Spelling correctors to improve production and diffusion of linguistic knowledge in Subsaharian Africa

Chantal Enguehard

Spelling correctors
to improve production and diffusion of linguistic knowledge
in Subsaharian Africa

Chantal Enguehard
Laboratoire d'Informatique de Nantes-Atlantique – Nantes – France
Chantal.Enguehard@univ-nantes.fr

1. Introduction

The countries of the Sahel appear among poorest on earth and their rates of illiteracy are very high (83% in Niger, 60% in Mali, 57% in Senegal). However, literacy of the population is an essential factor to develop a country, it thus appears essential to set up strategies to improve the elimination of illiteracy of the populations.
We quickly review the recent history of transcription of African languages and the problems encountered in the production and the dissemination of linguistic information and texts. Then we propose the development of specific data-processing tools to capitalize and diffuse linguistic knowledge.

2. Situation

2.1. Short historical background

The colonial period

The colonization of African countries endangered emergent systems of transcription which existed already, in particular local alphabets well adapted to the local languages: the writing of Bamum which king Njoya creates and makes thrive at the end of the nineteenth century in Cameroun, the Mandé syllabary which appeared between 1833 and 1930, the Nko alphabet or the Arab writing [Dalby 1986]. This period is also marked by the appearance of many monographs on the African languages, works of missionaries or administrators colonial, i.e. people certainly of goodwill but without any qualification in linguistics. It should be stressed that these amateurish works ended up imposing the use of the Latin alphabet. .

The colonizers had different behaviours with regard to the local languages in teaching. In the former Belgian Congo (today RDC[1]), some African languages were taught whereas at the same time they did not have any rights to a place in the schools of French colonies.

Schooling in the European languages was not a success. In 1817, Jean Dard [Gaucher 1968], a French teacher at the mutual school of Saint-Louis of Senegal of which he was the director, noted "the failure of a unilingual French teaching leading the pupils to read and to write French without understanding it" [Dumont 1983]. He then began to teach using the Wolof language. In March 25, 1829, the school commission sent by the governer Dubelin to inspect the school condemned and put an end to this teaching innovation.

This assimilationist policy of France of that time contributed to worsening the social and economic underdevelopment of Africa by maintaining most of the population in chronic illiteracy. How to learn to read and write in a foreign language that one does not understand?

Postcolonisation

Very early, African intellectuals understood that increasing literacy of the population, a factor essential to develop a country, must necessarily be based on teaching in maternal languages[2]. After decolonization, considerable efforts were thus carried out to transcribe African languages, but the delay on the matter, associated with the confusion introduced by amateur linguists, considerably complicated the task. It is necessary to fully appreciate the difference between the linguistic situation of great international languages like French or English, on which thousands of linguists have made very thorough studies, which have profited from national academies of languages for several centuries, and the linguistic situation of African languages whose natural historical passage from the oral to the

[1] Democratic Republic of Congo

[2] In 1937, during his conference to the Chamber of Commerce of Dakar, the future President of the Republic of Sénégal, Léopold Sédar Senghor, recommended to teach the maternal languages.

written was brutally interrupted, then largely corrupted by exogenous elements. Work to be carried out on African languages represents an immense task that is crucial to conclude in order to benefit the population in terms of development and education.

For 50 years, much work of great importance has been undertaken, but linguistics is a difficult science because it relates to the totality of speakers, it takes time (and means) to diffuse its results, have them reach the population and finally be adopted.

Work to define the alphabets (generally under the aegis of UNESCO) reflects these difficulties: the alphabet defined for the Mande languages in Bamako in 1966, though remarkable since it allowed a certain harmonization with the west-Mande of Senegal and Gambia, shocked certain practices and was not harmonized with other alphabets proposed for other languages (Fula and Songhaï in particular). During the following decade, different alphabets were created in the various Mandephone states, each one diverging in one way or another from the alphabet of Bamako. In July 1978, a new meeting led to the creation of an "African Reference Alphabet" based on conventions of the IPA (International Phonetic Association) and the IAI (International African Institute). This alphabet makes possible to note all the phonic possibilities of the African languages. Lastly, a few months later, during a workshop on the harmonization of the orthography of the Manden (various Mande alphabets which had developed in divergent ways in the 1970s) new problems had been dealt with: tones, segmentation and elision, punctuation, etc. The alphabet adopted was again improved in June 1981 with rules for tonal notation.

