Proceedings of the Workshop on Natural Language Processing in
the Biomedical Domain, Philadelphia, July 2002, pp. 21-28.
Association for Computational Linguistics.

Accenting unknown words in a specialized language

Pierre Zweigenbaum and **Natalia Grabar**
DIAM — STIM/DSI, Assistance Publique – Hôpitaux de Paris
& Département de Biomathématiques, Université Paris 6
{ngr,pz}@biomath.jussieu.fr

Abstract

We propose two internal methods for accenting unknown words, which both learn on a reference set of accented words the contexts of occurrence of the various accented forms of a given letter. One method is adapted from POS tagging, the other is based on finite state transducers.

We show experimental results for letter *e* on the French version of the Medical Subject Headings thesaurus. With the best training set, the tagging method obtains a precision-recall breakeven point of 84.2±4.4% and the transducer method 83.8±4.5% (with a baseline at 64%) for the unknown words that contain this letter. A consensus combination of both increases precision to 92.0±3.7% with a recall of 75%. We perform an error analysis and discuss further steps that might help improve over the current performance.

1 Introduction

The ISO-latin family, Unicode or the Universal Character Set have been around for some time now. They cater, among other things, for letters which can bear different diacritic marks. For instance, French uses four accented *e*s (*éèêë*) besides the unaccented form *e*. Some of these accented forms correspond to phonemic differences. The correct handling of such accented letters, beyond US ASCII, has not been immediate and general. Although suitable character encodings are widely available and used, some texts or terminologies are still, for historical reasons, written with unaccented letters. For instance, in the French version of the US National Library of Medicine's Medical Subject Headings thesaurus (MeSH, (INS, 2000)), all the terms are written in unaccented uppercase letters. This causes difficulties when these terms are used in Natural Language interfaces or for automatically indexing textual documents: a given unaccented word may match several words, giving rise to spurious ambiguities such as, *e.g.*, *marche* matching both the unaccented *marche* (*walking*) and the accented *marché* (*market*).

Removing all diacritics would simplify matching, but would increase ambiguity, which is already pervasive enough in natural language processing systems. Another of our aims, besides, is to build language resources (lexicons, morphological knowledge bases, etc.) for the medical domain (Zweigenbaum, 2001) and to learn linguistic knowledge from terminologies and corpora (Grabar and Zweigenbaum, 2000), including the MeSH. We would rather work, then, with linguistically sound data in the first place.

We therefore endeavored to produce an accented version of the French MeSH. This thesaurus includes 19,971 terms and 9,151 synonyms, with 21,475 different word forms. Human reaccentuation of the full thesaurus is a time-consuming, error-prone task. As in other instances of preparation of linguistic resources, *e.g.*, part-of-speech-tagged corpora or treebanks, it is generally more efficient for a human to correct a first annotation than to produce it from scratch. This can also help obtain better consistency over volumes of data. The issue is then to find a method for (semi-)automatic accentuation.

The CISMeF team of the Rouen University Hos-

pital already accented some 5,500 MeSH terms that are used as index terms in the CISMeF online catalog of French-language medical Internet sites (Darmoni et al., 2000) (www.chu-rouen.fr/cismef). This first means that less material has to be reaccented. Second, this accented portion of the MeSH might be usable as training material for a learning procedure.

However, the methods we found in the literature do not address the case of 'unknown' words, *i.e.*, words that are not found in the lexicon used by the accenting system. Despite the recourse to both general and specialized lexicons, a large number of the MeSH words are in this case, for instance those in table 1. One can argue indeed that the compilation of a larger lexicon should reduce the proportion of unknown words. But these are for the most part specialized, rare words, some of which we did not find even in a large reference medical dictionary (Garnier and Delamare, 1992). It is then reasonable to try to accentuate automatically these unknown words to help human domain experts perform faster post-editing. Moreover, an automatic accentuation method will be reusable for other unaccented textual resources. For instance, the ADM (Medical Diagnosis Aid) knowledge base online at Rennes University (Seka et al., 1997) is another large resource which is still in unaccented uppercase format.

cryomicroscopie	*dactylolyse*
decarboxylases	*decoquinate*
denitrificans	*deoxyribonuclease*
desmodonte	*desoxyadrenaline*
dextranase	*dichlorobenzidine*
dicrocoeliose	*diiodotyrosine*
dimethylamino	*dimethylcysteine*
dioctophymatoidea	*diosgenine*

Table 1: Unaccented words not in lexicon.

