Developing a Lexicon for a New French Spell-checker

Thierry Fontenelle
Microsoft Speech & Natural Language Group
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
USA
thierryf@microsoft.com

Abstract
This paper describes some of the problems the lexicographer was faced with when developing the lexicon for a new French spell-checker which shipped in the Microsoft Office 2003 Service Pack 2 in 2005. The new functionalities of this spell-checker are described, focusing on the implementation of the French spelling reform recommended by linguistic authorities such as the French Academy. The impact of these functionalities on the structure and content of the underlying speller dictionary is discussed. It is also shown that many of the decisions that are made during the design phase about issues like tokenization and word-breaking influence the contents of the dictionary as well as the performance of the speller.

1 Introduction

Developing a lexicon for an application like a spell-checker is a lexicographical activity which is both similar to and different from the more 'traditional' lexicographical task of writing a paper dictionary. It is undoubtedly different insofar as the lexicographer does not write definitions or select examples, since spell-checker dictionaries are normally not visible to the end-user and do not contain these items of information. Since spell-checkers are also mainly used to check the spelling of words, no grammatical information is generally included either and the problems revolving around sense distinctions are much less acute than in a traditional dictionary. The relationship between the speller lexicon and morphology is much more crucial, however, and the lexicographer who is working on the development of a word list for this kind of proofing tool needs to be well-versed in the morphology of the language he or she is working on in order to make sure all the inflected forms of every single lexical item included in the lexicon are appropriately generated. The lexicographer also needs to make decisions as to which items are going to be granted entry status, which means that the primary material he or she uses is made up of corpus data and concordances and frequency lists which make it possible to identify the neologisms which deserve lexicalization, or the named entities (city names, first names, famous people's names...) which users expect their spellers to recognize (see also Fontenelle 2004).

2 Spelling reform

It is not uncommon for languages to undergo a spelling reform and this naturally has huge implications for the compilation of reference works. In the case of French, the spelling reform in fact dates back to 1990, when the main linguistic authority, the French Academy, published a list of changes whose aim was to simplify the spelling and remove a number of inconsistencies. This reform was endorsed by most linguistic authorities, but the public (especially in France) originally failed to start applying it. It is only towards the end of the 1990s that some magazines and scientific publications in France started applying some of the changes. Standard dictionaries like the *Petit Robert* or the *Petit Larousse* first adopted a cautious attitude, the former eventually integrating a fair amount of the recommended changes, but keeping the traditional ('old') spelling as the primary access key and inserting a label "on écrirait mieux" (the following form would be better:) in front of the new spelling, at the end of the dictionary entry. In 2002, Hachette took a decisive step in fully integrating the spelling reform in its main series of monolingual dictionaries. Meanwhile, the new spelling was increasingly taught in French-speaking Belgium and, in March 2005, the Quebec government encouraged all its teachers to consider the 'new' spelling as valid, alongside the 'old' (traditional) spelling. This is in line with all the official recommendations, which stipulate that "aucune des deux graphies ne peut être tenue pour fautive» (Dictionnaire de l'Académie française (9e édition) dans les fascicules du Journal officiel, 22.05.93.). This explains why the default setting of the new speller is precisely the spelling configuration which accepts both the 'old' spelling and the 'new' one.

Note that the changes impact something like 2,000 words (which represents about 20,000 inflected forms). The following table gives a few examples:

Traditional ('old') spelling	'New' spelling
brûler	bruler
accroître	accroitre
aiguë	aigüe
ambiguë	ambigüe
apparaître	apparaitre
chaîne	chaine
contre-attaquer	contrattaquer
gèreras	gèreras
suggérerait	suggérerait
porte-monnaie	portemonnaie
penalties	pénaltys
ruisselle	ruissèle
whiskies	whiskys
matches	matchs

Table 1. Old vs. new spelling

As can be seen, the changes mainly concern the use of the circumflex accent, which disappears in a number of words like *connait, disparait, bruler, cout, enchainer*, the concatenation of some words which used to be hyphenated, the use of accents, or a number of irregular plural for loan-words which now behave like any other French word taking -s in the plural (*whiskys, matchs, gentlemans...*). The use of accents also reflects the real pronunciation (which is why it is now recommended to write *gèreras* or *opèrerai*, with a grave accent instead of the acute accent of *géreras, opérerai*). The changes are described in detail in the 6/12/1990 French Official Journal.