It thus required fifteen years to define an adequate alphabet. As this alphabet comprises characters not appearing in the Latin alphabet (for example ɛ, ɔ, ɲ, ŋ) it met great resistance because of technical problems arising when typing and editing texts, etc. It seemed simpler to continue to use only the Western Latin alphabet. Thus, in Guinea, this common alphabet was adopted only since 1988. In Burkina Faso and Ivory-Coast, it is only partially used since the letter 'ɲ' is replaced by 'ny'.

2.2. Technical evolution

As the special characters of the African alphabets do not appear in the ASCII representation, many fonts were created in the 1980s: glyphs of some useless characters (&, @, <, etc.) are redrawn to make appear the special characters. No harmonization having prevailed, the special characters do not always replace the same initial characters, making almost impossible the exchange and electronic treatment of documents. Thus texts in 7 languages collected in 2004 in Niger and Mali use no less than 25 different fonts [Enguehard 2004]!

The appearance of the Unicode standard in 1992 represents a considerable progress because it makes possible to represent more than one million different characters, i.e. all the characters of all the languages. This new encoding of characters authorizes all languages to reach the first stage of computerization of a language: the storage of documents in an electronic form that allows their analytical treatment [Chanard 2001].

However, some problems remain (see the illustrations Table 1):

Case 1: Certain characters still missing, this is the case of the 'r' with a horizontal bar, (Kanuri language[3] - decree 213-99 of the Republic of Niger). This letter can be composed by the superposition of a hyphen '-' on the letter 'r' but, as it is not possible to specify the position of a sign in relation to the other, the resulting glyph is not always satisfactory. In addition, users are tempted to use hyphens of different code-points according to glyph appearance (hyphen, dash), which harms a single representation of this character.

Case 2: Generally, many characters with diacritic signs have several possible encodings. Thus 'e' with acute accent is represented by the code U+00E9, but also by the combination of the letter 'e' and the code of the acute accent: U+0065 U+0301. These multiple encodings complicate the analytical treatment of texts. Moreover it happens that the software displays these accentuated characters differently, according to the codes. In particular, a character already presenting an additional sign (like the point above ‘i’), superimposed at a diacritic sign, is displayed by superimposing the point and the diacritic sign, which is incorrect: the point above ‘i’ should be replaced by the diacritic. Thus, the character ‘J with caron’ (Tamasheq language[4] - decree 214-99 of the Republic of Niger) is coded by the single code U+01F0 in its tiny form. The capital form, curiously, is absent from the Unicode alphabet. The display of the superposition of the capital ‘J’ with the caron is disappointing: the caron accent appears like a flat accent, or it is not located correctly above the 'J '.

Case 3: Certain languages use digraphs, signs made up of two letters but whose capital form reveals only the first letter in upper case. The Hausa language, for example, uses 9 digraphs (decree 212-99 of the Republic of Niger). A text can mix words of different languages, or proper names, in which can appear the same combinations of letters as

[3] One of the national languages of the Republic of Niger

[4] One of the national languages of the Republic of Niger

these digraphs whereas they are separate letters. In English, for example, the word "gyroscope" includes the continuation of letters "gy" but it is not a question of the digraph 'gy'. ɔ

These problems cannot be managed by fonts. The best solution is the definition of specific codes for the absent characters, all the characters with diacritics, and all the digraphs. This solution seems possible since, in the Unicode standard, many codes are still not affected.