We first review existing methods (section 2). We then present two trainable accenting methods (section 3), one adapted from part-of-speech tagging, the other based on finite-state transducers. We show experimental results for letter *e* on the French MeSH (section 4) with both methods and their combination. We finally discuss these results (section 5) and conclude on further research directions.

2 Background

Previous work has addressed text accentuation, with an emphasis on the cases where all possible words are assumed to be known (listed in a lexicon). The issue in that case is to disambiguate unaccented words when they match several possible accented word forms in the lexicon – the *marche*/*marché* examples in the introduction.

Yarowsky (1999) addresses accent restoration in Spanish and in French, and notes that they can be linked to part-of-speech ambiguities and to semantic ambiguities which context can help to resolve. He proposes three methods to handle these: N-gram tagging, Bayesian classification and decision lists, which obtain the best results. These methods rely either on full words, on word suffixes or on parts-of-speech. They are tested on 'the most problematic cases of each ambiguity type', extracted from the Spanish AP Newswire. The agreement with human accented words reaches 78.4–98.4% depending on ambiguity type.

Spriet and El-Bèze (1997) use an N-gram model on parts-of-speech. They evaluate this method on a 19,000 word test corpus consisting of news articles and obtain a 99.31% accuracy. In this corpus, only 2.6% of the words were unknown, among which 89.5% did not need accents. The resulting error rate (0.3%) accounts for nearly one half of the total error rate, but is so small that it is not worth trying to guess accentuation for unknown words.

The same kind of approach is used in project RÉACC (Simard, 1998). Here again, unknown words are left untouched, and account for one fourth of the errors. We typed the words in table 1 through the demonstration interface of RÉACC online at www-rali.iro.umontreal.ca/Reacc/: none of these words was accented by the system (7 out of 16 do need accentuation).

When the unaccented words are in the lexicon, the problem can also be addressed as a spelling correction task, using methods such as string edit distance (Levenshtein, 1966), possibly combined with the previous approach (Ruch et al., 2001).

However, these methods have limited power when a word is not in the lexicon. At best, they might say something about accented letters in grammatical affixes which mark contextual, syntactic constraints.

We found no specific reference about the accentuation of such 'unknown' words: a method that, when a word is not listed in the lexicon, proposes an accented version of that word. Indeed, in the above works, the proportion of unknown words is too small for specific steps to be taken to handle them. The situation is quite different in our case, where about one fourth of the words are 'unknown'. Moreover, contextual clues are scarce in our short, often ungrammatical terms.

We took obvious measures to reduce the number of unknown words: we filtered out the words that can be found in accented lexicons and corpora. But this technique is limited by the size of the corpus that would be necessary for such 'rare' words to occur, and by the lack of availability of specialized French lexicons for the medical domain.

We then designed two methods that can learn accenting rules for the remaining unknown words: (i) adapting a POS-tagging method (Brill, 1995) (section 3.3); (ii) adapting a method designed for learning morphological rules (Theron and Cloete, 1997) (section 3.4).

3 Accenting unknown words

3.1 Filtering out know words

The French MeSH was briefly presented in the introduction; we work with the 2001 version. The part which was accented and converted into mixed case by the CISMeF team is that of November 2001. As more resources are added to CISMeF on a regular basis, a larger number of these accented terms must now be available. The list of word forms that occur in these accented terms serves as our base lexicon (4861 word forms). We removed from this list the 'words' that contain numbers, those that are shorter than 3 characters (abbreviations), and converted them in lower case. The resulting lexicon includes 4054 words (4047 once unaccented). This lexicon deals with single words. It does not try to register complex terms such as *myocardial infarction*, but instead breaks them into the two words *myocardial* and *infarction*.

A word is considered unknown when it is not listed in our lexicon. A first concern is to filter out from subsequent processing words that can be found in larger lexicons. The question is then to find suitable sources of additional words.

We used various specialized word lists found on the Web (lexicon on cancer, general medical lexicon) and the ABU lexicon (abu.cnam.fr/DICO), which contains some 300,000 entries for 'general' French. Several corpora provided accented sources for extending this lexicon with some medical words (cardiology, haematology, intensive care, drawn from the current state of the CLEF corpus (Habert et al., 2001), and drug monographs). We also used a word list extracted from the French versions of two other medical terminologies: the International Classification of Diseases (ICD-10) and the Microglossary for Pathology of the Systematized Nomenclature of Medicine (SNOMED). This word list contains 8874 different word forms. The total number of word forms of the final word list was 276 445.