Spelling reforms can have different philosophies. The spelling reform which the German language underwent a few years ago, for instance, is an "either-or" decision (users decide to apply the new spelling or to stick to the old one). None of the official texts stipulates that the old spelling and the new one can co-exist in the same text. This is why spell-checkers which support the German spelling reform offer two options, allowing users to apply either flavor (pre- or post-reform spelling). In the Netherlands, a Dutch spelling reform was promulgated by the main linguistic authority, the *Nederlandse Taalunie*, in October 2005. This spelling reform also has serious implications for dictionary publishers since the 'old' spelling became invalid as soon as the reform was enforced and only the new spelling can be accepted by teachers. Reference works then need to be updated as soon as possible and proofing tools need to be adapted so that a number of words get flagged as invalid, even if, until October 15, 2005, these words were considered perfectly correct. Unlike the German spelling reform, the Dutch reform forces those who develop proofing tools to remove pre-reform spellings from their lexicons, since no option allows users to go on using the 'old' spelling. The French spelling reform allows much more flexibility since all the linguistic authorities agree that the old forms and the new forms should be considered as valid and this, as we are going to see, impacts the way dictionaries are compiled as well as the information which is provided for each lexical entry. In the new spell-checker, a simple dialog box with three options is put at the users' disposal, to enable them to select one of the following options:

(a) apply only the traditional ('old') spelling (i.e. 'new' forms will be red-squiggled)
(b) apply the 'new' (rectified) spelling only (i.e. the 'old' forms will be red-squiggled)
(c) consider the old and new forms as valid (which is the default option)

These options have forced the lexicographers to introduce additional information into the electronic dictionaries. Codes have been added to indicate whether a given word form is valid with the pre-reform setting, the post-reform setting or with the third setting allowing pre- and post-reform spellings. The task was complicated by the fact that the reform can impact a whole lemma or only some of its inflected forms. The word pair *nénuphar-nénufar* (water lily) is a case in point. The traditional spelling was *nénuphar.* The lexicographer had to create a new entry for the new spelling (*nénufar*) and to assign a code indicating that this lemma and all its inflected forms correspond to the new spelling. The lexicon includes the following information:

nénuphar	D0	(D0= code for the traditional, 'old' spelling)
nénufar	D1	(D1= code for the new spelling)

For a word like *whisky*, however, the situation is slightly more complex. The reform stipulates that *whisky* is now no longer considered as a loan word in French and that it now takes on the usual plural marker (+s), like any other regular noun in French (*maison – maisons*). The irregular form *whiskies* will therefore be marked as 'old' spelling in the dictionary. The singular word *whisky* will of course be valid in all cases (D0 and D1), which forces the lexicographer to assign the codes at the level of the inflected forms:

whisky	D0, D1
whiskies	D0
whiskys	D1

In other cases, new morphological paradigms had to be created to allow the generation of new forms. For instance, verbs like *céder, gérer, posséder* or *opérer*, which used to form their future or conditional tenses with an acute accent (*céderai, gérerait, posséderions, opérera*) now make use of a grave accent, which is in fact more consistent with the actual pronunciation of these verbs (*cèderai, gèrerait, possèderions, opèrera*). These forms are represented with the following codes in the French speller lexicon:

