such entries fully accented. Here we see clues that technology may directly influence accent usage.
In the world of everyday writing, printing, and publishing, the use of accents is conditioned by habits, vague orthographic principles, personal preference, and even technology. While no issue of orthography needs to be fully legislated or documented, one senses that the characteristic French vigilance for proper linguistic form falls short in accentuation issues. This in turn directly impacts the field of natural language processing.

We will now examine another accent-related phenomenon which also has far-reaching implications for the French text processing industry.

DEACCENTED HOMOGRAPHS

A "homograph" is a word which has one lexical form, or spelling, but which is inherently ambiguous, representing several meanings. French homographs are plentiful; here are a few examples:

- est: inflection of the verb 'être' ("to be")
  noun meaning "east"

- noyer: verb meaning "to drown"
  noun meaning "nut tree"

- mises: 2 inflections of the verb 'miser' ("to bid")
  plural of the noun meaning "a positioning"
  1 inflection of the verb 'mettre' ("to put")

Many orthographic works have a good inventory of such homographs.

More interesting to us is the concept of "deaccented homographs". These are sets of words which are not true homographs, but do become homographs when accents are missing. For example, the words "jeune" and "jeûne" are not homographic when written as such, but are deaccented homographs when the circumflex is left off the latter.

In the headwords of the Petit Robert there are over 720 sets of deaccented homographs. Note that this number only includes base forms, but no inflections of any kind. There are thousands of sets of deaccented homographs in French when all inflected forms are taken into account.

The classification of such homographs reveals several different types. For example, most of the Robert headword deaccented homographs are of the type "noun/er-verb participle" such as:

"abîme, abîmé"
"bouche, bouché"
"courbe, courbé"
"pénicille, pénicillé". These pairs are closely related semantically and share the same orthographic derivations. Formed from the most common verb conjugation paradigm, they constitute a very productive set.

Another kind of deaccented homograph set is one where ambiguity is introduced by the addition of a prefix such as "dé-" or "ré-":

-denier, dénier"
-recréer, récréer"
-repartir, répartir"
-reformer, réformer".

Yet another is caused by different spellings for the same word, based on pronunciation, and often occurs when words come into French from another language. Usage often vacillates between the type of accent on a particular letter:

-"foène, foène"
-"prèle, prèle"
-"poèle, poèle"

or with the presence/absence of accents:

-"gable, gâble"
-"tépidarium, tépidarium"
-"rodeo, rodéo"

or even the shifting of accents from one letter to another:

-"taïisme, taoïsmes".

Such problems have caused much difficulty in the past, and in 1976 a French government departmental order decreed that variations like:

-referendum, référendum"
-"événement, événement"
-"épitre, épître"
-"crûment, crument"

would be tolerated as long as a confusing homograph would not result (as would be the case with "tâche, tache" or "dites, dites").

Another common set of unaccented homographs results from the logographic role of accents mentioned earlier. More examples of such pairs include:

-"sûr, sûr"
-"mur, mûr"

Verbal conjugations are particularly productive in the creation of other types of deaccented homographs:

-"tue, tué"
-"prises, prisés"
-"rimes(pl), rîmes"
"ne, né"
"rites(pl), rites"
"jeune, jeûné"

Finally, there are many sets of deaccented homographs whose convergence can be best described as "accidental":
"mais, mais"
"mâcon, maçon"
"mème, mémé"

Obviously, these are but a few of the thousands of deaccented homographs in the French language. These words only differ by accent information—sometimes with a complete change in meaning, sometimes with none. Accents, then, are often crucial to the resolution of homographs. It is when these accents are missing from text that computer-based language processing becomes problematic.

THE PROBLEM

As we have seen, it is not uncommon for words to be deaccented, especially when there are capital letters involved. Most people who know French are not seriously hindered by the absence of accents in a French text. Context can usually serve to eliminate any confusion that may arise. For example, one would not have any difficulty understanding:

Il travaille aux Etats-Unis.
Voilà l'Ecole militaire.
A Paris il neige très peu.

We can reason, upon seeing a sign saying SALLE DES CONGRES, that it probably announces a convention center, and not an eel warehouse. In some cases, though, even with our capacities for reasoning and processing contextual clues, a misunderstanding may still result. Both readings of each of these newspaper headlines are plausible:

UN POLICIER TUE
AUGMENTATION DES RETRAITES

Even whole sentences may not give enough information to a human for correct understanding of deaccented text. Consider:

ELLE EST A LA POURSUITE D'UN DIPLOME.

LES PREMIERS COLONS D'AMERIQUE S'ADAPTERENT VITE A LA CONSOMMATION DU MAIS.

IL Y A CETTE ANNEE UNE RECOLTE SURE POUR LE VIGNOBLE BORDELAIS.
Clearly the accent is sometimes necessary, even to the human reader:

(DIPÔME or DIPÔMÉ?)
(COLONS or CÔLONS?)
(SURE or SûRE?).