After application of this list to the MeSH, 7407 words were still not recognized. We converted these words to lower case, removed those that did not include the letter *e*, were shorter than 3 letters (mainly acronyms) or contained numbers. The remaining 5188 words, among which those listed in table 1, were submitted to the following procedure.

3.2 Representing the context of a letter

The underlying hypotheses of this method are that sufficiently regular rules determine, for most words, which letters are accented, and that the context of occurrence of a letter (its neighboring letters) is a good basis for making accentuation decisions. We attempted to compile these rules by observing the occurrences of *eéèêë* in a reference list of words (the *training set*, for instance, the part of the French MeSH accented by the CISMeF team). In the following, we shall call *pivot letter* a letter that is part of the *confusion set eéèêë* (set of letters to discriminate).

An issue is then to find a suitable description of the context of a pivot letter in a word, for instance the letter *é* in *excisée*. We explored and compared two different representation schemes, which underlie two accentuation methods.

3.3 Accentuation as contextual tagging

This first method is based on the use of a part-of-speech tagger: Brill's (1995) tagger. We consider

each word as a 'string of letters': each letter makes one word, and the sequence of letters of a word makes a sentence. The 'tag' of a letter is the expected accented form of this letter (or the same letter if it is not accented). For instance, for the word *endometre* (*endometer*), to be accented as *endomètre*, the 'tagged sentence' is *e/e n/n d/d o/o m/m e/è t/t r/r e/e* (in the format of Brill's tagger). The regular procedure of the tagger then learns contextual accentuation rules, the first of which are shown on table 2.

Brill Format	Gloss
(1) e é NEXT2TAG i	*e.i* ⇒ e → é
(2) e é NEXT1OR2TAG o	*e.?o* ⇒ e → é
(3) e é NEXT1OR2TAG a	*e.?a* ⇒ e → é
(4) e é NEXT1OR2WD e	*e.?e* ⇒ e → é
(5) e é NEXT2TAG h	*e.h* ⇒ e → é
(6) é è NEXTBIGRAM n e	*éne* ⇒ é → è
(7) é e NEXTBIGRAM m e	*éme* ⇒ é → e
(8) e é NEXTBIGRAM t r	*etr* ⇒ e → é
(9) é e NEXT1OR2OR3TAG x	*é.?.?x* ⇒ é → e
(10) e é NEXT1OR2TAG y	*e.?y* ⇒ e → é
(11) e é NEXT2TAG u	*e.u* ⇒ e → é
(12) e é SURROUNDTAG t i	*tei* ⇒ e → é
(13) é è NEXTBIGRAM s e	*ése* ⇒ é → è

Table 2: Accentuation correction rules, of the form 'change t_1 to t_2 if *test* true on x [y]'. NEXT2TAG = second next tag, NEXT1OR2TAG = one of next 2 tags, NEXTBIGRAM = next 2 words, NEXT1OR2OR3TAG = one of next 3 tags, SURROUNDTAG = previous and next tags,

Given a new 'sentence', Brill's tagger first assigns each 'word' its mots frequent tag: this consists in accenting no *e*. The contextual rules are then applied and successively correct the current accentuation. For instance, when accenting the word *flexion*, rule (1) first applies (*if* e *with second next tag* = i, *change to* é) and accentuates the *e* to yield *fléxion* (as in ...*émie*). Rule (9) applies next (*if* é *with one of next three tags* = x, *change to* e) to correct this accentuation before an *x*, which finally results in *flexion*. These rules correspond to representations of the contexts of occurrence of a letter. This representation is mixed (left and right contexts can be combined, *e.g.*, in SURROUNDTAG, where both immediate left and right tags are examined), and can extend to a distance of three letters left and right, but in restricted combinations.

