Lexicalization for Proofing Tools

Thierry Fontenelle
Microsoft Corporation
Natural Language Group
One Microsoft Way
Redmond, 98052
USA
Email: thierryf@microsoft.com

Abstract
This paper discusses some of the crucial issues related to the construction of word lists for proofing tools such as spelling checkers. The more specific problem of derivational morphology is described and the paper argues that efforts made over the past few years to build generative morphology components have mainly been driven by the need to recognize neologisms and novel forms in an analysis perspective. The inevitable over-generation which has been neglected on too many occasions forces computational linguists and lexicographers to resort to hybrid methods when constructing their lexicons for spell-checkers, which cannot afford mis-analyzing erroneous input or suggesting aberrant forms.

1. Introduction: Spell-Checking - The Lexicographer's Nightmare

Lexicographers who compile a learner's or a collegiate dictionary know they have to respect very strict guidelines as to the maximum number of entries they are allowed to compile, since the number of pages is normally determined at the beginning of the project. Users do not expect to find the latest neologism or a rare or archaic term in a learner's dictionary. This situation is slightly different with tools such as spell checkers, since a word is normally going to be flagged if it is not included in the electronic dictionary used by the speller (see the red squiggles in the Microsoft Office System programs such as Word). Since users are by definition very demanding, the temptation is great to include as many lexical items as possible. The coverage of a spell checker also usually includes frequent named entities (city names, country names, frequent first names, famous people's names...), which means that the dictionary must be continuously maintained and extended. With time, however, users have come to admit that no spell checker can include all possible words in a language and that a certain percentage of "red squiggles" is both inevitable and acceptable.

Lexicographers who create a lexicon for a spell-checker are faced with a series of problems. First of all, the initial word list must be extended with the help of large corpora and powerful tools to exploit them. Frequency information is obviously a key criterion (though not the only one) to decide whether a word should be included or not. Another aspect which should be borne in mind is that spell-checking covers both analysis (recognizing whether a given form is valid or not) and generation (whenever possible, a spell-checker normally offers one or more suggestions meant to replace an erroneous word). Two options at least are possible: either the lexicon contains a list of lemmas and a list of possible affixes, with rules describing how the former and the latter can be combined, or it contains a pre-compiled list of lemmas and their inflected forms, which the speller uses as a static resource against which an input text can be checked. The former method rests upon advances in what has come to be known as generative morphology, to which we would now like to turn our attention.

2. Computational Derivational Morphology

Derivational morphology has been a hot topic in traditional linguistics (Corbin 1987; de Caluwe & Taeldeman 2003) and in computational linguistics (Gaussier 1999; Sproat 1992) for a number of years. The emphasis has generally been laid on analysis, with the professed aim to develop robust natural language processing (NLP) systems which can recognize a wide range of morphological phenomena, while limiting the size of the lexicon (a constant goal among computational linguists). NLP researchers have proposed a number of techniques and formalisms which can cope with users' natural creativity while recognizing that it is impossible to identify and record all the possible words that can be produced in a given language (a task which could be compared to the Myth of Sisyphus since new words are created every day and users resort to powerful derivational mechanisms such as prefixation and suffixation to create these neologisms, the majority of which instantly fall into oblivion and only a small percentage of which will eventually make their way into traditional dictionaries).

Most linguists would probably agree that an NLP lexicon need not contain all possible words starting with the prefix *hyper-* and that rules such as *hyper* + N → N or *hyper* + Adj → Adj would be sufficient to analyze neologisms as in the following sentences extracted from our French corpus[1]:

Au cours d'une période d'	hyperinflation	la monnaie et le crédit s'accroissent à un rythme exponentiel, détruisant tous les liens existant entre valeur réelle et valeur nom
Ce phénomène s'explique par une	hyperpolarisation	qui empêche le neurone postsynaptique de générer un potentiel d'action.
Les déserts	hyperarides	ou absolus sont les plus rares: ils couvrent moins de 6millions de km2.

Provided the lexicon already contains information on the adjectives *aride* or the nouns *inflation* and *polarisation*, it should therefore not be necessary to include *hyperpolarisation*, *hyperinflation* or *hyperaride*, since the derived words share common syntactic and morphological properties with their base. This view is heavily influenced by the analysis perspective adopted by most computational linguists who are anxious not to miss any novel usage.

