HAL
open science

VARIATIONS LEXICOGRAPHIQUES EN ALBANAIS CONTEMPORAIN, A L'EPREUVE DU TAL

Odile Piton, Remzi Pernaska

To cite this version:

Odile Piton, Remzi Pernaska. VARIATIONS LEXICOGRAPHIQUES EN ALBANAIS CONTEMPORAIN, A L'EPREUVE DU TAL. Journées de Linguistique de Corpus 2013, équipe LiCoRN de l'Université de Bretagne Sud, Sep 2013, Lorient, France. pp.1-19. hal-01088895v2

HAL Id: hal-01088895

https://hal.science/hal-01088895v2

Submitted on 2 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Variations lexicographiques en Albanais contemporain, à l'épreuve du TAL

Odile Piton[1] & Remzi Përnaska[2],
[1]Laboratoire SAMM, Université Paris1 Panthéon-Sorbonne, France
Odile.Piton@univ-paris1.fr
[2] Laboratoire Forell, Université de Poitiers, France
r.pernaska@club-internet.fr

Résumé

Depuis plusieurs années, nous développons des outils pour le traitement automatique de la langue albanaise. Après avoir esquissé l'historique des réformes ayant conduit à l'élaboration d'une langue nationale, nous étudions la conformité des textes étudiés à la langue littéraire. Ayant constaté des différences, et la fréquence de variations orthographiques, voire grammaticales, nous proposons une méthode dynamique de traitement, constituée d'un ensemble de grammaires morphologiques d'annotations. Nous montrons qu'elles sont de nature à élargir le traitement automatique à de nombreuses variations que nous analysons.

I- Introduction

La communauté scientifique est très active dans le domaine des technologies de la langue où collaborent linguistes et informaticiens. Nous nous intéressons au traitement de la langue albanaise. Fruit d'une décision politique, cette langue nationale, telle qu'elle est définie, semble bien adaptée à un traitement automatique. Des études linguistiques en vue du TAL de la langue littéraire albanaise sont menées depuis quelques années (Lagji 2006), (Piton, Lagji 2008), (Kabashi 2003, 2005, 2011), (Trommer, Kalluli 2004). Des ressources sont développées (Murzaku), (Piton Përnaska 2006), (Piton O., Lagji K., Përnaska, 2007).
Basée sur l'unification du guègue et du tosque, des variations orthographiques, voire syntaxiques, apparaissant néanmoins dans les textes, et sont sources de problèmes, la non reconnaissance d'un terme entraînant l'échec de traitements ultérieurs. Dans ce travail nous exposons la méthode que nous employons pour élargir le traitement de la langue littéraire. Nous présentons brièvement la langue albanaise puis traitons les problèmes rencontrés.

II- Des dialectes à la langue littéraire

A- Le guègue et le tosque, deux familles de dialectes

La langue albanaise est composée de deux dialectes: le tosque et le guègue, qui furent peu ou pas écrits durant des siècles. Les dialectes albanais, qui peuvent être considérés comme deux variations d'une même langue ancienne, ont divergé à l'époque ottomane. Dans cette région où coexistaient Musulmans, Orthodoxes et Catholiques, la population était organisée en millets selon sa religion. Durant les 400 ans de l'occupation ottomane, l'enseignement est quasi inexistant. Son organisation est dévolue aux millets. Les religieux musulmans organisent l'instruction religieuse

en turc à Istanbul, les Orthodoxes le font en grec dans le Sud pour les tosques, et c'est en caractères latins qu'écrivent les Catholiques guègues du Nord. Ce sont donc des alphabets différents que pratique la minorité albanaise instruite. A ceci s'ajoutent les différences dialectales dues à l'isolement régional des communautés. Citons Michel Roux : « Au début du XXe siècle, les Albanais se trouvaient dans une situation de retard culturel marqué par rapport aux peuples voisins, eu égard à un illettrisme presque général, à la quasi absence de codification de la langue et à l'extrême faiblesse de la production écrite. », néanmoins « l'énorme retard de l'alphabétisation ne doit pas inciter à considérer les Albanais comme un peuple inculte, comme si la culture orale n'était pas une culture. »

La langue porte avec vigueur - et s'en nourrit-, les fleurons de l'oralité que sont la poésie, le récit populaire et l'épopée épique. Les deux dialectes, eux-mêmes diversifiés, développent au cours des siècles des spécificités régionales, morphosyntaxiques, phonétiques et lexicales. Michel Roux : « Les parlers du Nord diffèrent notablement de ceux du Sud, ou tosques, du point de vue phonologique, morphologique et lexicologique, sans toutefois que ces différences empêchent l'intercompréhension. » « Sous l'Empire ottoman,[les Albanais] vivaient ce qu'on peut appeler une situation de polyglossie, c'est-à-dire avec un système de communication comportant plusieurs langues aux emplois nettement délimités et hiérarchisés, avec une relation forte entre statuts sociaux et niveaux de compétence linguistique. La langue albanaise, bien qu'il fût possible de l'écrire en caractères arabes, grecs ou latins, restait pour l'essentiel confinée au domaine de l'expression orale. […] Pulvérisé en dialectes, infiltré de milliers de mots turcs, slaves ou autres (avec de considérables différences régionales et confessionnelles dans cette imprégnation), l'albanais ne connaissait alors aucune codification lexicale, orthographique ***et grammaticale. »***

Auguste Dozon, Consul de France à Ioannina, s'intéresse à la linguistique et dans ses différents postes, notamment en Albanie, il recueille des récits (contes, chansons, proverbes) ce qui le met dans l'obligation d'étudier la langue. Nous donnons un extrait de ses observations, de son travail pour concevoir un alphabet, rédiger une grammaire, recueillir des textes. Il donne ses impressions sur les récits « chkipetars » recueillis, la langue, les influences italienne, turque et grecque. Son travail est un témoignage de l'état de la langue dans les années 1870. Il raconte de manière pittoresque dans quelles conditions il a recueilli les récits.

> « Je me suis vu entraîné, sans en avoir eu aucunement le projet, à étudier l'albanais. » « La préface de la grammaire rendra compte du système orthographique que j'ai été conduit à adopter, faute de mieux. » « Je connais un homme, — il était naguère dans ma maison, c'était un de mes kavas, musulman, né à Prévéza d'une mère Grecque et d'un père Albanais, échappé jadis au massacre des Gardikiotes par Ali-Pacha, — qui sait l'une et l'autre langue, mieux le grec, et a en outre la mémoire très-bien garnie de contes, qu'il ne fait aucune difficulté de dire, dans son jargon gréco-épirote. Et parmi les nombreux Albanais aussi bilingues, on en trouverait sans doute plus d'un autre également propre à servir d'agent de transmission entre les deux peuples, dont les fictions présentent d'ailleurs la plus grande ressemblance. Parmi les quatre élèves du gymnase d'Ianina que j'ai eus successivement pour maîtres, et sous la dictée de qui j'ai écrit, les uns m'ont répété ce qu'ils avaient appris dans leurs familles, un autre s'en allait le soir dans une auberge fréquentée par les voyageurs de son pays, et s'y faisait raconter ce qu'il me rapportait le lendemain. »
>
> « Il n'est rien que je n'aie écrit moi-même, — et cela en exerçant un contrôle perpétuel et sur les mots, et sur la syntaxe, et parfois même sur la rédaction, — sous la dictée d'un Chkipetar, notamment des quatre étudiants dont il a été question plus haut, et qui s'étaient pliés à ma fantaisie, tout extraordinaire qu'elle leur parût peut-être. Ces jeunes gens savaient passablement le grec, en connaissaient la technologie grammaticale, et c'est par l'intermédiaire de cette langue qu'ils ont pu me fournir les explications pratiques les plus nécessaires ; quant aux théoriques, il en est que je cherche encore, même après de persévérantes études. On me comprendra si l'on songe qu'aucun Chkipetar de Turquie, à l'exception de Kristophoridhis, n'a encore réfléchi sur sa langue, ne sait

l'écrire, et ne croit possible ou même utile de le faire ; s'il a le goût et le moyen de s'instruire, il n'aspire (je parle des chrétiens) qu'à posséder le grec, seul instrument d'éducation qu'il ait à sa portée. Éloigné de ses parents, c'est en cette langue qu'il communique avec eux. Comme tous les idiomes, surtout ceux qui ne sont point cultivés, l'albanais se partage en une infinité de dialectes, plus ou moins caractérisés. »

Auguste Dozon : La littérature populaire chez les Chkipes ou Albanais (extraits)

Il crée un système d'écriture de l'albanais, dont nous donnons un exemple.