3.4 Mixed context representation

The 'mixed context' representation used by Theron and Cloete (1997) folds the letters of a word around a pivot letter: it enumerates alternately the next letter on the right then on the left, until it reaches the word boundaries, which are marked with special symbols (here, ^ for start of word, and $ for end of word). Theron & Cloete additionally repeat an out-of-bounds symbol outside the word, whereas we dispense with these marks. For instance, the first *e* in *excisée* (*excised*) is represented as the mixed context in the right column of the first row of table 3. The left column shows the order in which the letters of the word are enumerated. The next two rows explain the mixed context representations for the two other *e*s in the word. This representation

Word	Mixed Context=Output
^ e x c i s é e $ 2 . 1 3 4 5 6 7 8	x ^ c i s e e $=e
^ e x c i s é e $ 8 7 6 5 4 2 . 1 3	e s $ i c x e ^=é
^ e x c i s é e $ 8 7 6 5 4 3 2 . 1	$ e s i c x e ^=e

Table 3: Mixed context representations.

caters for contexts of different sizes and facilitates their comparison.

Each of these contexts is unaccented (it is meant to be matched with representations of unaccented words) and the original form of the pivot letter is associated to the context as an output (we use the symbol '=' to mark this output). Each context is thus converted into a transducer: the input tape is the mixed context of a pivot letter, and the output tape is the appropriate letter in the confusion set *eéèêë*.

The next step is to determine minimal discriminating contexts (figure 1). To obtain them, we join all these transducers (OR operator) by factoring their common prefixes as a *trie* structure, *i.e.*, a deterministic transducer that exactly represents the training set. We then compute, for each state of this transducer and for each possible output (letter in the con-

fusion set) reachable from this state, the number of paths starting from this state that lead to this output.

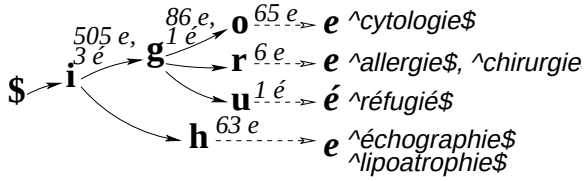

Figure 1: Trie of mixed contexts, each state showing the frequency of each possible output.

We call a state *unambiguous* if all the paths from this state lead to the same output. In that case, for our needs, these paths may be replaced with a shortcut to an exit to the common output (see figure 1). This amounts to generalizing the set of contexts by replacing them with a set of minimal discriminating contexts.

Given a word that needs to be accented, the first step consists in representing the context of each of its pivot letters. For instance, the word *biologie*: *$igoloib^*. Each context is matched with the transducer in order to find the longest path from the start state that corresponds to a prefix of the context string (here, *$igo*). If this path leads to an output state, this output provides the proposed accented form of the pivot letter (here, *e*). If the match terminates earlier, we have an ambiguity: several possible outputs can be reached (*e.g.*, *hémorragie* matches *$ig*).

We can take absolute frequencies into account to obtain a measure of the *support* (confidence level) for a given output O from the current state S: how much evidence there is to support this decision. It is computed as the number of contexts of the training set that go through S to an output state labelled with O (see figure 1). The accenting procedure can choose to make a decision only when the support for that decision is above a given threshold. Table 4

Context	Support	Gloss	Examples
$igo=e	65	–ogie	*cytologie*
$ih=e	63	–hie	*lipoatrophie*
$uqit=e	77	–tique	*amélanotique*
u=e	247	-eu-	*activateur, calleux*
x=e	68	-ex-	*excisée*

Table 4: Some minimal discriminating contexts.

shows some minimal discriminating contexts learnt from the accented part of the French MeSH with a high support threshold. However, in previous experiments (Zweigenbaum and Grabar, 2002), we tested a range of support thresholds and observed that the gain in precision obtained by raising the support threshold was minor, and counterbalanced by a large loss in recall. We therefore do not use this device here and accept any level of support.

Instead, we take into account the *relative frequencies* of occurrence of the paths that lead to the different outputs, as marked in the trie. A probabilistic, majority decision is made on that basis: if one of the competing outputs has a relative frequency above a given threshold, this output is chosen. In the present experiments, we tested two thresholds: 0.9 (90% or more of the examples must support this case; this makes the correct decision for *hémorragie*) and 1 (only non-ambiguous states lead to a decision: no decision for the first *e* in *hemorragie*, which we leave unaccented).

Simpler context representations of the same family can also be used. We examined *right contexts* (a variable-length string of letters on the right of the pivot letter) and *left contexts* (*idem*, on the left).