This approach is used for the various parsers used by the Microsoft grammar checkers. The following is an example of a morphological rule which, given an unknown input ending in *-el*, makes it possible to analyze it as an adjective while recognizing the base noun from which it is derived:

```
{Word      "DER_Adj_el"
 Senses
  {Bits     J_al Masc Skip Sing
   SenseNo  100
   Cat      "Adj"
   Morph_op  "M_any cultur el -> cultur e ;"
             "M_any professionn el -> profession ;"
             "M_any déme ntiel -> déme nce ;"
   Postype  ADJ
   Infl     "Adj-cruel"
   NextMorphemes (NONE INFL_adj_fem INFL_adj_masc_plur
INFL_adj_fem_plur)
   PrevCat  NOUN
   PrevMorphemes (Stem) }
 }
```

Figure 1: Derivational rule of adjectives in *-el*

The rule, called DER_Adj_el, applies to the stem (PrevMorpheme) of a Noun (PrevCat = category of the previous morpheme) and generates an adjective which has the same inflectional paradigm (Infl) as the adjective *cruel* (i.e. the feminine singular form has a double *ll+e*, as in *cruelle*...). A list of possible morphemes which can be attached to this adjective is given in the NextMorpheme field, which accounts for the various forms of the –*el* adjective, e.g. *cruel, cruelle, cruels, cruelles*...

This rule is powerful enough to correctly analyze *prudentiel* (< prudence), *superficiel* (< superficie), *carentiel* (< carence), *possessionnel* (< possession), or *jurisprudentiel* (< jurisprudence). However, it is too powerful insofar as it incorrectly analyzes the string "viel" as *vie* (life) + *-el* in the following sentences extracted from our unedited French corpus[2]:

Il était une fois, il y a très longtemps, vivait dans un petit village caché au fond de la forêt, un	viel	homme et son fils.
On y voit un	viel	homme, chef du village, qui refuse de laisser partir le sel a dos de yacks car la coutume veut que ce transport se fasse dans des conditions bien precises, et avec la faveur des dieux.
Mais malgrés ce	viel	adage .
Ou alors faut-ol que je copie un	viel	MSDos 6.20 sur mon disque dur ?

The risk is clear here: using a generative morphology rule such as the one illustrated above to produce a run-time analysis of the input text would force the spell-checker to accept a spelling mistake as valid and possibly to generate erroneous suggestions ("*viel*" is here a typo and should in all cases be replaced by *vieil* (old)). Such over-generation must absolutely be avoided and a trade-off must be sought between robust analysis techniques and the need to be linguistically accurate.

3. A Hybrid Approach

The compilation of a lexicon is only one facet of the development of a spell-checker. To make sure mistakes can be identified and corrected, it is also essential to have a typology of errors in order to determine which patterns of mistakes are frequent and how certain strings of letters should be rewritten if the input word is not present in the lexicon. Such a typology of errors would for instance reveal that, in order to find the correct spelling of a misspelled word ending in *–ction*, priority should be given to a rule rewriting *–ction* into *–xion* (connection → connexion).[3]

The dictionary-building approach we adopted to compile a spell-checker's lexicon is a hybrid method combining a generative mechanism to help identify potentially interesting derived forms, and extensive corpus analysis whose primary aim is to sift through the output of the generative morphology analyses. The lexicographer's judgment is essential in this process since the identification and rejection of erroneous forms is a sine qua non. The resulting lexicon would otherwise be populated with aberrant forms which would never be flagged (squiggled) or would be suggested as possible replacement forms. The correct words which pass the frequency and relevance tests are then lexicalized, i.e. integrated into the static resource which will be used by the spell-checker.[4]

Besides the obvious practical applications described here, viz. the compilation of an electronic word list for a spell-checker, such an approach also helps shed light on the types of constraints which should be taken into account when implementing or refining generative morphology components. The following examples illustrate some of the constraints which must be imposed to limit the power of such a component and reduce over-generation.

3.1. Don't let any morphology rule apply to one-letter words

Ignoring this rule would be very dangerous since a speller could recognize the following erroneous forms and fail to flag them:

extrai ← extra + i (frequent mistake for *extrait* or *extrais*, where the final letter is omitted by many people)

C'est avec une grande difficulté que je m'	extrai	de mon lit.

The letters of the alphabet are granted entry status and are assigned the Noun part-of-speech. It is therefore essential to block the rule which makes it possible to combine *extra* with nouns to generate other nouns (as in *extraterritorialité…*).