	Langage	Minus lettre		Capital lettre	
		display	codes	display	codes
'r' with bar	Kanuri	ɾ	U+0072 (r) U+0335 (short stroke)	Ɍ	U+0052 (R) U+0335 (short stroke)
		ɾ	U+0072 (r) U+0336 (long stroke)	Ɍ	U+0052 (R) U+0336 (long stroke)
e with acute accent	French	é	U+00E9	É	U+00C9
		é	U+0065+U+0301	É	U+0065+U+0301
'j' with caron	Tamasheq	ǰ	U+01F0		
		ǰ	U+006A U+030C	J̌	U+004A U+030C
digraphes	Hausa[5]	fy	U+0066 (f) U+0079 (y)	Fy	U+0046 (F) U+0079 (y)
		gw	U+0067 (g) U+0077 (w)	Gw	U+0047 (G) U+0077 (w)
		gy	U+0067 (g) U+0079 (y)	Gy	U+0047 (G) U+0079 (y)
		kw	U+006B (k) U+0077 (w)	Kw	U+004B (K) U+0077 (w)
		ky	U+006B (k) U+0079 (y)	Ky	U+004B (K) U+0079 (y)
		ƙw	U+00199 (ƙ) U+0077 (w)	Kw	U+00198 (Ƙ) U+0077 (w)
		ƙy	U+00199 (ƙ) U+0079 (y)	Ky	U+00198 (Ƙ) U+0079 (y)
		sh	U+0073 (s) U+0068 (h)	Sh	U+0053 (S) U+0068 (h)
		ts	U+0074 (t) U+0073 (s)	Ts	U+0054 (T) U+0073 (s)

Table 1 : some special characters

2.3. Linguistic situation

Few linguists

The scientific description of the African languages is far from being completed and much essential work still remains to be carried out: the languages present many regional variants, concurrent transcriptions of the same word are current, for a single language the official orthographical systems are generally not harmonized from one country to another, or even inside the same state. This situation naturally led to the existence of different practices[6] within the same language. As we saw, these studies must imperatively be undertaken by professional linguists in order to

[5] One of the national languages of the Republic of Niger

[6] Here again the heavy colonial past should be noted. In Wolof, the statement [nja:y] is transcribed ndiaye in Senegal (colonized by French), and njie in Gambia (colonized by English). It should be noted that such a situation was avoided between Mauritania and Senegal, whose systems of transcription were respectively founded on principles of morphophonologic or phonological nature. A project of the ACCT (Agency of Cultural and Technical Assistance), now AIF (Intergovernmental Agency of Francophonie) allowed, in the 1990s, Senegalese and Mauritanian experts to meet, in Dakar as well as Nouakchott, and to harmonize the system of transcription of Wolof in the two concerned countries.

develop coherent systems of transcription, likely to establish the local languages and to contribute to their introduction into the education system.

But in poor countries where a very small part of the population is educated, linguists are extremely rare whereas the descriptive linguistic work that remains to be carried out is gigantic. The countries concerned do not have the economic means to train linguists, or to give the means to trained linguists to work under good conditions: the majority of work I described profited from the support of international organizations like UNESCO, national ones (V.S.O.), or private ones like SIL International.

It should be added that these countries face such huge underdevelopment problems (access to drinking water, food, education, health), that they did not have the means of making decisions taken on linguistic matters compulsory through binding legislative measures: the official systems of transcription are not always respected and the decrees are generally ignored.

No linguistic resources

The endemic economic weakness, combined with the lack of qualified people, has impeded the production of linguistic knowledge necessary to the proper use of a language (dictionaries, grammars, etc). Thus, the majority of the languages do not benefit from any monolingual dictionary[7], which constitutes a paradoxical situation since they are often equipped with several bilingual dictionaries.

This situation is worsened by the westerners who, without bad intention but with a remarkable lack of conscience, still nowadays contribute to scrambling the tracks. For instance, books about botany, which are largely diffused throughout the world, often propose names of plants in several African languages. These names, transcribed by botanists without linguistic competence, do not comply at all with the correct transcription rules of these languages (most of the time, they are not even written with the adequate alphabet) [Enguehard 2003]. More seriously, on the web, many amateur sites give rudiments in various languages. Their authors transcribe words without complying with the rules of good transcription.... because they are unaware of their existence.