3.5 Evaluating the rules

We trained both methods, Brill and contexts (mixed, left and right), on three training sets: the 4054 words of the accented part of the MeSH, the 54,291 lemmas of the ABU lexicon and the 8874 words in the ICD-SNOMED word list. To check the validity of the rules, we applied them to the accented part of the MeSH. The context method knows when it can make a decision, so that we can separate the words that are fully processed (f, all *e*s have lead to decisions) from those that are partially (p) processed or not (n) processed at all. Let f_c the number of correct accentuations in f. If we decide to only propose an accented form for the words that get fully accented, we can compute recall R_f and precision P_f figures as follows: $R_f = \frac{f_c}{f+p+n}$ and $P_f = \frac{f_c}{f}$. Similar measures can be computed for p and n, as well as for the total set of words.

We then applied the accentuation rules to the 5188 accentable 'unknown' words of the MeSH. No gold standard is available for these words: human validation was necessary. We drew from that set a random sample containing 260 words (5% of the total)

which were reviewed by the CISMeF team. Because of sampling, precision measures must include a confidence interval.

We also tested whether the results of several methods can be combined to increase precision. We simply applied a consensus rule (intersection): a word is accepted only if all the methods considered agree on its accentuation.

The programs were developed in the `Perl5` language. They include a trie manipulation package which we wrote by extending the Tree::Trie package, online on the Comprehensive Perl Archive Network (www.cpan.org).

4 Results

The baseline of this task consists in accenting no *e*. On the accented part of the MeSH, it obtains an accuracy of 0.623, and on the test sample, 0.642. The Brill tagger learns 80 contextual rules with MeSH training (208 on ABU and 47 on CIM-SNOMED). The context method learns 1,832 rules on the MeSH training set (16,591 on ABU and 3,050 on CIM-SNOMED).

Tables 5, 6 and 7 summarize the validation results obtained on the accented part of the MeSH. *Set* denotes the subset of words as explained in section 3.5. *Cor.* stands for the number of correctly accented words.

Not surprizingly, the best global precision is obtained with MeSH training (table 6). The mixed context method obtains a perfect precision, whereas Brill reaches 0.901 (table 5). ABU and CIM-SNOMED training also obtain good results (table 7), again better with the mixed context method (0.912–0.931) than with Brill (0.871–0.895). We performed the same tests with right and left contexts (table 6): precision can be as good for fully processed words (set f) as that of mixed contexts, but recall is always lower. The results of these two context variants are therefore not kept in the following tables. Both precision and recall are generally slightly better with the majority decision variant. If we concentrate on the fully processed words (f), precision is always higher than the global result and than that of words with no decision (n). The n class, whose words are left unaccented, generally obtain a precision well over the baseline. Partially processed words (p) are always those with the worst precision.

training set	cor.	recall	precision±ci
MeSH	3646	0.899	0.901±0.009
ABU	3524	0.869	0.871±0.010
CIM-SNOMED	3621	0.893	0.895±0.009

Table 5: Validation: Brill, 4054 words of accented MeSH.

context	set	cor.	recall	precision±ci
right	n	1906	0.470	0.747±0.017
	p	943	0.233	0.804±0.023
	f	324	0.080	1.000±0.000
	tot	3173	0.783	0.784±0.013
left	n	743	0.183	0.649±0.028
	p	500	0.123	0.428±0.028
	f	1734	0.428	1.000±0.000
	tot	2977	0.734	0.736±0.014
mixed	n	7	0.002	1.000±0.000
	p	0	0.000	0.000±0.000
	f	4040	0.997	1.000±0.000
	tot	4047	0.998	1.000±0.000
majority decision (0.9)				
mixed	n	2	0.000	1.000±0.000
	p	0	0.000	0.000±0.000
	f	4045	0.998	1.000±0.000
	tot	4047	0.998	1.000±0.000

Table 6: Validation: different context methods, MeSH training, 4054 words of accented MeSH.

Precision and recall for the unaccented part of the MeSH are showed on tables 8 and 9. The global results with the different training sets at breakeven point, with their confidence intervals, are not really distinguishable. They are clustered from 0.819±0.047 to 0.842±0.044, except the unambiguous decision method trained on MeSH which stands a bit lower at 0.800±0.049 and the Brill tagger trained on ABU (0.785). If we only consider fully processed words, precision can reach 0.884±0.043 (ICD-SNOMED training, majority decision), with a recall of 0.731 (or 0.876±0.043 / 0.758 with MeSH training, majority decision).