opérer	D0, D1
opérera	D0
opèrera	D1
opérons	D0, D1

3 Feminine job titles

Attitudes vis-à-vis feminine job titles differ widely in the French-speaking and the English-speaking world. Forms that are marked as masculine in English are more and more frequently replaced by gender-neutral forms to avoid any kind of sexist bias (*chairman* used to be the only form, then was progressively opposed to *chairwoman*; nowadays, *chair* is often preferred because it does not include any indication about the gender of the person – see Dumond, 2005). In French, one single form (usually a masculine form) used to be employed to refer to both male and female professions (*docteur, auteur, gouverneur...*). Quebec started using explicitly feminine forms in the 1980s, in an attempt to recognize women's professional status and their right to have jobs which until then were predominantly seen as male strongholds. The use of feminine suffixes like *-e* (*gouverneure, auteure, docteure, chercheure...*) or *-trice* (*thanatopractrice, factrice...*) has now become more and more frequent in the press (Dister, 2004) and official decrees in French-speaking Belgium even force government and other official publications to make use of these feminine forms (Commission de féminisation du Conseil supérieur de la langue française, 2005). Dister & Moreau (2006) have shown that language has evolved over the past 15 years: they examined a corpus of descriptions presenting candidates to the European Parliament and observed that the use of feminine job titles had increased significantly between the European elections of 1989 and those of 2004, whether in France or in French-speaking Belgium.

Here again, taking into account these feminine forms forced the lexicographers to add new morphological codes to existing entries to make sure that the new forms were appropriately generated. In the case of *écrivain, député* or *professeur*, for example, it was necessary to add a code indicating that these words belong to the morphological class of items that form their feminine form by adding the suffix *-e*, like "ami" (amie) (*professeure, écrivaine, députée*).

4 Relationship between the word-breaker and the contents of the dictionary

Traditionally, a spell-checker checks the spelling of forms appearing in a text against a list of valid forms included in its lexicon. Words that are not included in this lexicon are then usually considered as mistakes and flagged by the speller. This presupposes that the speller should know what a word is. This is the job of a "word-breaker", or tokenizer, whose aim is to break a text down into smaller units that can be consumed by an NLP system such as a spell-checker or a grammar checker. Although tokenization is frequently considered as a rather uninteresting task for the researcher (see also Grefenstette & Tapanainen 1994, Mikheev 2003, Fontenelle 2005a), it is a crucial process since it also determines what kind of "words" should go into the lexicon. Everyone will probably agree that elision in French, which is a very frequent phenomenon, should be seen as involving two distinct words, i.e. the elided determiner (or conjunction...) followed by a word beginning with a vowel (or a mute h) which has triggered this elision (*l'école l'* + *école*). This means that, in these cases, the apostrophe is considered as a breaking character, i.e. a character that signals that a boundary should be inserted immediately after it. For the lexicographer, this obviously means that the words *l'* and *école* should be distinct entries and that the dictionary should not include the string *l'école* (otherwise, it should also contain *d'école, qu'école*, etc and all the possible combinations of words starting with a vowel and one of the few (about 15) elided words containing an apostrophe (*l', d', m', s', n', t'*... – See Fontenelle, 2005a for more details about this oversimplification, since there are of course words in which the apostrophe is not a breaking character, *aujourd'hui* or *prud'homme* being cases in point).

Hyphens pose similar problems. They belong to the "word" in a significant amount of cases and the lexicon definitely needs to include hyphenated words like *après-midi, grand-mère, tire-bouchon,* or *rendez-vous* (at least when *rendez-vous* is a noun). But the hyphen is also used in productive deictic forms and with clitics in cases like "*cette maison-là*", "*ce livre-ci*", "*donne-moi la main*", "*prends-en encore deux*"... Even if one tried to predict all the possible combinations and to lexicalize them (all nouns can theoretically appear with *-là* or *-ci* in deictic constructions), the size of the lexicon would be unmanageable and the dictionary would contain about 15 million entries. Moreover, hyphens are also used to combine words from open classes in constructions like "*le match France-Espagne*" or "*les relations employeurs-employés*". It is clear that if the hyphen is considered as a non-breaking character, the full string *France-Espagne* will be emitted as one single token and it will need to be lexicalized if one wants the speller to consider it as a correct combination. This is clearly impossible (and undesirable) since it would entail adding to the lexicon gazillions of forms like *France-Belgique, France-Italie, France-Allemagne, Belgique-Italie, Belgique-Luxembourg,*

Belgique-Suisse, etc... Not lexicalizing these compounds is a solution adopted by some spell-checkers, but this means that users have to see a large amount of combinations that are then flagged by the tool: a large number of false flags can rapidly become annoying and real mistakes then tend to pass unnoticed by the user, even if the tool flags them.