Kyénœ tri mótra, ňœ nga ató m'e vógœlya kyœ kyoúhey Fatimé, içte m'e boúkourœ nga tœ dúa.
L'écriture actuelle est :
Qenë tri motra, një nga ato më e vogëla që quhej Fatime ishte më e bukur nga të dya.
Il y avait trois sœurs, l'une d'elles, la plus petite qui s'appelait Fatimé, était plus belle que les deux [autres].

Nous empruntons la carte linguistique de l'albanais à l'ouvrage de Gut, Brunet-Gut, Përnaska.

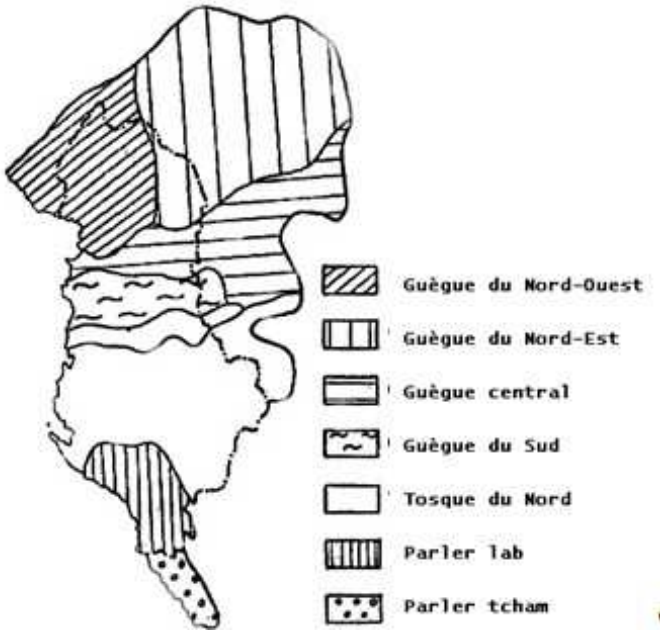

Figure I : Carte linguistique de l'albanais

Tel est l'outil de communication en Albanie à la fin de l'Empire ottoman.
Au XIX° l'analphabétisme est presque général, la première école en albanais ouvre en 1887. A la chute de l'état ottoman, tout est à faire en Albanie dans le domaine de l'instruction. Il faut créer des écoles et former un corps enseignant. Les dialectes ne peuvent fournir le support de l'enseignement. Se pose le problème de l'unification de la langue et de sa codification grammaticale et lexicale.

B- Unification de l'Albanais

Déjà au XIXème siècle, avec le mouvement de la Renaissance nationale de l'Albanie –la Rilindja Kombëtare-, un comité de lettrés avait exprimé la nécessité de concevoir un alphabet pour l'albanais. Ceci donna lieu à des travaux et des publications en albanais. Plus tard, en 1908, le Congrès de Monastir a facilité le choix définitif de l'alphabet latin. En 1909 le Congrès d'Elbasan prend des décisions importantes : le choix de l'albanais comme langue officielle, la création d'écoles en langue albanaise, et la création de la première école normale pour former les enseignants. De nombreuses personnalités collaborent à la conception de grammaires et de

dictionnaires. De 1900 à 1920, on édite environ vingt grammaires (Osmani). La Commission littéraire de Shkodra de 1916-1917 codifie les premières règles, puis le Congrès de l'éducation de Lushnja, en 1920, établit les bases de l'orthographe. Il faut unifier la langue. Un premier projet est publié. « L'étape la plus marquante en vue de la normalisation date de 1967 avec la publication d'un ensemble de règles orthographiques […], visant à représenter une langue nationale uniforme » (Lloshi). Il sert de base à la rédaction de la directive de 1972 publiée dans le cadre du Congrès de l'orthographe (Kostallari).

La langue albanaise tend vers une écriture phonétique, mais certaines successions de lettres et certains sons peuvent donner lieu à diverses interprétations. Il faut des règles. Les règles d'orthographe sont conçues au sein d'un système cohérent et complet, en liaison étroite avec le système de la langue littéraire albanaise en général, et reflète sa structure phonétique, morphologique et lexicale. Le Congrès, composé de 87 membres, rédige une centaine de règles, qui sont mises en œuvre dans les grammaires et dictionnaires. La langue ainsi définie s'appelle la « langue littéraire », ce n'est pas encore la langue de l'homme de la rue. C'est en principe celle de l'administration, des médias et de l'école.

Nathalie Clayer relève que, durant les années 20, l'école en albanais peine à s'imposer face à l'aura de la culture et de la langue grecque. Les grécophones du Sud renâclent à renoncer à l'école en grec. Les tentatives d'instaurer des cours en albanais se heurtent à l'insuffisance de la formation des enseignants. La langue est réformée, mais elle est encore loin d'être la langue de tous. L'imprégnation va se faire au cours du XX° siècle. Alors qu'en 1945, il y avait 80 % d'analphabètes en Albanie, en 1991, 70 % des jeunes albanais ont terminé l'école secondaire.

C- La langue littéraire albanaise

L'albanais est une langue indo-européenne, écrite en caractères latins. L'alphabet comporte trente-six lettres, soit sept voyelles et vingt-neuf consonnes, dont neuf digrammes : dh, gj, ll, nj, rr, sh, th, xh, zh. Nous devons expliquer quelques aspects du codage informatique des textes. Les logiciels que nous utilisons ignorent les particularités de la langue albanaise. Les digrammes sont codés comme une suite de deux caractères, et, si cela simplifie le codage du texte, les logiciels de tri ne respectent pas l'ordre alphabétique de l'albanais. Cela complique notablement le travail de comparaison avec les dictionnaires papier. Concrètement c'est source d'erreurs et d'oublis. Par ailleurs, si certains caractères latins n'appartiennent pas à l'alphabet albanais, ils peuvent néanmoins être rencontrés lors de l'emprunt de mots étrangers. Prenons le cas du w. Dans les textes, nous avons trouvé walkman et New York. Le dictionnaire albanais transcrit ce dernier en Nju Jork, mais il semble se développer une tendance à emprunter les entités nommées étrangères, sans effectuer de transcription dans l'alphabet albanais. Il s'ensuit que l'alphabet utilisé en pratique est étendu à tout l'alphabet latin.

L'apostrophe est utilisée dans un nombre précis de cas, tels la négation « s' » et l'interrogation « ç' », ainsi que dans un nombre limités de contraction de clitiques. L'accent tonique est indiqué uniquement dans les dictionnaires. Quant à la ponctuation elle est similaire à la ponctuation du français.

Ja	pse	babi	rosaku	dhe	mami	rosa	ishin	të merakosur	dhe	
Voilà	pourquoi	papa	le canard	et	maman	la cane	étaient	inquiets	et	
prisnin	kur	do të çelte	veza	e paçelur	pak	më	e madhe	se	të	tjerat
attendaient	quand	ouvrirait	l'œuf	non-ouvert	un peu	plus	grand	que	les	autres

Exemple de texte albanais

La morphologie de l'albanais est principalement de type flexionnel. L'albanais connaît le masculin, le féminin, et le neutre - rare et en voie de disparition. Pour le traitement du syntagme nominal, nous avons pris en compte l'ordre nom/adjectif. Dans ce cas, le nom se décline mais seule l'éventuel « article de connexion » de l'adjectif se fléchit. L'ordre inverse -moins fréquent- adjectif nominalisé (et muni d'une flexion nominale) suivi du nom en position adjectivale n'est pas encore traité.