Consensus combination of several methods (table 8) does increase precision, at the expense of recall. A precision/recall of 0.920±0.037/0.750 is

ABU training (strict)				*majority decision (0.9)*		
set	cor.	recall	precision±ci	cor.	recall	precision±ci
n	368	0.091	0.864±0.033	111	0.027	0.860±0.060
p	227	0.056	0.668±0.050	77	0.019	0.524±0.081
f	3164	0.780	0.964±0.006	3585	0.884	0.951±0.007
tot	3759	0.927	0.929±0.008	3773	0.931	0.932±0.008
CIM-SNOMED training				*majority decision (0.9)*		
n	176	0.043	0.752±0.055	57	0.014	0.803±0.093
p	114	0.028	0.425±0.059	51	0.013	0.300±0.069
f	3400	0.839	0.959±0.007	3607	0.890	0.948±0.007
tot	3690	0.910	0.912±0.009	3715	0.916	0.918±0.008

Table 7: Validation: mixed contexts, strict (threshold = 1) and majority (threshold = 0.9) decisions, 4054 words of accented MeSH.

training set	cor.	recall	precision±ci
MeSH	219	0.842	0.842±0.044
ABU	204	0.785	0.785±0.050
CIM-SNOMED	218	0.838	0.838±0.045
Combined methods			
mesh/Brill + mesh/majority	195	0.750	0.920±0.037
mesh/Brill + mesh/majority$_f$	185	0.712	0.930±0.036
mesh+abu+cim-snomed/Brill + mesh/majority	178	0.685	0.927±0.037

Table 8: Evaluation on the rest of the MeSH: Brill, estimate on 5% sample (260 words).

obtained by combining Brill and the mixed context method (majority decision), with MeSH training on both sides. The same level of precision is obtained with other combinations, but with lower recalls.

5 Discussion and Conclusion

We showed that a higher precision, which should make human post-editing easier, can be obtained in two ways. First, within the mixed context method, three sets of words are separated: if only the 'fully processed' words f are considered (table 9), precision/recall can reach 0.884/0.731 (CIM-SNOMED, majority) or 0.876/0.758 (MeSH, majority). Second, the results of several methods can be combined with a consensus rule: a word is accepted only if all these methods agree on its accentuation. The combination of Brill mixed contexts (majority decision), for instance with MeSH training on both sides, increases precision to 0.920±0.037 with a recall still at 0.750 (table 8).

The results obtained show that the methods presented here obtain not only good performance on their training set, but also useful results on the target data. We believe these methods will allow us to reduce dramatically the final human time needed to accentuate useful resources such as the MeSH thesaurus and ADM knowledge base.

MeSH training (strict)				*majority decision*		
set	cor.	recall	precision±ci	cor.	recall	precision±ci
n	19	0.073	0.731±0.170	8	0.031	0.727±0.263
p	15	0.058	0.429±0.164	11	0.042	0.458±0.199
f	174	0.669	0.874±0.046	197	0.758	0.876±0.043
tot	208	0.800	0.800±0.049	216	0.831	0.831±0.046
ABU training (strict)				*majority decision*		
n	30	0.115	0.882±0.108	13	0.050	0.929±0.135
p	32	0.123	0.711±0.132	11	0.042	0.786±0.215
f	153	0.588	0.845±0.053	194	0.746	0.836±0.048
tot	215	0.827	0.827±0.046	218	0.838	0.838±0.045
CIM-SNOMED training				*majority decision*		
n	27	0.104	0.818±0.132	14	0.054	0.824±0.181
p	19	0.073	0.487±0.157	9	0.035	0.321±0.173
f	168	0.646	0.894±0.044	190	0.731	0.884±0.043
tot	214	0.823	0.823±0.046	213	0.819	0.819±0.047

Table 9: Evaluation on the rest of the MeSH: mixed contexts, estimate on same 5% sample.

It is interesting that a general-language lexicon such as ABU can be a good training set for accenting specialized-language unknown words, although this is true with the mixed context method and the reverse with the Brill tagger.

A study of the 44 errors made by the mixed context method (table 9, MeSH training, majority decision: 216 correct out of 260) revealed the following errors classes. MeSH terms contain some English words (*academy*, *cleavage*) and many Latin words (*arenaria*, *chrysantemi*, *denitrificans*), some of which built over proper names (*edwardsiella*). These loan words should not bear accents; some of their patterns are correctly processed by the methods presented here (*i.e.*, unaccented *eae$*, *ella$*), but others are not distinguishable from normal French words and get erroneously accented (*rena* of *arenaria* is erroneously processed as in *rénal*; *académy* as in *académie*). A first-stage classifier might help handle this issue by categorizing Latin (and English) words and excluding them from processing. Our first such experiments are not conclusive and add as many errors as are removed.