Whereas the human may have little difficulty solving problems such as these, the computer is definitely at a disadvantage. The expertise and experience of human understanding far outstrip computational capacity. Yet computers are increasingly being called upon to handle such "fuzzy" and "grey" areas of accent usage as missing or out-of-order accents.

Special algorithms and methods must be developed to allow computers to "reaccent" deaccented text when necessary. Many different types of natural language processing software require this capability. We will now mention some of these applications.

The correct use of accents is problematic even for Francophones. Entering accented characters with word processors is often awkward—requiring several keystrokes—and therefore error-prone. Still, many word processors do not yet allow for accent error detection and correction. Some do not even permit accented capital letters; others do, but flag such words as erroneous.

Some text formatters have the ability to split words into hyphenated syllables. As with word processors, they can be very unforgiving programs when run against text not exhibiting the accentuation characteristics expected by the developers.

There is an increasing number of text-to-phonetic-representation and text-to-speech utility programs. The phonographic accents discussed earlier must be present in the input text to assure correct acoustic rendition of certain words.

The storage, management, and retrieval of terminology also requires the consideration of accent information. The resolution of deaccented homographs and accented alphabetization must be central to the design of a useable terminology management system.

Some natural language processing applications do varying degrees of morphological, syntactic, and even semantic parsing of the input text. Such programs are used for database querying, natural language translation, text
summarization, grammar and stylistics checking, and intelligence gathering. They, too, require accented input or the ability to reaccent input.

What is the extent of the problem? While it is true that most texts have accents in them, some texts may have suspect or missing accent information. Often TELEX messages, for example, are in all upper case and have no accents. Technology must compensate in such cases. In French technical text only 3% to 6% of all letters are accented. When full reaccentuation is required, though, the task is more demanding: a full 45% of all letters are the five vowels and "c", which can all potentially take accents. Every word is suspect in such an environment.

THE SOLUTIONS

There are several approaches to the reaccentuation of text. The following is by no means an inventory of all possible algorithms, but rather the ones I have successfully implemented.

The most obvious approach is a straightforward one: just calculate all of the word’s possible accented permutations, and sift through or filter out the extraneous ones. Such an approach done on the word "plat" would produce the three possibilities: "plat", "plât", and "plât". The last two could be discarded. This approach, though, quickly multiplies the number of possibilities. For example, consider these words and the number of permuted reaccentuation possibilities each has:

- **plat:** 3
- **eau:** 60
- **événement:** 625
- **élémentaire:** 5625
- **événementellement:** 12,500
- **hétérogénéité:** 18,750.

This factorial explosion, or overgeneration, of possible forms makes the permutations approach, despite its simplicity, impractical.

One could constrain the number of results from this approach by adding rules such as:
- "e" cannot have a circumflex or grave at the end of a word
- "è" cannot begin a word.

However, the use of such ad-hoc rules introduces a whole level of complexity and data management issues to the solution. The set of such rules could easily become open-ended, unintuitive, and internally conflicting.
A "cut-and-paste" morphology approach could be used where morphemes are identified, affixes stripped, and forms reaccented. For example, one could use rules like:

- if a word ends in "-atre", try it with "-âtre"
- if a word begins with "de-", try it with "dé-
- if a word begins with "aero-", try it with "aéro-"

Again, these rules could potentially become unmanageable due to their complexity and internal interference.

Some word processors have a list of all possible inflected words in French, usually at least 1,750,000 words, in their spelling dictionaries. The creation and updating of such a list is a very large undertaking indeed, as French is very rich in inflected forms. For words (and unaccented variants) not contained in the list, a straightforward word-list retrieval system cannot provide reasonable reaccented alternatives.

Another approach is one which I call "tiling n-graphs". It involves the use of n-graphs, or sequences of n letters. A tetragraph, for example, is a sequence of four letters. To "tile" n-graphs is to overlay them, much as tiles overlap each other on the roof of a house. We will work through an example of reaccentuation by tiling tetragraphs.

First, though, we identify and deaccent all tetragraphs in the French language. Mathematically, there is a possibility for well over 2.5 million sequences of four French letters (including one character for word boundaries). In practice, however, we only end up with some 7,300 deaccented tetragraphs in French which can take accents. Each deaccented tetragraph has a few reaccentuation possibilities, or tetragraph substitutions. Any tetragraph not in the dictionary cannot take an accent.

Here are a few examples of entries from the French tetragraph dictionary: (# is the word-boundary character)

```
acad  #gai  depe  aine  gue#
----  ----  ----  ----  ----
acad  #gai  dépe  aine  gué
açad  #gai  dépe  aîné  gué
    #gai  dépé  aîné  gué
    dépé  aîné  aîné
```

This is how tetragraph tiling is done:

1) the input word (with its boundaries) is "exploded" into tetragraphs
2) each successive tetragraph is looked up in the tetragraph dictionary

3) successful tetragraph substitutions are saved and carried on to the next iteration, while those which fail because of contradiction to pre-existing conditions are discarded.