La lemmatisation reconnaît une dizaine de catégories, principalement verbe, nom, adjectif, adverbe, conjonction, préposition, pronom[1]… Des traits décrivent le nombre, le genre, le temps, la personne, le cas, etc.

III- Vers le traitement automatique de la langue albanaise

A- Ressources linguistiques pour la langue littéraire albanaise

Le travail avec la plateforme NooJ (Silberztein) requiert la description du vocabulaire de la langue traitée, et de toutes ses flexions. A cet effet, elle fournit des outils sous forme de grammaires de flexion et de dérivation. Nous avons présenté les travaux effectués (Piton Përnaska 2005), (Piton, Lagji, Përnaska 2007). Les formes sont décrites au moyen de catégories et de traits syntactico-sémantiques. Les listes obtenues sont compilées et enregistrées en tant que « dictionnaires électroniques ». Il ne s'agit pas ici de version électronique de dictionnaire papier, mais de l'enregistrement électronique « statique » des entrées fléchies de dictionnaires. Nos dictionnaires électroniques ont été créés à partir du Dictionnaire de l'Académie des Sciences de 1984 "Fjalor i shqipes së sotme" (Dictionnaire de l'albanais d'aujourd'hui) d'environ 34000 mots, rédigé sous la direction du professeur Androkli Kostallari, du "Fjalor shqip-frëngjisht" de 1998 (Dictionnaire albanais-français) de Vedat Kokona d'environ 35000 mots, et du lexique de « Parlons Albanais » de Christian Gut, Agnès Brunet-Gut et Remzi Përnaska (5285 mots). Ceci concerne la plus grande partie du vocabulaire.

Mais ceci ne suffit pas pour identifier tous les mots d'un texte. Certains sont construits à partir d'entrées et de règles de construction. Illustrons cette remarque sur le français. Le dictionnaire enregistre les verbes dire, faire, écouter, écrire, mais pas redire, refaire, réécouter, réécrire, ni rerefaire, reredire etc. Il en est de même en albanais. Ces mots requièrent un traitement dynamique que la plateforme NooJ peut effectuer. En albanais, ceci concerne aussi les expressions numériques, la préfixation de verbes, la déclinaison de certains noms propres et des sigles, et le traitement de phénomènes d'agglutination en albanais concernant l'impératif[2] ou la création dynamique de formes composées d'une expression numérique suivie d'une autre forme [3].

L'enregistrement des flexions, associé aux outils dynamiques mentionnés, fournit une base de graphies auxquels sont comparées les formes rencontrées lors de la lemmatisation.

La plateforme NooJ effectue cette lemmatisation, et produit la liste des formes non reconnues. Parmi elles se trouvent le plus souvent des noms propres, mais aussi les lacunes des dictionnaires, ainsi que les variantes lexicographiques qui sont l'objet de notre présentation.

Remarquons la spécificité des entités nommées en albanais. Elles se déclinent, et les termes

---

1 Ceci diffère de la grammaire albanaise qui regroupe sous une même catégorie adjectif et pronom. Il convient d'insister sur la spécificité du traitement linguistique en vue du TAL, qui impose des distinctions supplémentaires.

2 L'impératif à l'affirmatif agglutine, voire insère, le pronom accusatif : donne-moi: ***më*** *jep* > *jep**më*** ; donnez : *jepni* ; donnez-moi: ***më*** *jepni.*> *jep**mëni.*** ; donne-le lui ***ia*** *jep* > *jep**ia*** ; donnez-le lui ***ia*** *jepni* > *jep**ia**ni*.

3 Nous avons détaillé ces graphes utilisés dynamiquement pour les listes ouvertes –par exemple pour analyser les formes comme «n fishoj » qui signifie « n-upler », n étant un nombre cardinal (Piton et al. 2007).

d'origine étrangère, tels les noms de lieux ou de personnes, sont (en principe) transcrits afin de mettre leur prononciation en conformité avec la phonétique albanaise[4]. Leur flexion peut se présenter sous forme concaténée, ou séparée du mot par un tiret[5], comme dans le cas des sigles[6].

B- Corpus

Le corpus a été soit scanné, soit transmis par l'auteur sous forme de fichier, soit repris sur Internet. La plateforme NooJ importe chaque texte, et le code en format Unicode. Les textes choisis sont tous édités postérieurement à la réforme, mais la date de rédaction est inconnue. Ce sont : Kush e solli Doruntinën (Qui a ramené Doruntine), 1979, d'Ismail Kadaré ; Deklarata e përgjithshme mbi të drejtat e njeriut (la Déclaration universelle des droits de l'homme), 1995 ; Rreth botës për 80 ditë (Le tour du monde en 80 jours) de Zhyl Vern (Jules Vernes), 2002 ; Autobusi blu (L'autobus bleu), 2003, de Gérard Alba ; Fjalor i teologjisë biblike, (Vocabulaire de théologie biblique), 2005, de Xavier Léon Dufour ; Apokalips (Apocalypse), 2006, d'Enver Kushi ; Camera obscura : Origjina e botës, (La chambre obscure : l'origine du monde) 2007, de Luan Rama.
Le texte de la déclaration universelle des droits de l'homme est écrit dans la pure langue littéraire. Les autres présentent tous des particularités liées au domaine abordé –comme le vocabulaire théologique qui dans la langue littéraire fait références à des personnages et des événements historiques, et utilise beaucoup d'abréviations. La plupart des textes sont écrits dans une langue mâtinée d'emprunts dialectaux qui ne sont donc pas reconnus.

IV- Différences dialectales

A- La langue littéraire albanaise et ses dialectes

La langue littéraire n'est pas coupée de ses dialectes. Tout comme Gjinari et Beci, Anastas Dodi dresse une étude phonologique de leurs relations. Il estime que les plus anciennes traces attestées montrent que guègue et tosque furent proches, mais que « par suite du morcellement féodal aux XVII et XVIII[ee] siècles, les dialectes évoluèrent dans des voies divergentes. » Il distingue l'analyse savante du linguiste de la connaissance du locuteur moyen. « L'appartenance dialectale des affixes n'est transparente que pour le linguiste. » Le système affixal de la langue littéraire, qui a intégré des traits septentrionaux et des traits méridionaux interagit avec les dialectes, « les suffixes d'origine dialectale deviennent productifs ».
La place de l'accentuation joue un rôle essentiel en albanais. Pour les mots empruntés au turc, langue oxytonique, le tosque garde l'accentuation sur la dernière syllabe, alors que le guègue le déplace comme pour kafé qui devient káfe en guègue. L'unification de la norme de la prononciation standard, associée à certains choix lexicaux, a des conséquences sur la place et la « cohabitation » des voyelles e (dialecte septentrional) et ë (dialecte méridional). « Par exemple, à côté de brënda "dedans", il y a brendi "contenu"; mëndje "esprit" – mendoj "penser", mendim "pensée" ». « De même les groupes de voyelles ie, ye, ua, ue sont prononcés tels quels dans le dialecte méridional, tandis qu'ils sont monophtongués dans la plupart des parlers du dialecte septentrional en i:, y:, u. ». « L'intégration de ces formations aux traits dialectaux dans la langue littéraire est devenue plus facile à l'état actuel par le fait que ie, ye, ua, ue fonctionnent

4 François Mitterrand s'écrit *Fransua Miteran*. Jacques Chirac s'écrit *Zhak Shirak* et se déclinent : "une réunion secrète avec *Zhak Shirak**un***", "le parti de *Zhak Shirak**ut***" " la tombe du *Gjeneral**it*** (Général) à *Kolombei le dëzegliz* (Colombey-les-Deux-Eglises), en *Lorenë* (Lorraine)".

5 *Libri për De Golin* : le livre sur De Gaulle. ***in*** marque l'accusatif de De Gol.