Another class of errors are related with morpheme boundaries: some accentuation rules which depend on the start-of-word boundary would need to apply to morpheme boundaries. For in-

stance, *pilo/erection* fails to receive the *é* of *r^e=é* (*^érection*), *apic/ectomie* erroneously receives an *é* as in *cc=é* (*cécité*). An accurate morpheme segmenter would be needed to provide suitable input to this process without again adding noise to it.

In some instances, no accentuation decision could be made because no example had been learnt for a specific context (*e.g.*, accentuation of *céfalo* in *cefaloglycine*).

We also uncovered accentuation inconsistencies in both the already accented MeSH words and the validated sample (*e.g.*, *bacterium* or *bactérium* in different compounds). Cross-checking on the Web confirmed the variability in the accentuation of rare words. This shows the difficulty to obtain consistent human accentuation across large sets of complex words. One potential development of the present automated accentuation methods could be to check the consistency of word lists. In addition, we discovered spelling errors in some MeSH terms (*e.g.*, *bethanechol* instead of *betanechol* prevents the proper accentuation of *beta*).

Finally, further testing is necessary to check the relevance of these methods to other accented letters in French and in other languages.

Acknowledgements

We wish to thank Magaly Douyère, Benoît Thirion and Stéfan Darmoni, of the CISMeF team, for providing us with accented MeSH terms and patiently reviewing the automatically accented word samples.

References

[Brill1995] Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

[Darmoni et al.2000] Stéfan J. Darmoni, J.-P. Leroy, Benoît Thirion, F. Baudic, Magali Douyere, and J. Piot. 2000. CISMeF: a structured health resource guide. *Methods Inf Med*, 39(1):30–35.

[Garnier and Delamare1992] M. Garnier and V. Delamare. 1992. *Dictionnaire des Termes de Médecine*. Maloine, Paris.

[Grabar and Zweigenbaum2000] Natalia Grabar and Pierre Zweigenbaum. 2000. Automatic acquisition of domain-specific morphological resources from thesauri. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, pages 765–784, Paris, France, April. C.I.D.

[Habert et al.2001] Benoît Habert, Natalia Grabar, Pierre Jacquemart, and Pierre Zweigenbaum. 2001. Building a text corpus for representing the variety of medical language. In *Corpus Linguistics 2001*, Lancaster.

[INS2000] Institut National de la Santé et de la Recherche Médicale, Paris, 2000. *Thésaurus Biomédical Français/Anglais*.

[Levenshtein1966] V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklandy*, pages 707–710.

[Ruch et al.2001] Patrick Ruch, Robert H. Baud, Antoine Geissbuhler, Christian Lovis, Anne-Marie Rassinoux, and A. Rivière. 2001. Looking back or looking all around: comparing two spell checking strategies for documents edition in an electronic patient record. *J Am Med Inform Assoc*, 8(suppl):568–572.

[Seka et al.1997] LP Seka, C Courtin, and P Le Beux. 1997. ADM-INDEX: an automated system for indexing and retrieval of medical texts. In *Stud Health Technol Inform*, volume 43 Pt A, pages 406–410. Reidel.

[Simard1998] Michel Simard. 1998. Automatic insertion of accents in French text. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, Grenade.

[Spriet and El-Bèze1997] Thierry Spriet and Marc El-Bèze. 1997. Réaccentuation automatique de textes. In *FRACTAL 97*, Besançon.

[Theron and Cloete1997] Pieter Theron and Ian Cloete. 1997. Automatic acquisition of two-level morphological rules. In Ralph Grishman, editor, *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 103–110, Washington, DC, March-April. ACL.

[Yarowsky1999] David Yarowsky. 1999. Corpus-based techniques for restoring accents in Spanish and French text. In *Natural Language Processing Using Very Large Corpora*, pages 99–120. Kluwer Academic Publishers.

[Zweigenbaum and Grabar2002] Pierre Zweigenbaum and Natalia Grabar. 2002. Accenting unknown words: application to the French version of the MeSH. In *Workshop NLP in Biomedical Applications*, pages 69–74, Cyprus, March. EFMI.

[Zweigenbaum2001] Pierre Zweigenbaum. 2001. Resources for the medical domain: medical terminologies, lexicons and corpora. *ELRA Newsletter*, 6(4):8–11.