In the new speller we developed, we considered the hyphen as a breaking character, but lexicalized all the forms that all linguists would recognize as one word (*tire-bouchon, grand-père, vis-à-vis, porte-avion, souffre-douleur, gagne-pain...*). Even if the speller is verifying individual tokens on either side of the hyphen, which makes it possible to recognize as a valid form strings like "*France-Belgique*", we still want to flag common mistakes like "*des tires-bouchons*" or "des *portes-avions*" (the first element of the compound, *tire-* and *porte-*, is invariable and the speller needs to flag erroneous plurals, even though, in isolation, *tires* and *portes* are perfectly valid forms). By being able to flag only the real mistakes and avoiding attracting the user's attention to valid and productive compounds, we managed to reduce the number of false flags by 74% compared to the preceding speller.

5 Injecting grammatical knowledge into the dictionary to improve suggestions

In this last section, I would like to describe an innovative way of integrating some grammatical information into the speller lexicon in order to make the suggestions more relevant. A spell-checker indeed has two main functions: it should spot mistakes, but it should also try to suggest the most likely word form to replace the erroneous input. Computing the suggestions is usually an algorithmic process based upon the concept of "edit distance" (Levenhstein, 1968), which measures the number of character manipulations that were necessary to turn a correct word into an incorrectly spelled one: deleting, adding, transposing or replacing a character are the most common manipulations (for instance, *information infomation* illustrates character deletion).

Because the linguistic intelligence of a spell-checker is limited, several suggestions are frequently offered when several valid words that are close to the input are identified in the lexicon. If the input is *ofer*, an English speller can suggest *offer* (because a character had been deleted) as well as *over* (because a character was replaced by another one), which are both one manipulation away from the erroneous input. It is usually up the user to decide on which form corresponds to what they actually meant.

A study we conducted on the most frequent spelling mistakes revealed that, in addition to the usual manipulations described above, some people tend to forget apostrophes, which results in an undesirable concatenation of two words. Re-inserting or restoring the apostrophe is therefore a phenomenon which seems to be typical of French (and probably other languages which make use of elided forms, like Italian). It is indeed necessary to flag the mistake in the following example and the speller should know that an apostrophe should be inserted after the first letter to generate the appropriate suggestion:

lécole → l'école

As we have seen above, the list of elided forms is limited to around 15 items (*d', m', j', n', t', l', qu'...*). It is however rather dangerous to insert an apostrophe anywhere as soon as an erroneous string of characters starts with one of these letters (*d, m, j, n, t, l, qu...*). If, for

example, the user writes *sallons*, it would seem perfectly logical to suggest the valid French words *salons*, *sablons* or *sillons*, since these words can be found at a very short edit distance from the input. Yet, if the speller inserted an apostrophe after the initial *s*, the suggestion would be *s'allons*: *s'* and *allons* are possible in isolation, but users would probably raise an eyebrow if they saw this type of suggestion because the pronoun *s'* is limited to a co-occurrence with 3rd person singular verbs. Since *allons* is a 1st person plural verb, suggesting *s'allons* is grammatically incorrect.

This type of constraint is grammatical, but the dictionary used by the speller includes information on person, number and part of speech. Since our speller 'knows' that *allons* is a 1st person plural verb, we have implemented a set of constraints based upon the presence of this information in the dictionary, which makes it possible to limit the insertion of apostrophes to well-defined sets of cases. Suggestions that would be considered as ridiculous can then be avoided. In the new version of the speller, we specified for example that *m'*, *t'* or *s'* must be followed by a verb or a pronoun starting with a vowel, but never by a noun. So if the input is:

Pierre regarde le simages.

the word *simages* will be squiggled and only *images* will be offered as a suggestion. The speller does not try to suggest *s'images* since *s'* can only be followed by a verb or a pronoun and *images* is only assigned the Noun part of speech in the lexicon). When the user selects *images*, the grammar checker fires and corrects the agreement mistake (*le images les images*).