6 *Skuadrat e SS**-ve*** **:** les escouades de SS. ***-ve*** marque le génitif de SS.

aujourd'hui non plus comme des diphtongues, dont ils tirent leur origine, mais comme des groupes de voyelles formés de deux voyelles à part qu'on peut diviser par la frontière syllabique. »
Concernant la distinction entre les consonnes r et rr, le dialecte guègue la maintient alors que le dialecte tosque a tendance à passer généralement de rr à r. « Même s'ils sont hétéro-dialectaux, les traits phonétiques s'intègrent à la structure de la langue en l'élargissant. »
Miço Samara se penche sur l'aspect lexical. Il constate l'interaction réciproque et les rapports entre la langue littéraire et les dialectes. Il fait « une appréciation générale du type de ces rapports et du lexique régional qui s'est intégré dans la structure de la langue littéraire, en tant que porteur des particularités des variantes littéraires du Nord et du Sud en matière de phonétique, de formation des mots et de sémantique. »
Ceci rejoint nos observations sur notre corpus de textes postérieurs à la réforme. La langue littéraire a inséré des éléments dialectaux divers. Elle est plus diversifiée que ses dialectes et apporte au locuteur des éléments lexicaux porteurs de traits phonologiques ou affixaux auxquels il applique des paradigmes de dérivation et de flexion de la langue littéraire, voire du guègue ou du tosque.

B- Traits distinctifs entre le guègue et le tosque

Ces trais sont souvent à l'origine de variations rencontrées entre les formes non reconnues et les formes standards. L'ouvrage de Gut, Brunet-Gut et Përnaska établit une liste de différences entre le guègue et le tosque[7]. Nous en reprenons ici quelques-unes, les illustrerons sur nos observations et montrerons quelques-uns des moyens permettant de les traiter.

Le rhotacisme

Il s'agit du *n* intervocalique du guègue qui est souvent rhotacisé en tosque : *-n-* > *-r- zuna* > *zura.*

Le groupe vocalique *ue/ua*

Certaines formes en *ua-* du tosque -ou du littéraire- correspondent en guègue à *-u-, -ue-, -aue-, -ou-.*

La nasalisation

La voyelle *ë* accentuée du tosque -ou du littéraire- suivie des consonnes *m* ou *n* se nasalise en guègue. Nos textes en attestent. La langue littéraire ne mentionne pas la nasalisation.

Groupes consonantiques

Aux groupes de consonnes *mb, nd, ng* du tosque correspondent en guègue à *m, n* ; au groupe final *nj* correspond *j* en guègue.

Accentuation

L'accentuation est différente pour certains mots : *baba, disa*, cela a des conséquences sur la flexion : *babanë/babain*.

L'infinitif

En guègue l'infinitif périphrastique du type *me* + *Participe Passé* est remplacé en tosque et en langue littéraire par le subjonctif ou un infinitif du type *për të* : *me punue* > *të punoj* ou *për të*

7 Lloshi décrit en une page « les différences dialectales les plus frappantes » entre le guègue et le tosque.

punuar. De manière générale, le *participe passé* est utilisé non seulement pour les temps composés, mais aussi pour l'infinitif. Les affixes utilisés varient selon le dialecte.

La conjugaison

Il existe des différences de conjugaison. En guègue le futur est formé par le présent du verbe avoir suivi de l'infinitif guègue, tandis qu'en tosque et en langue littéraire le futur est formé par la particule do suivie du subjonctif : *kam me ardhë > do te vij*. Certains verbes en *t* à la première personne du présent de l'indicatif sont en *s* pour les guègues. La langue littéraire a parfois opté pour la forme tosque et parfois pour la forme guègue. Il s'ensuit que les auteurs inclinent à utiliser les deux : *zbritnin/zbrisnin.*

Le pluriel

Aux pluriels tosques en *–ë* correspondent des pluriel guègues en *–a.* il s'ensuit que les deux sont utilisés dans les textes.

C- Variations observées

Laissant de côté les fautes de frappe, à côté des variations qui s'expliquent par les règles listées plus haut, nous avons rencontré de nombreuses erreurs de substitution de morphèmes à la phonétique proche -par exemple entre les sonnantes sourdes-, des ajouts ou suppressions de e muet (ë), des concaténations, et des erreurs portant sur les signes diacritiques. Nous en donnons des exemples dans un tableau.

Forme A	Forme B	A mis pour B	B mis pour A
	ë	*malsor* mis pour *malësor* *puntor* mis pour *punëtor*	*ardhëshme* mis pour *ardhshme* *kokëulur* mis pour *kokulur*
a	o	*labaratorët* mis pour *laboratorët*	*mitroloz* mis pour *mitraloz*
ai	aji	*kallai* mis pour *kallaji*	
ç	c	*elektriçist* mis pour *elektricist*	
ç	sh	*çpoi* mis pour *shpoi*	
ck	sk	*guacka* mis pour *guaska*	
d	t	*gjigande* mis pour *gjigant*	*limonatë* mis pour *limonadë*
e	ë	*shembëllim* mis pour *shëmbëllim*	*ëngjëj* mis pour *engjëj*
ë	a	*rrezëtinin* mis pour *rrezatonin* *buzëgaz* mis pour *buzagaz*	
f	jf	*qef* mis pour *qejf* *çakërrqef* mis pour *çakërrqejf*	
g	k	*guzhinës* mis pour *kuzhinë*	*karrike* mis pour *karrige*
g	gj	*gipsi* mis pour *gjipsi*	
ia	ja	*familiar* mis pour *familjar* *lagjia* mis pour *lagjja*	*vetmja* mis pour *vetmia*
ij	i	*amullije* mis pour *amullie* *mamija* mis pour *mamia*	*makiazh* mis pour *makijazh*
(n)je	(n)ie	*djela* mis pour *diela* *kuzhinjerët* mis pour *kuzhinierët* *djegjes* mis pour *djegies*	
k	q	*rrëshkisnin* mis pour *rrëshqitnin*	
l	ll	*stambolitë* mis pour *stambollitë*	*llotari* mis pour *lotari*
nj	gn	*manjetizimi* mis pour *magnetizimi*	
o	oo	*alkolet* mis pour *alkoolet*	
p	b	*llampadarë* mis pour *llambadar*	

r	rr	*pasthirmë* mis pour *pasthirrmë*	*parreshtur* mis pour *pareshtur*
s	z	*entusiazmuar* mis pour *entuziazmuar*	*autobuz* mis pour *autobus*
sh	s	*realisht* mis pour *realist*	
th	dh	*drithma* mis pour *dridhmë*	
xh	gj	*xhestit* mis pour *gjestit*	
xh	zh	*xhveshur* mis pour *zhveshur*	
y	yj	*hynia* mis pour *hyjnia*	
zh	sh	*zhkarravinat* mis pour *shkarravinat*	*shurmshëm* mis pour *zhurmshëm*

Table 1 : variations relevées

Nous trouvons beaucoup d'erreurs multiples sur un même terme : endëroj mis pour ëndërroj, et endra mis pour ëndrra, çakërqef mis pour çakërrqejf, çirozë mis pour cirrozë. Notons également que certaines erreurs proviennent de transcription de termes étrangers – tels Jeruzalem mis pour Jerusalem, alkolet mis pour alkoolet ou mitroloz mis pour mitraloz -. Parmi ces variations, les plus fréquentes sont, de loin, celles concernant le ë – insertion, suppression, ou confusion avec e-, la confusion entre r et rr, et les erreurs de conjugaison, de participe passé, ou de pluriel.
La méthodologie développée précédemment en vue de la reconnaissance automatique de variantes archaïques du XVII siècle de l'anglais, nous est apparue transposable aux variations décrites ici. Nous allons présenter les principes généraux de cette méthode.

V- Variantes lexicales

A- L'outil dynamique

Nous formons l'hypothèse que la méthode employée pour une étude effectuée dans une approche diachronique, peut être reprise dans une approche synchronique, pour fournir une méthodologie de traitement de variations lexicographiques d'une langue dont les variations régionales perdureraient. La plateforme NooJ fournit, par ses grammaires morphologiques d'annotation, un outil apte à traiter de variations affixales. Présentons la méthode sur le français. La réforme de 1990 instaure des modifications orthographiques comme : « Le i suivant ll ne s'entendant pas, on peut écrire -iller et -illère dans joaillier, marguillier, quincaillier, serpillière ».
Nous la transcrivons ainsi : « si un lexème qui devrait se terminer par la séquence illier/s (ou illière/s) se termine par la séquence iller/s (ou illère/s), alors il est validé dynamiquement au moyen d'une grammaire morphologique ». Soit une telle grammaire G1.