Here is an example showing: the source word "#evenement#"; its tetragraphs and reaccented variants supplied from the deaccented tetragraph dictionary; and iterated matching and discarding of reaccented variant strings. The single reaccented form #événement# is the final result. {} means the entry is not in the dictionary—use the tetragraph exactly as it appears, with no reaccented variants possible.

#evenement#

<table>
<thead>
<tr>
<th>#eve</th>
<th>even</th>
<th>vene</th>
<th>enem</th>
<th>neme</th>
<th>emen</th>
<th>ment</th>
<th>ent#</th>
</tr>
</thead>
<tbody>
<tr>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>#évé</td>
<td>évén</td>
<td>vène</td>
<td>ènem</td>
<td>néme</td>
<td>émen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#évén</td>
<td>vène</td>
<td>nème</td>
<td>émén</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
#véne
#vévé
#évé
#évé
#éven
#événe
#événe
#événenem
#événeneme
#événenemen
#événenement
#événenement#

As can be seen, in such an example as this one, we have avoided overgeneration by applying data on the immediate context of French accented letters. In fact, the tetragraphs capture inherently the immediate lexical environment of
accented letters. Using this data, any new word can be successfully analyzed.

Here are the numbers of reaccented possibilities supplied by tiling tetragraphs on the words shown in the permutations examples:

- plat: 1
- eau: 1
- événement: 1
- élémentaire: 4
- éventuellement: 1
- hétérogénéité: 2.

This tiling approach can be used with digraphs, trigraphs, or any other length of n-graphs. For our particular implementation's memory and disk storage requirements, processing with tetragraphs proved to be the best approach. Dynamic updating of the n-graph dictionary can be easily accomplished automatically as new forms are found while handling input texts. The tetragraphs and their data can be stored in an indexed file, hash table, lexical trie, or finite-state automaton.

This tiling n-graph approach, while avoiding overgeneration to a great degree, nonetheless does generate incorrect forms. For example, consider these 4 forms generated for "élémentaire":

- élémentaire, élémentaire, élémentaire, élémentaire.

One way to constrain the generation of such forms is to assign to each tetragraph substitution a preference or weight based on its frequency of occurrence in French. If we consider the forms mentioned above, and redo the tiling with weighted substitution tetragraphs, we reject all incorrect reaccented forms for "élémentaire" and "hétérogénéité" by a wide margin.

One other reaccentuation approach involves the lexicon. Natural language processing applications usually consult an on-line dictionary or lexicon. While containing application-specific data, this same lexicon can also provide reaccentuation data for known words. In lexicons storing entries in alphabetic order, keyword letters and their accents can be considered separately, rather than together. Of primary consideration is the letter; the accent, though important, is secondary. The lexicon's data can then be sequentially organized, primarily by the entries, and if necessary based on any accents. Deaccented homographs can then be stored contiguously in the lexicon. Once a direct read into the lexicon has been performed, either successfully or unsuccessfully, comparatively inexpensive sequential reads can be performed to collect any
or all deaccented homographs. Any accent information on the original word can be consulted to filter out inappropriate results.

USER INTERFACE

All of the above approaches have been implemented in practical systems with varying degrees of success. Ultimately, though, they must fit the requirements of a human user. The user's needs can best be met when he is allowed to select the degree of reaccentuation needed. Reaccenting requires noticeable overhead, so it is in the user's best interest to select a level of processing best suited to the source text's characteristics.

The user can control the system based on letter-case accenting considerations. REACCENT NOTHING means that the input text is assumed to be correctly accented. REACCENT CAPITALS ONLY means that the input text has deaccented capital letters, so the system will automatically reaccent capitals. REACCENT ALL LETTERS means that the input text is missing accents on both lowercase and uppercase letters. Any accents present will be accepted as valid, but any unaccented letters are reaccented when possible.

The user can also control processing based upon the reliability of the input text itself. TRUST ACCENTS means that the accent information is expected to be there, and also to be correct. TRUST OVERT ACCENTS means that if a word has an accent anywhere on it, it is safe to assume that all accent information present for that word is correct. TRUST NOTHING means that no accent information (i.e. the presence or absence of accents on each letter) is considered reliable, and all accents are ignored.

SUMMARY

Although French accentuation is a complex orthographic problem complicated by inconsistent usage and constrained by technology, there are ways to minimize its impact on French natural language processing applications by providing for user-controlled reaccentuation of texts deficient in accents.
FOOTNOTE REFERENCES


BIBLIOGRAPHY