This adulterated knowledge is widely diffused throughout the world (botany books are present in libraries, Internet sites are open to all) whereas reference resources produced by African linguists, are not accessible because they still remain too expensive for the people.

Conscious of the vital stake that the elimination of illiteracy of the population represents, the States of West Africa have however implemented a policy of significant linguistic planning, choosing their national languages, standardizing the alphabets and the rules of transcription, supporting centres of production of little books on different subjects: medicine, hygiene, management, rights of children, etc., and including since a few years the teaching of national languages. However, the participants in the recent national conference on the promotion of the languages in Mali[8], deplored the lack of writing in national languages: people are taught reading and writing in their language but they lose their knowledge little by little because they almost never have the occasion to read, and what there is to read does not always correspond to their interests. In particular there are few books for children, almost no comic strips. It thus appears vital to encourage and support the production of writing in national languages.

In addition, linguistic knowledge inculcated in the elementary school cannot be perpetuated. Who would correctly write his own language while having left the school at age 14, while never having access to a dictionary or grammar, and seldom encountering written texts? Under such conditions, to write an African language constitutes a feat that is the prerogative of rare elites.

The production of linguistic reference resources and their wide diffusion constitutes a major goal to make possible to African people the acquisition of a culture of writing, and to thus escape chronic illiteracy.

3. processing, an asset for the African languages

3.1. Assumptions

It is imperative to develop low cost strategies (human and financial) to make possible the constitution and diffusion of linguistic resources.

We make the assumption that the people writing texts in a national language (school books, technical books, newspapers, Internet sites) could maintain a good quality of language and produce texts in greater quantity if they had access to the linguistic resources they need.

---

[7] We can cite an exception : the monolingual Zarma dictionnary : Isufi Alzuma Umaru, "Kaamuusu Kayna", éd. Alpha, 1996.

[8] Bamako, 15-17 January 2004

However, any text, at one time or another of its development, is composed and stored in an electronic medium. We make the assumption that this systematic use of the computers represents an opportunity to seize because they represent a place to diffuse and collect linguistic data, and offer modern and elaborate tools that are able to accelerate the slow work of language transcription.

3.2. Facilitating the dissemination of linguistic information

Linguistic knowledge can be presented in the form of hypertexts downloadable from the sites of the institutions, but producing hypertexts represents an too huge amount of work to make it possible in a short-term implement. It seems more judicious to create orthographical correctors within the text editors because it is possible to produce such correctors with little human work but based on corpora texts.

In addition, even if the African languages are currently not well represented on the Web [Diki-Kidiri 2003], the solution of representing and displaying special characters and the development of adapted editors, should allow the creation of Internet sites written in African languages for the speakers of these languages.

3.3. Facilitating the production of texts

Users can also be solicited to provide their production to institutions in order to collaborate in the constitution of corpora. This collaboration supposes that the produced texts are coded according to the Unicode standard still badly known in Africa out of the community of the data processing specialists. It can be set up only thanks to the development of adapted tools, which is not currently the case.

The majority of the special characters are not available on the usually distributed keyboards. It is possible to compose them by using their code, but this solution obviously lacks ergonomics (it is necessary to remember the codes, typing several keys to obtain a character). Within a research project of the AUF joining together the University of Nouakchott, the University of Dakar and the ISTI, virtual Unicode keyboards were developed in Balante, Bambara, Pulaar, Serer and Wolof (http://www.termisti.refer.org/ltt/ltt.htm). These keyboards make it possible to obtain the special characters required by striking only one or two keys. The generated code is the Unicode code of the character. It is easy to materialize the characters on a physical keyboard by posing a mask comprising these characters. The development, and the diffusion of these keyboards constitutes a significant progress in the production of texts since the task of composition is carried out under satisfactory ergonomic conditions and the text is directly encoded according to the Unicode standard. Unfortunately the choice of the site of the special characters on the keyboard is governed by no general principle, the risk of proliferation of competitor keyboards is real. It is thus urgent to define rules of attribution of the keys of keyboard to the new characters.