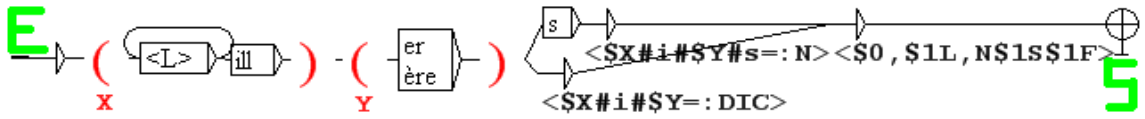

Figure 2 : Graphe morphologique d'annotation G1

Le principe est de fractionner le mot étudié selon des séquences telles X et Y ci-dessus, puis ces séquences modifiées sont testées par comparaison avec l'ensemble des dictionnaires électroniques, noté DIC, ou avec certaines catégories. Des traits spécifiques peuvent être mentionnés dans cette comparaison. Si la vérification est validée, alors une entrée est créée dynamiquement.
Le chemin débute au point d'entrée E (flèche située à l'extrémité gauche), et s'achève au point de sortie S (ballon situé à l'extrémité droite). L'hypothèse de travail est que la plateforme NooJ est utilisée pour le français, et dispose de dictionnaires électroniques comportant les mots joaillier, marguillier, quincaillier, serpillière, mais ignore les nouvelles formes joailler, marguiller,

quincailler et serpillère. Le traitement usuel conduit à l'inscription de ces mots dans une liste de formes inconnues. Par exemple : joailler,UNKNOWN.
La forme n'est pas reconnue par l'ensemble des dictionnaires. Le graphe G1, qui est affecté d'une priorité inférieure à celle des dictionnaires, est alors appliqué. Un mot est traité par ce graphe s'il existe un chemin de reconnaissance qui permet de passer de l'entrée E à la sortie S. Soumettons-lui le mot joailler. Le mot est au singulier, c'est le chemin inférieur qui est suivi. Le graphe découpe le mot à l'endroit où « manque » le i, et utilise les variables $X et $Y. (symbole $ n'est pas visible, mais il est présent.) Le symbole # exprime la concaténation. DIC représente l'ensemble des dictionnaires électroniques dont nous disposons.

« joaill » « er »
↓ ↓
$X#i#$Y joaillier

$X#i#$Y représente donc le mot joaillier. L'expression <$X#i#$Y=:DIC> teste si le mot « joaillier » est une forme présente dans DIC. Dans le cas négatif le parcours s'arrête. Dans le cas positif le parcours se poursuit et le mot est étiqueté comme une variante, : joailler est une variante de joaillier[me]. Remarquons que les séquences découpées $X et $Y ne sont pas des mots du dictionnaire, c'est l'expression entre crochets qui indique quel est le mot recherché.

B- Traitement de variantes lexicales en Albanais

Les dictionnaires de la langue littéraires s'avérant insuffisants pour traiter les textes, il faut envisager d'enrichir les dictionnaires, ce qui nécessite de disposer de la liste des formes à insérer[8] et opérer une transformation coûteuse en temps. Nous avons une autre possibilité : observant de nombreuses régularités dans les transformations opérées, nous pouvons construire des règles. Nous devons nous doter d'un moyen d'analyse des formes inconnues. Ceci nous conduit à décrire la transformation avec un nombre restreint de propriétés : la position, la transformation et la condition[9].

Position
La transformation est en position préfixale, suffixale ou infixale.
Transformation
La séquence concernée est une séquence de lettres.
La transformation peut être une insertion, une suppression, ou une modification de séquence.
Condition : Test à effectuer sur le mot.
Ce test peut concerner l'ensemble des dictionnaires notés DIC, ou bien une catégorie telle que codifiée dans les dictionnaires, catégorie à laquelle peuvent être adjoints des traits.

Règle 1 : La description de la position, la condition et la transformation.

Ceci est compatible avec l'outil dynamique décrit plus haut. Nous disposons donc de règles décrivant les conditions et les transformations, quant à la position elle est à décrire explicitement dans le graphe. De plus nous pouvons décrire la succession de plusieurs transformations. Nous

me La séquence <$0,$1L,C1S$1F> crée une annotation comportant la séquence traitée –notée $0-, le lemme identifié –noté $1L-, et la catégorie –notée $C- de ce lemme suivie de tous les traits syntaxiques –notés $1S- et sémantiques –notés $1F- de la forme reconnue, soit : « joailler,joaillier,N+Genre=m+Nb=s ».
8 Ceci suppose le problème résolu.
9 Nous observons des similitudes avec le travail développé à l' ATILF CNRS par Souvay et Pierrel pour LGeRM Lemmatisation des mots en moyen français.

détaillons la méthode sur quelques cas.

Traitement du participe passé guègue

Nous avons observé quelques formes de participes passés guègues dans les textes étudiés. Ils sont regroupés dans la Table 2. En langue littéraire et en tosque, les verbes en *oj* (*lulëzoj* fleurir) ont leur participe passé en *uar* – *lulëzuar*-, tandis qu'en guègue ils sont en *ue* : *lulëzue*, voire *lulzue*.

Concordance for Text Apokalips initial.not

Reset | Display: 6 | characters / word forms | before, and 5 after. Display: Matches | Outputs

Text	Before	Seq.	After
	kena dalë në Zallin e Kirit`.'...	Jam ligshtue	boll për sëmundjen tande... Kam
	koha, si ato letra. Princi im!	Ka ardh	pranvera dhe në kopsht ka
	Ka ardh pranvera dhe në kopsht	ka lulzue	kajsia. ...Asht mesnatë, i dashur
	e di, tha Princi. Shpejt kam	me ardhë	ndryshe në ato vise, tha
	shpresë me t'pa... A mundesh	me dërgue	fotografinë që kena dalë në
	boll për sëmundjen tande... Kam shpresë	me u takue	këtë verë. Më ka marrë

Query 6/6

Table 2 : Exemples de participes passés guègues

Position :	La transformation est en *position suffixale.*
Transformation :	*ue* > *uar*
Condition	La forme construite est reconnue comme le participe passé d'un verbe : V+PP

Règle 2 : Traitement du participe passé guègue

Position 1 :	La transformation est en *position infixale.*
Transformation 1 :	insertion de *ë*
Position 2 :	La transformation est en *position suffixale.*
Transformation 2 :	*ue* > *uar*
Condition	La forme construite est reconnue comme le participe passé d'un verbe : V+PP

Règle 3 : Traitement du participe passé guègue avec insertion de ë

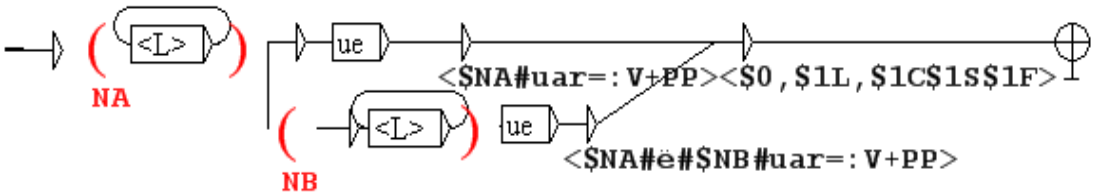

Figure 3 : Graphe morphologique d'annotation G2

La règle 2 permet de reconnaître *takue* mis pour *takuar,* ou *lulëzue* mis pour *lulëzuar*, tandis que la règle 3 permet de traiter *lulzue*.

Nous exhibons ci-dessus le graphe morphologique G2 correspondant aux règles 2 et 3. Le chemin supérieur traduit la règle 2 et peut traiter la forme *lulëzue*. Comme expliqué précédemment, le mot est découpé afin d'isoler et remplacer la partie à transformer.