Injecting this kind of «linguistic intelligence» into a speller lexicon takes this proofing tool one step closer to a grammar checker and improves the overall proofing experience.

6 Conclusion

We have described some of the features of the new French spell-checker made available to Microsoft Office users. Some of the problems faced by the lexicographers who compile the lexicon for such a tool have been described and discussed, ranging from decisions as to what should be lexicalized, to the problems posed by the need to offer several spelling reform settings. The relationship between the underlying word-breaker, which produces the tokens consumed by the speller, and the types of lexical items that should be stored in the lexicon, has also been discussed, as well as the integration of richer information into the lexicon, which gradually tends to blur the traditional distinction between spell-checker and grammar checker.

References

Cerquiglini, B. (réd.) (1999), *Femme, j'écris ton nom... Guide d'aide à la féminisation des noms de métiers, titres, grades et fonctions*. CNRS. INALF.

Commission de féminisation du Conseil supérieur de la langue française. (2005), *Mettre au féminin. Guide de féminisation des noms de métier, fonction, grade ou titre*. Ministère de la Communauté française de Belgique. Bruxelles. 2ème édition.

Conseil Supérieur de la Langue Française, 'Les Rectifications de l'orthographe', *Journal Officiel de la République Française*. N°100, 6 décembre 1990.

Dister, A. (2004), 'La féminisation des noms de métier, fonction, grade ou titre en Belgique francophone. État des lieux dans un corpus de presse', in Purnelle, G., Fairon, C., Dister, A. (réd.) *Actes des*

JADT 2004 – 7èmes Journées internationales d'Analyse statistique des Données Textuelles, Presses universitaires de Louvain, pp. 313-324.

Dister, A., Moreau, M.-L. (2006), 'Dis-moi comment tu féminises, je te dirai pour qui tu votes – Les dénominations des candidates dans les élections européennes de 1989 et de 2004 en Belgique et en France', in *Langage et Société*, March 2006.

Dumond, Val (2005), *Just Words – The US and THEM thing: A Guide to inclusive spoken and written English*, Muddy Puddle Press.

Fontenelle, Th. (2004), 'Lexicalization for proofing tools' in Williams, G. & Vessier, S. (eds) *EURALEX 2004 Proceedings (11th EURALEX International Congress)*, Université de Bretagne-Sud, Lorient, pp. 79-86.

Fontenelle, Th. (2005a), 'Identifying tokens: Is word-breaking so easy?', in Hiligsmann, Ph., Janssens, G., & Vromans, J. (eds.) *Woord voor woord. Zin voor zin. Liber Amicorum voor Siegfried Theissen*, Ghent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde, pp. 109-115.

Fontenelle, Th. (2005b), 'Dictionnaires et outils de correction linguistique', in Fontenelle, Th. (coord.) *Revue Française de Linguistique Appliquée*, numéro spécial sur les «Dictionnaires: nouvelles approches, nouveaux modèles». 2005, Vol. X-2, pp. 119-128.

Fontenelle, Th. (2006), 'Les nouveaux outils de correction linguistique de Microsoft', in Fairon, C. & Mertens, P. (eds) *Actes de TALN 2006*, Louvain/Leuven.

Grefenstette, Gregory & Tapanainen, Patsy (1994), 'What is a word, what is a sentence? Problems of tokenization', *Proceedings of the 3rd Conference on Computational Lexicography and Text Research – COMPLEX'94*, Budapest, Hungary, 7-10 July 1994, pp. 79-87.

Levenshtein, V. (1965), 'Binary Codes Capable of Correcting Deletions, Insertions and Reversals'. 707/709, *Soviet Physics Doklady* 10.

Mikheev, A. (2003), 'Text Segmentation', in Mitkov, R., (ed.), *The Oxford Handbook of Computational Linguistics*, OUP, pp. 201-218.