The usually available text editors (like Word or Open Office) are carried out in international languages (English, French, Spanish, etc). It is obviously possible to use such editors to write other languages but then unforeseen difficulties emerge. First of all, it is difficult to use a software if one does not have a command of the language in which its interface is written: the user must function in a bilingual mode, which can have consequences on his cognitive operation, one of the languages being able to influence the words and the selected syntactic structures for the drafting in the other language. Then, the complementary linguistic functionalities, like the correction of the spelling, are unusable as soon as language is changed. For the moment, there are no editors developed for the less dominant languages, like African languages. However, the progress made in the encoding of characters, the displaying of text, and the development of virtual keyboards makes it possible to consider the development of such software.

4. Objectives

We focus on the development of spelling correctors produced from corpus of texts, and tools allowing linguists to easily strip corpora of texts in order to enrich the linguistic resources.

4.1. Spelling correctors

Short state of the art

Spelling correctors constitute a research orientation since the years 1960 [Kukich 1992]. They are now usually used by general public because the current text editors often integrate one, and they bring a considerable comfort during the drafting of texts. These correctors function according to an interactive mode in which the user intervenes, contrary to the automatic spelling correctors as in the field of the optical character recognition (and of which we are not concerned here).

An interactive spelling corrector functions following several stages:

1. detection of the errors;
2. selection of the possible corrections;
3. ordering the possible corrections and proposition to the user;
4. effective correction of the text respecting the choice of the user.

1. The detection of errors is carried out while considering one by one the words of the text to correct, in an isolated way. Each word of the text is compared with the words of the lexicon (which contains the words of the language, with their inflections). Any word not found in the lexicon is regarded as erroneous. This technique is very simple to implement but presents the disadvantage of not detecting the errors transforming a word into another word present in the lexicon. For instance, in the sentence "I wrote this brook ". The word "book", was transformed into "brook", which is obviously erroneous. The rate of such real-word errors increases with the size of the lexicon. The increase in the size of the lexicon thus contributes, paradoxically, to degrade the performances of the spelling corrector. Only taking into account the context of the words can help to avoid this major pitfall. The first experiments in this direction were founded on the calculation of trigrams (on the words). This theoretically valid approach presents the major disadvantage in requiring an enormous corpus which is not always possible to constitute [Mays 1991]. The most recent work falls under the field of lexical cohesion. But to simply test the lexical chains on a sentence leads the system to detect errors that, for nine out of ten cases, are not errors [Hirst 1998]. A new algorithm exploiting the various semantic relations between words (synonymy, meronymy, frequency of co-occurrences, etc), and extending the concept of vicinity, formerly restricted with the sentence to one or more paragraphs, seems capable of much better performances [Hirst 2003].

2. When an error is detected, the corrector selects a series of words likely to be the correct version of the chain to be corrected. These words are selected according to various techniques (calculation of the minimal editing distance, of the key of similarity, or measurement of the phonological distance).

3. The ordering of the possible corrections takes into account the measurement used during the previous stage of selection, as well as statistical measurements (like the frequency of appearance of the words, or the word most frequently selected when previously meetings with the same error).

Lastly, an interactive stage makes possible the user to supervise the correction. He can adopt one of the three following attitudes:

- to correct the erroneous word by selecting one o f the proposed corrections;
- to modify the erroneous word;
- to not correct; in this last case, he can add this word to his personal dictionary.

Spelling correctors encounter two major difficulties. First of all, inopportune concatenations of words, or the insertion of a delimitor (space character, punctuation) inside a word make very delicate the selection of candidates for the correction. This difficulty is however not too awkward in an interactive correction because these typing errors are easily corrected by the user. The update of the lexicon constitutes a more significant problem: the languages evolve rather quickly, the use of a corrector based on an several years old lexicon reveals that number of words usually used are wrongfully diagnosed as erroneous because they are loans, neologisms, or new derivations of words.

Specific needs

There are already spelling correctors for some African languages (Microsoft announced in 2004 the marketing of an spelling corrector for Word in Kiswahili), but they are generally very simple: they use existing spelling correctors, providing a lexicon corresponding to the targeted language [Van der Veken 2003]. These spelling correctors locate the errors by scanning the words of the text in an isolated way and, even if they render services, they will fatally encounter the previously underlined problems. We think that a spelling corrector adapted to the African languages must take into account the contexts of the words in order to not limit its performances.