« *lulëz* »
↓
$NA#uar *lulëzuar*

L'expression $NA#uar représente le mot « *lulëzuar* ». « V+PP » est un verbe au participe passé. La séquence <$NA#uar=:V+PP> recherche si « *lulëzuar* » est reconnu comme un verbe au participe passé. La réponse est positive[10].
Précisons qu'en albanais de nombreux verbes ont une forme directe qui peut être transitive ou intransitive, et une forme réflexive (ou médio passive) qui partagent le même participe passé. Il s'agit ici respectivement de *lulëzoj* (fleurir) et de *lulëzohem* (être fleuri, s'épanouir). Le passé composé de la forme directe se conjugue avec l'auxiliaire avoir (*kam*), tandis que le passé composé de l'autre forme se conjugue avec être (*jam*). Cette distinction syntaxique ne peut se faire au niveau morphologique, mais pourra se faire lors d'une étape syntaxique ultérieure.
La variante du texte comporte une double différence : *lulzue>lulëzue>lulëzuar*. Ceci correspond à la règle 3, c'est le chemin inférieur de G2 qui la transcrit. La séquence est fractionnée ainsi :

« *lul* » « *z* »
↓ ↓
$NA#ë#$NB#uar *lulëzuar*

L'expression $NA#ë#$NB#uar représente le mot « *lulëzuar* » . La séquence est reconnue et l'annotation est créée.

Texte initial	*Ka* ***ardh*** *pranvera dhe në kopsht ka* ***lulzue*** *kajsia*
Transcription en langue littéraire	*Ka* ***ardhur*** *pranvera dhe në kopsht ka* ***lulëzuar*** *kajsia*
Traduction	« a venu » le printemps et dans le verger a fleuri l'abricotier
Annotations (extrait)	**ardh**,vij,V+intrans+Temps=PP+FR="venir" **lulzue**,lulëzoj,V+intrans,+Temps=PP+FR=fleurir **lulzue**,lulëzohem,V+Temps=PP+NA+refl+FR="s'épanouir"

Table 3 : Passage guègue

Si l'alternative *ua/ue* est répandue, cela tient au grand nombre de verbes en *oj/ohem* qui sont concernés. Actuellement ils sont respectivement enregistrés dans notre liste au nombre de 2086 et 918, sur un total de 7175, soit 42% des verbes. Nous avons aussi rencontré d'autres variations pour les verbes telle *shkrumbosur* pour *shkrumbuar -brûlé*.
Il faut se garder d'une analyse superficielle, car nous trouvons (sur Internet) le verbe *shkrumbos*. L'explication tient à l'existence du doublon *shkrumbos* (participe passé *shkrumbosur)* mis pour *shkrumboj* (participe passé *shkrumbuar*), et même du triplet *shkrumbëzoj*, (participe passé *shkrumbëzuar*). La langue littéraire a retenu *shkrumboj* et sa variante *shkrumbëzoj*. L'utilisation de *shkrumbos* reste malgré tout naturelle pour le locuteur qui ne vérifie pas chaque mot dans le dictionnaire. Ce cas illustre notre propos, et nous permet de préciser les limites du traitement automatique que nous proposons, qui n'a pas vocation à être assimilé à une étude linguistique. Nous traitons des textes soumis, nous ne faisons pas l'exégèse du texte. C'est donc bien une transcription simplifiée que nous proposerons dans ce cas. Sans exhiber l'existence de *shkrumbos*, nous proposerons la reconnaissance de la forme non littéraire *shkrumbosur* comme une forme utilisée « à la place » de *shkrumbuar*, mais nous sommes conscients que ceci est un traitement et pas une explication, laquelle inclurait l'existence du verbe *shkrumbos* non retenu

10 La forme est annotée par la séquence <$0,$1L,C1S$1F>. Il serait aisé d'introduire un trait syntaxique supplémentaire tel +geg pour indiquer que la forme reconnue est guègue. Il suffirait de remplacer la séquence <$0,$1L,C1S$1F> par la séquence <$0,$1L,$C$1S$1F+geg>.

dans la langue littéraire. La situation de doublon entre un verbe de la langue littéraire et un verbe dialectal est très fréquente. La variation *osur/uar* est incluse dans nos règles.

Traitement du rhotacisme et de la nasalisation

Nous avons *zanin* mis pour *zërin*. *Zanin* est découpé en trois.

« *z*»	« *in* »	
↓	↓	
$NA#ër#	$NB	*zërin*

Rhotacisme et nasalisation : *zërin* > *zanin*. Nous le traitons par le chemin inférieur du graphe. La première sortie du chemin inférieur teste la séquence sans rhotacisme : $NA#ën#$NB soit *zënin*, il n'est pas trouvé dans le dictionnaire, d'où échec. La deuxième sortie teste la séquence avec rhotacisme $NA#ër#$NB soit *zërin*. Il trouve l'accusatif déterminé de *zë, zëri* (la voix). Et crée l'annotation : zanin,zë,N+Cas=kallëz+TDef=shquar+Genre=m+Nombre=s.

Position :	La transformation est en *position préfixale.*
Transformation :	an > ër
Condition	La forme construite est reconnue comme une forme du dictionnaire : DIC

Règle 4 : Traitement du rhotacisme avec nasalisation en position préfixale

Position :	La transformation est en *position préfixale.*
Transformation :	an > ën
Condition	La forme construite est reconnue comme une forme du dictionnaire : DIC
Exemple	*andërroj* > *ëndërroj*

Règle 5 : Traitement de la nasalisation en position préfixale

Position :	La transformation est en *position infixale.*
Transformation :	an > ër
Condition	La forme construite est reconnue comme une forme du dictionnaire : DIC
Exemple	*zanin* > *zërin*

Règle 6 : Traitement du rhotacisme avec nasalisation d'un infixe

Position :	La transformation est en *position infixale.*
Transformation :	an > ën
Condition	La forme construite est reconnue comme une forme du dictionnaire : DIC
Exemple	*tand* > *tënd*

Règle 7 : Traitement de la nasalisation en position d'un infixe

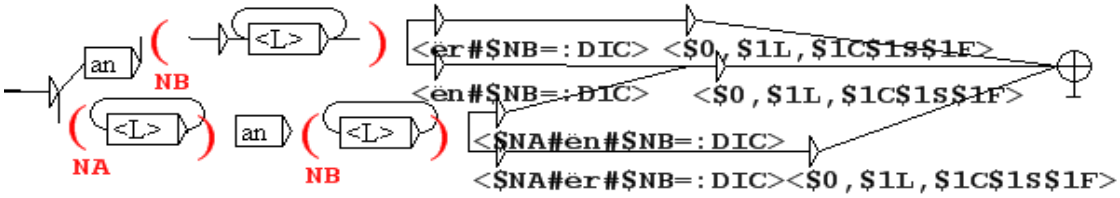

Figure 4 : Graphe morphologique d'annotation G3

Le lexème *andërroj* est testé par le chemin supérieur. NB reçoit le fragment de mot qui suit *an*, soit *dërroj*. Deux chemins sont proposés. Le premier teste la séquence avec rhotacisme soit :

ër#$NB soit *ërdërroj*. Il y a échec. Le deuxième chemin teste la séquence sans rhotacisme ën#$NB soit *ëndërroj,- je rêve-* et crée deux annotations, l'une pour le subjonctif et l'autre pour l'indicatif soit : andërroj,ëndërroj,V+Ind+intrans+Temps=PR+Perso=1+Nombre=s+FR="rêver"

Texte initial	***andërroj*** *të dëgjoj* ***zanin*** ***tand*** *të* ***ambël***
Transcription en langue littéraire	***ëndërroj*** *të dëgjoj* ***zërin*** ***tënd*** *të* ***ëmbël***
Traduction	je rêve d'entendre voix tienne suave
Annotations (extrait)	andërroj,ëndërroj,V+Ind+intrans+Temps=PR+Perso=1+Nombre=s+FR=rêver zanin,zë,N+Cas=kallëz+TDef=shquar+Genre=m+Nombre=s+FR=voix. tand,yt,DET+Poss+Genre=m+Nombre=s+Perso=2+kallëz+FR=ton ambel,ëmbël,A+Genre=m+Nombre=s+FR=sucré

Table 4 : Passage guègue avec rhotacisme et nasalisation

Traitement du pluriel

L'échange entre le pluriel en a et le pluriel en ë se traite simplement : il suffit de pouvoir reconnaître en fin de nom pluriel *ë, ët, ëve, ësh* au lieu de *a, at, ave, ash*, ou réciproquement comme dans notre exemple où le pluriel de *vagon* (wagon) en *ë* a été remplacé par un pluriel en *a*. (Dans le même texte, nous observons, pour ce mot, six occurrences de pluriel en *a,* et cinq occurrences de pluriel en *ë*.) Nous avons une règle pour chacune de ces quatre formes, et quatre règles symétriques pour la transformation *a* > *ë*. Nous en présentons une en Règle 8. Le graphe global correspondant est le graphe morphologique G4.