3.2. Do not let morphology rules apply to monosyllabic words

If highly productive suffixes such as *co-* or *per-*, which frequently combine with a very wide range of nouns, were allowed to combine with monosyllabic words, the following forms, attested in our unedited corpus, would be considered as valid:

conu ← co + nu (lit. 'co-nude'; frequent mistake for *connu* – known)
transfer ← trans + fer (lit. 'trans-iron'; frequent mistake for *transfert*)
permi ← per + mi (lit. 'per + musical note *mi*'; frequent mistake for *permis* or *permit*)

Comme son nom l'indique, Allocine n'est pas né sur Internet mais le web lui a	permi	d'élargir considérablement ses activités.
Si oui, m'est-il	permi	d'utiliser et de modifier ces images comme je le souhaite.
Leur habileté leur a	permi	de redresser le pays tout en conservant le fruit de plusieurs années de lutte pour la Révolution.
Implant Soulaines, à une cinquantaine de kilomètres à l'est de Troyes, ce centre est	conu	pour recevoir, jusqu'en 2025, un million de mètres cubes de dchets faiblement et moyennement radioactifs
Quelle est la vitesse approximative en Meg à la minute d'un	transfer	de données ?

3.3. Don't let morphology rules apply to words which start with a consonant if the same string with a double consonant is already lexicalized

This is based upon the frequent occurrences of a single consonant instead of a double consonant in our typology of errors. The following are cases of aberrant forms which would otherwise be allowed, given the fact that a general rule such as *co-* + Noun/Verb is very productive in French (*coactionnaire, coauteur, codemandeur, coscénariste, coproducteur, cofinancer, coexister, coprésider, cofonder...*):
comission ← co + mission (instead of *commission*)
comettre ← co + mettre (instead of *commettre*)

La	Comission	dispose d'un mois pour donner son aval ou ouvrir une enquête plus approfondie.
Ne peuvent ils jamais reflechir avant de	comettre	des actes de ce genre.

3.4. Do not let rules apply to vulgar, slang, or archaic words

Some rules such as 'privative *a-* + Noun/Adj', which generate learned words, should not be allowed to fire if they apply to vulgar or slang terms. The following is a case in point:
avit[5] ← a + vit (frequent spelling mistake for *avait* – had)

les gens peuvent vraiment voir de quoi ca	avit	l'air.
Mais, mai 68 l'	avit	stoppé dans son élan.

In fact, the rule '*a* + Noun/Adj → Noun/Adj' should only be allowed to apply to "learned" words with a Greek or Latin origin (*alogique, achromique, anorganique, anionique...*). This presupposes that the concept of "learned" base or stem be available in the lexicon, and that register level be coded in the lexical database to mark vulgar or slang terms.

This seems to be a sine qua non to investigate the productive nature of some rules, such as those analyzed by Aliquot-Suengas (2003:44), who points out that the French suffix *–ade* is frequently used to denote culinary preparations, but is never attached to learned bases.

3.5. Do not let rules apply if the input string is the normalized, unaccented version of an already existing word

Our typology of French spelling mistakes indicates that missing accents account for a high percentage of errors. In the following example, incorrectly analyzed as the prefix *auto-* combined with the noun *rite*, the final 'e' has lost its accent (a phenomenon frequently observed in newsgroups):
autorite ← auto + rite (instead of *autorité* – authority)

Et le Siège étant depuis le 13eme siècle le nom qui designe la place ou le juge s'assied et par consequent étant l'image de son	autorite.
Ils n'avaient pas accorde toute l'attention necessaire au fonctionnement des commissions de securite dependant de leur	autorite.

4. Constraining Derivation

As we have seen above, it is essential to constrain the rules which enable the NLP system to generate the forms which can be analyzed by the speller, or which can appear in the list of suggestions offered to the user. In addition to register information, word length or number of syllables, purely semantic information should also ideally be used to block undesirable derivations. As pointed out by Anderson (1988:149), whose remark is reproduced by Kerkleroux (2003:22), "lexical rules have access to the thematic relations associated with particular arguments". For instance, Kerkleroux shows that deverbal nouns in *–eur* require that the verbal base include a proto-agent-like argument. Corbin (1987) has also shown that adjectives in *–eux* refer to a property seen as endogenous, i.e. caused by factors which are internal to the entity denoted by the base noun. Namer (2003) points out that such adjectives are semantically incompatible with the suffix *–is(er),* which generates change-of-state verbs. Adjectives which can combine with *–is(er)* must express a property which can be acquired (change of state), which excludes adjectives in *–eux*, since the latter are based upon an endogenous property (*boutonneux, venteux*). One immediately sees the impact this can have upon the breadth and depth of the lexical-semantic description of lexical items in a large coverage lexicon if one wishes to develop robust, efficient and precise spell-checkers.