In addition, it must have additional functionalities compared to the usual spelling correctors in order to take into account the linguistic context of the African languages. On the one hand, it can take part in the dissemination of linguistic information by accompanying the proposals for corrections of additional linguistic information; on the other hand, it can contribute to the constitution of linguistic resources by collecting data intended for linguists.

Lastly, a spelling corrector encounters, by definition, many texts, and is provided with a lexicon which enables to identify the words absent from the lexicon. We mentioned that a user can add words that are correct, but absent from general lexicon, to his personal lexicon. We wish to exploit this process of enrichment in order to help the institutions responsible for the languages to increase the official lexicon of a language. When adding a word to the personal lexicon, the spelling corrector can memorize this word in a file, as well as the sentence in which it appears. The user is highly encouraged to transmit this file to the institution charged with work on the language. The

information contained in this type of file can be used to enrich the lexicon of the language (see “support for the production of linguistic resources”).

This spelling corrector is, to a certain extent, independent of the language since the language dependent treatments (like the calculation of the inflections and derivations of words for example) are described in generic modules which use information gathered in the electronic lexicon of the language. Each item is, as far as possible, accompanied by its grammatical category, by its mode of inflection and by its possible derivations in order to be able to extend the lexicon to all the forms. Contextual information is also integrated into the lexicon in the form of probabilities. To adapt this spelling corrector to a new language, it is thus enough to change the lexical resources.

This spelling corrector is under development at the University of Nantes, it will be compatible with the usually used text editors (like Word) as well as the editors of electronic messages. The development of the pilot version should be completed in 2004. The tests carried out at the beginning of 2005 will lead to the realization of a new version in 2006.

4.2. Support for the production of linguistic resources

We already underlined the lack of linguistic resources on African languages. It is vitally important to support their development, because these resources constitute the basis spelling correction software. As the work to be carried out is very significant, and the qualified people are rare, it is necessary to provide tools to help the production of these resources:

The enrichment of the lexicon within the institution in charge of work on the language can be facilitated by the development of an adequate software integrating the examination of corpus. This software has several objectives:

- integration of the contributions of the users of a spelling corrector;
- enrichment of the items of the lexicon (lexical category, definitions, examples of use, etc.)
- automatic calculation of statistical data (trigrams on the symbols, frequency of appearance of the words, etc),
- generating concordances for the words chosen by the user.

People in charge of the maintenance of the electronic resources would see their task facilitated by the use of software to note new information (grammatical category, definition, etc.) in adapted forms. Such information could be suggested by the software following a phase of training on the language starting from the corpora. This information could be directly registered in the electronic lexicon of the language.

4.3. Constitution of linguistic resources

We have as a project to develop spelling correctors for several African languages (Bambara, Kanuri, Tamajaq, Fulfulde, Wolof, Hausa and Zarma). We collected textual resources from institutions, journalists and publishers in order to initialize the electronic lexicon of each language. Linguists specialized in these languages checked that these texts are written in accordance with the rules of transcription and orthography in order not to skew the quality of our corrector.

Language	Number of words
Bambara	89 684
Kanuri	79 336
Tamajaq	350 010
Fulfulde	24 088
Hausa	139 239
Zarma	74 398
Wolof	In evaluation process

Table 2 : collected corpus

The constitution of electronic linguistic resources starting from these rough corpora proceeds in several stages: It is necessary to standardize the texts which, although written in correct language and respecting the decrees in force, were produced with "redrawn" font to display the special characters, which makes impossible their electronic exploitation just as they are. We thus modified the coding of these special characters in order to respect the Unicode standard. We then marked out the texts with XML[9] mark-ups according with the XCES[10] standard.

[9] eXtended Markup Language http://www.w3.org/XML/

These textual resources are supplemented by some generally bilingual lexicons (Bambara-French lexicon on laws, for example), and by bilingual dictionaries (Bambara-French, Hausa-French, Wolof-French) or monolinguals ones (Zarma).