Position :	La transformation est en *position suffixale.*
Transformation :	ëve > ave
Condition	La forme construite est reconnue comme un nom au pluriel : N+p
Exemple	*vagonave* > *vagonëve*

Règle 8 : Traitement du pluriel ave > ëve.

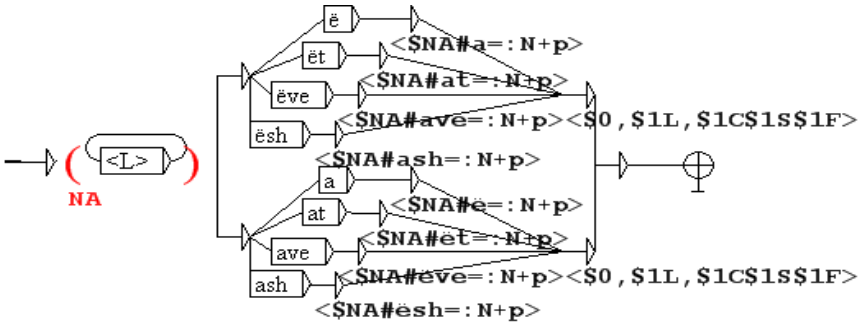

Figure 5 : Graphe morphologique d'annotation G4.

Texte initial	*poshtë* ***vagonave*** poshtë + ablatif (*rrjehdhore*) *Nga çatitë e* ***vagonave*** génitif (*gjinore*) *zbritën nga* ***vagonat*** nominatif (*emërore*)
Transcription en langue littéraire	*poshtë* ***vagonëve*** (sous les wagons) *Nga çatitë e* ***vagonëve*** (des toits des wagons) *zbritën nga* ***vagonët*** (descendirent des wagons)
Annotations (extrait)	vagonave,vagon,N+Cas=rrjedh+TDef=pashquar+Genre=m+Nombre=p vagonave,vagon,N+Cas=gjin+TDef=pashquar+Genre=m+Nombre=p vagonat,vagon,N+Cas=emër+TDef=shquar+Genre=m+Nombre=p

Table 4 :Occurrences de pluriel en a (at, ave) au lieu de ë (ët, ëve)

Traitement des erreurs de graphie

Nous présentons dans la Figure 6 quelques chemins aptes à traiter le remplacement de consonnes. Nous avons systématiquement opéré les deux remplacements symétriques, par exemple s mis pour z, aussi bien que z mis pour s.

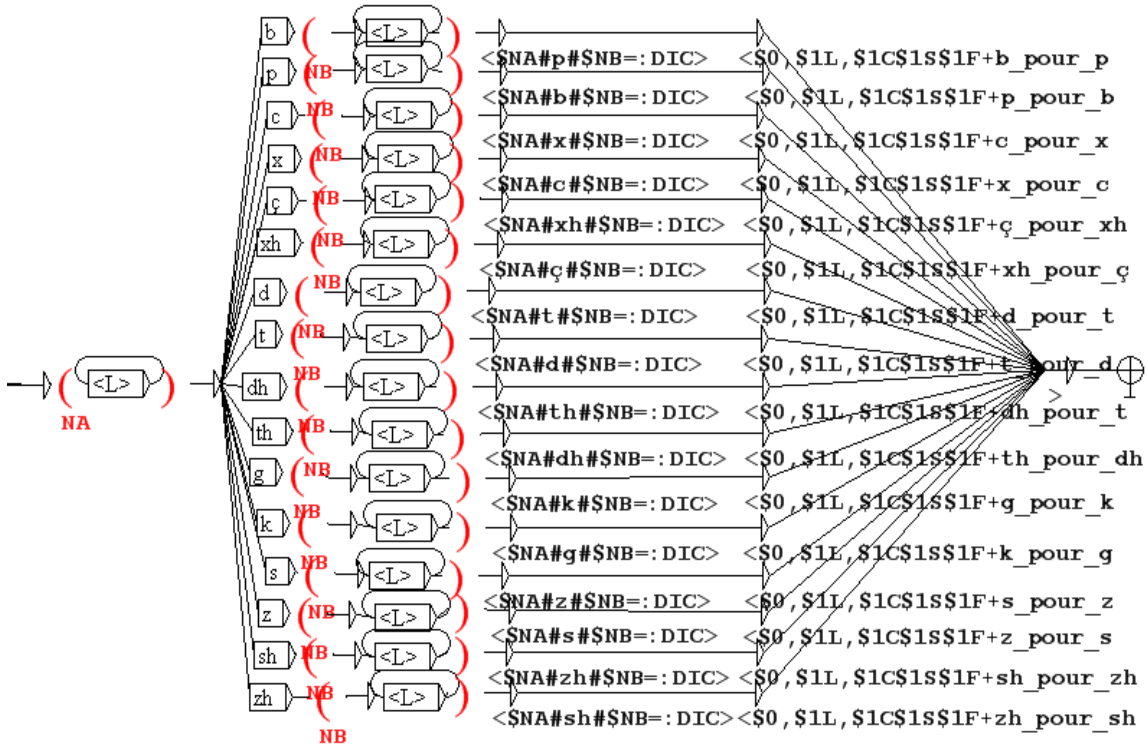

Figure 6 : Graphe morphologique G5

En pratique nous devons démultiplier les graphes, car les préfixes, infixes ou suffixes sont traités par des graphes (ou des chemins) distincts. De plus il faut également prévoir les cas extrêmement fréquents de transformations multiples, par exemple l'insertion ou la suppression d'un ë. A chaque règle est associé un chemin. La Figure 6 montre un exemple de regroupement de seize chemins traitant seize règles.

Le graphe G5 ne peut traiter que de séquences insérées. Nous montrons dans la Figure 7 le Graphe morphologique G6 qui permet de traiter les préfixes et infixes pour l'alternance *rr/r*. *pasthirmë > pasthirrmë, parreshtur> pareshtur.*

Position :	La transformation est en *position préfixe.*
Transformation :	r > rr
Condition	La forme construite est reconnue comme une forme du dictionnaire
Exemple	*rezatoni > rrezatoni*

Règle 9 : Traitement de l'alternance r > rr en position préfixe

Position :	La transformation est en *position infixe.*
Transformation :	rr > r
Condition	La forme construite est reconnue comme une forme du dictionnaire
Exemple	*parreshtu r> pareshtur*

Règle 10 : Traitement de l'alternance rr > r en position infixe

Les deux chemins supérieurs traitent des préfixes et les deux autres des infixes. Nous avons l'exemple de *rradhazi* à la place duquel il faut reconnaître *radhazi* (tour à tour), et le cas de *rezatoni* écrit pour *rrezatoni* (vous irradiez). L'erreur est au début du mot, ce sont les chemins 1 et 2 qui vont permettre d'isoler la séquence $NB=*adhazi*, et de tester la séquence r#$NB. Dans le second cas on isole $NB=*ezatoni* et on teste l'existence de rr#$NB.