5. Conclusion

I have described some of the pitfalls faced by lexicographers who compile word lists for a spell- checker and wish to extend the coverage of their lexicons. Generative morphology components in NLP systems are too often marred with over-generation problems because their development has generally focused on the recognition of novel senses and neologisms in robust parsers. The special nature of spell-checkers, which cannot afford accepting erroneous forms or suggesting aberrant words, forces the lexicographer to adopt hybrid methods, based upon automatic analysis of derived forms combined with extensive corpus analysis to validate or reject forms recognized by these automatic procedures. Such a hybrid method sheds new light onto the types of constraints morphological systems should take into account to limit over-generation.

Endnotes

[1] The corpus data we use for our various NLP projects include a wide gamut of data, ranging from newspaper data to transcribed conversations to business letters or newsgroup data, as well as email messages, academic textbooks or texts from the Encarta Encyclopedia.

[2] It is clear that the corpus used to acquire new words for a spelling checker cannot be edited. Since editing is usually done with the help of proofing tools, the risk would be big of modifying source data and biasing it with respect to a pre-existing spell-checker, by selecting only forms which are already included in this tool.

[3] Establishing the list of suggestions which are offered to the user to correct a misspelling is usually done algorithmically on the basis of a variant of Levenshtein's algorithm (Levenshtein 1965). This technique is based upon the concept of edit distance and computes the shortest path from the misspelling to a correct form found in the lexicon, the aim being to suggest the closest words, i.e. the words which can be retrieved by applying the smallest number of manipulations (insertion or deletion of characters, permutation of two letters...).These general principles are generally augmented with language-specific knowledge derived from the error typology (e.g. a transition such as *–als* → *-aux* makes it possible to suggest *chevaux* when the input is *chevals*, a frequent mistake made by non-native speakers).

[4] Corpora of several dozen million words are used to extract such data and a frequency threshold must be used in order to exclude extremely rare occurrences which would unnecessarily inflate the size of the lexicon or could mask frequent spelling mistakes. The lexicographer also has to determine whether a word is relevant in such a lexicon since some frequent forms may be banned from the lexicon because of prescriptive legislation or geo-political sensitive issues.

[5] *Vit* is a vulgar and archaic term in French for *penis.*

References

Adams, Valerie. 1973. *An Introduction to Modern English Word-Formation.* Longman. London and New York.

Aliquot-Suengas, Sophie. 2003. 'La productivité actuelle de la forme constructionnelle –ade', in *Langue Française*, Dal, G. (réd.), La productivité morphologique en questions et en expérimentations. Larousse. Décembre 2003. 38-55.

Anderson, S.T. 1988. 'Morphological Theory' in F. Newmeyer (ed.), *Linguistics: The Cambridge Survey.* Cambridge University Press. 146-191.

Corbin, Danielle. 1987. *Morphologie dérivationnelle et structuration du lexique.* Tübingen. Max Niemeyer Verlag.

Dal, Georgette. 2003. 'Productivité morphologique : définitions et notions connexes', in *Langue Française*, Dal, G. (réd.), La productivité morphologique en questions et en expérimentations. Larousse. Décembre 2003. 3-23.

De Caluwe, Johan & Taeldeman, Johan. 2003. 'Morphology in dictionaries', in P. van Sterkenburg (ed.) *A Practical Guide to Lexicography.* John Benjamins. Amsterdam/Philadelphia. 114-126.

Gaussier, Eric. 1999. 'Unsupervised learning of derivational morphology from inflectional lexicons', *in* Proceedings of the Workshop on Unsupervised Methods in Natural

Language Processing. *Association for Computational Linguistics. ACL'99. Univ. of Maryland, USA.*

Kerkleroux, Françoise. 2003. 'Morpho-logie: La Forme et l'Intelligible', in *Langages.* N°152. 12-32.

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

Namer, Fiammetta. 2003. 'Productivité morphologique, représentativité et complexité de la base : Le système MoQuête', in *Langue Française,* Dal, G. (réd.), La productivité morphologique en questions et en expérimentations. Larousse. Décembre 2003. 79-101.

Sproat, Richard. 1992. *Morphology and Computation.* Cambridge. MIT Press.