Dictionaries constitute a particularly valuable resource because they contain many useful semantic bonds for a spelling corrector (synonymy, antonymy, analogy).

5. Conclusion

The technical obstacles in the writing of African languages disappear thanks to emergence of the Unicode standard. Computing tools are not always adapted to these languages, but in a few years considerable progress should take place (effective development of spelling correctors, constitution and diffusion of electronic linguistic resources), allowing African people to use their own languages in the most modern environments of communication.

The capitalization of electronic texts will be also a factor of progress making possible certain applications. For instance, the measurement of the frequencies of the words in a corpus represents a new criterion to decide which words must appear in a basic lexicon, or in a dictionary. A concordance is a valuable help that makes it possible to distinguish the various meanings of a word, or to compare the contexts of use of two words. Texts of dialectal alternatives can be statistically compared. Other medium-term developments can also be considered, like the automatic extraction of terms, or help in translation.

6. Bibliography

[Bailleul 1996] Bailleul, C. "Dictionnaire bambara-français", éd. Donniya, Bamako, Mali, 1996.

[Chanard 2001] Chanard, C., Popescu-Belis A., "Encodage informatique multilingue : application au contexte du Niger". Les cahiers du RIFAL n°22 , pp.33-45, 2001.

[Dalby 1996] Dalby, D., "L'Afrique et la lettre", Centre Culturel Français, Lagos & Fête de la Lettre, Paris, 1986.

[Diki-Kidiri 2003] Diki-Kidiri, M., Atibakwa Baboya E. (2003) "Les langues africaines sur la toile", Les cahiers du RIFAL n°23 Le traitement informatique des langues africaines, pp. 5-32, novembre 2003.

[Dumont 1983] Dumont, P., "Le français et les langues africaines au Sénégal", Paris, Acct-Karthala, 1983.

[Enguehard 2003] Enguehard, C., Mbodj, C., "Recueillir et diffuser les noms des plantes dans les langues africaines". Les cahiers du RIFAL n°23 Le traitement informatique des langues africaines, pp.47-54, novembre 2003.

[Enguehard 2004] Enguehard, C., Mbodj, C., "des correcteurs orthographiques pour les langues africaines". BULAG n° 29 : La correction automatique : bilan et perspectives, pp.51-68, 2004.

[Gaucher 1968] Gaucher, J., "Les Débuts de l'enseignement en Afrique francophone. Jean Dard et l'école mutuelle de Saint-Louis du Sénégal" , Paris, Le livre africain, 198 pages, 1968.

[Kukich 1992] Kukich, Karen. "Techniques for automatically correcting words in text". ACM Computing Surveys, 24(4), pp.377–439. 1992.

[Hirst 1998] Hirst, G., St-Onge, D., "Lexical chains as representations of context for the detection and correction of malapropisms", Fellbaum, pp.305-332, 1998.
http://www.cs.toronto.edu/compling/Publications/Abstracts/Papers/Hirst+Budanitsky-2001-abs.html.

[Hirst 2003] Hirst, G., Budanitsky, A., "Correcting real-word spelling errors by restoring lexical cohesion", 2003.

[Mays 1991] Mays, E., Damerau, F. J., Mercer, R. L., "Context based spelling correction", *Information Processing and Management*, 27(5), pp.517-522, 1991.

[Van der Veken 2003] Van der Veken, A., de Schryver, G.-M., "Les langues africaines sur la Toile : études des cas haoussa, somali, lingala et isixhosa". Les cahiers du RIFAL n°23, pp.33-45, novembre 2003.

Chantal Enguehard is a Lecturer in Computer Sciences since 1993. She works in Natural Language Processing and is specialized in the field of the languages having few linguistic resources: non international languages like the African ones, or low quality international languages (like the catches of notes or small advertisements).
Since 1995, Chantal regularly works in the area of the Sahel, in particular in Senegal, Niger and Mali where she collaborates with the national institutions.

[10] Corpus Encoding Standard for XML http://www.xml-ces.org/