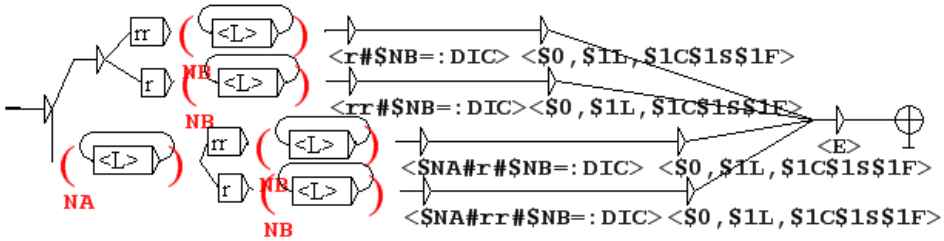

Figure 7 : Graphe morphologique G6

Les deux derniers chemins permettent de traiter les cas de *pasthirmë* mis pour *pasthirrmë* (interjection), et de *parreshtur* mis pour *pareshtur* (sans cesse). Détaillons le premier. La séquence est fractionnée en trois, $NA reçoit *pasthi* tandis que $NB reçoit *më*. C'est la séquence $NA#rr#$NB qui est testée et reconnue.

C- Mise en œuvre

Les règles concernant les positions préfixes et infixes sont générales, tandis que celles qui concernent les suffixes se rapportent souvent à une catégorie bien définie. Nous avons environ deux cent cinquante règles. Leur nombre augmente avec les observations.

La contrainte de cette méthode est de devoir contrôler les résultats obtenus par graphes. Plus le lexème traité est long, plus la probabilité d'avoir une réponse exacte est grande. A contrario, pour les lexèmes très courts cette méthode produit beaucoup de bruit. Nous disposons de possibilités pour le limiter. Il s'agit non seulement de développer des graphes de désambiguation, que nous ne présentons pas ici, dont le rôle est de bloquer les successions syntaxiquement interdites, et, ce faisant, d'éliminer une partie du bruit ; nous disposons d'une autre possibilité. En effet, les différents dictionnaires et graphes peuvent être organisés selon une hiérarchie de priorité[11]. La recherche de la concordance débute au niveau de priorité le plus élevé. En cas de succès, la recherche s'arrête après exploration totale des graphes du même niveau, toutes les correspondances sont listées, et la recherche cesse. Si aucune correspondance n'a été trouvée, NooJ opère la recherche parmi les graphes ou dictionnaires de niveau inférieur. Ceci se poursuit jusqu'au niveau de priorité le plus faible. Les formes non reconnues n'ont été trouvées à aucun niveau. Ceci permet de limiter le bruit, mais la répartition hiérarchique est à étudier soigneusement.

VI- Conclusion

La variabilité lexicale de la langue albanaise est réelle, nous avons même plusieurs fois observé des variations pour un même mot dans un même texte. Nous constatons que l'unification n'est pas achevée, et peut-elle l'être ? Imposée par un pouvoir autoritaire, elle reste soumise à la réalité de la langue qui ne se décrète pas.

Les dialectes ont développé des habitudes de résistance à l'environnement contraint qui fut le leur

11 Nous l'avons évoqué lors de la présentation du graphe G1.

durant les quatre siècles de l'occupation ottomane. Leur différenciation phonétique s'est confortée au cours du temps. Les règles de flexion et de dérivation lexicales, basées à la fois sur la morphologie et la phonétique ont produit des paradigmes différenciés que les locuteurs appliquent quotidiennement dans la sphère orale. Ce n'est pas en Albanie que l'on se serait gaussé des néologismes « bravitude » ou «abracadabrantesque », pas plus que l'on ne se serait préoccupé de les traiter de barbarismes dès lors qu'ils concatènent un suffixe et un radical à la sémantique compatibles. Les paradigmes ont contribué au maintien, à l'enrichissement et à la vitalité du guègue et du tosque, en compensant autant que possible l'absence de règles écrits.
La normalisation de la langue ne peut tout imposer. D'une part certains usages se maintiennent, et d'autre part les textes peuvent comporter des dialogues. Les traits régionaux sont une richesse de la langue, et donnent du relief aux textes. Ainsi, même si l'influence des médias est un facteur puissant d'évolution vers une langue commune, leur utilisation est-elles appelée à perdurer. Les outils de traitement automatique ne peuvent se contenter de traiter la langue littéraire, mais doivent intégrer des outils à même d'élargir le traitement.

Bibliographie

Clayer N., L'Albanisation de la Zone Frontière Albano-Grecque et ses Aléas dans l'Entre-Deux-Guerres, in "Südost-Forschungen 68 (2010) 328-348"

Dodi A., La Langue Littéraire Albanaise et ses Dialectes au Niveau Phonologique, in P.U.L.A. Nos ¾, CORTI 90, Actes du Colloque International des Langues Polynômiques, Université de Corse, 17-22 septembre 1990, J. Chiorboli Ed. 118-121

Doja A., Entre Invention et Construction des Traditions : l'Héritage Historique et Culturel des Albanais, in Nationalities Papers: The Journal of the Association for the Study of Nationalities 28, 3 (2000) 417-448

Dozon A., La Littérature Populaire chez les Chkipes ou Albanais, in Bulletin de Correspondance Hellénique Année 1878, Vol. 2 N°2 pp. 45-53

Gut Chr., Brunet-Gut A., Përnaska R., Parlons Albanais Paris L'Harmattan (1999)

Kabashi B., Modellierung und Implementierung einer LA-Grammatik für ein Fragment des Albanischen. Unpublisht Manuscript. Computerlinguistik, Univ. Erlangen-Nürnberg,1999, Automatische Wortformerkennung für des Albanische, Master's thesis in 'Linguistische Informatik' (Computational Linguistics), Univ. Erlangen-Nürnberg, 2003,

Kabashi B., Analiza Automatike e Fjalëformave të Gjuhës Shqipe, Seminari Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare, XXXIII, v. 23/1, Prishtinë, 2005,

Kabashi B., Disa propozime për Modelimin e Informacionit në Leksikografinë Kompjuterike, Seminari Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare, XXXIV, Prishtinë, 2005,

Kabashi B., Zeichen für Gjon Buzuku. Die Zusammenarbeit zwischen der Albanischen Linguistik, Nach 450 Jahren. Buzukus 'Missale' und seine Rezeption in unserer Zeit (Albanische Forschungen 25), Wiesbaden, 2007,

Kabashi B., Pronominal Clitics and Valency in Albanian. A Computational Linguistics Perspective and Modelling within the LAG-Framework, Valency. Theoretical, Descriptive and Cognitive Issues, Berlin, 2007

Kabashi B., Ein Valenzlexikon für das Albanische, in « 50 Jahre Albanologie an der Ludwig-Maximilians-Universität München », (23.-25.06.2011)

Kostallari A., Domi M., Çabej E., Lafe E., Drejtshkrimi i Gjuhës Shqipe, *(Orthographe de la langue albanaise)* 1972, Ribotim i Institutit Albanologjik të Prishtinës 1974. http://www.drejtshkrimi.net/

Lagji K., Etude sur le Statut du Mot en Albanais dans le Cadre des Traitements Automatiques des Langues. In Annotation Automatique de Relations Sémantiques et Recherche d'Informations: vers de Nouveaux Accès aux Savoirs. Université Paris-Sorbonne, 27-28 octobre, Paris, (2006)

Lloshi Xh., Albanian, in Handbuch der Südosteuropa-Linguistik, Uwe Hinrichs,Uwe Büttner Ed., 1999, Harrassowitz Verlag . Wiesbaden, 277-299

Murzaku A., Inverse Dictionary of Albanian, 32,005

Murzaku A., Finding Patterns in Texts: a Quantitative Analysis of Qualitative Language Changes, in « 50 Jahre Albanologie an der Ludwig-Maximilians-Universität München », (23.-25.06.2011)

Osmani T., Priku M., Njësimi i Alfabetit dhe i Gjuhës Letrare së Përbashkët–Dy Probleme Gjuhësore Parësore të Viteve të para të Shekullit XX, in « 50 Jahre Albanologie an der Ludwig-Maximilians-Universität München », (23.-25.06.2011)

Piton O., Lagji K., Përnaska R., Electronic Dictionaries and Transducers for Automatic Processing of the Albanian Language. NLDB 2007. Springer Verlag Eds. LNCS 4592. pp 407-413.

Piton O., Mesfar S., Pignot H., Automatic Transcription of 17th Century English into Contemporary English with NooJ: Method and Evaluation, NooJ 2011, Dubrovnik, http://hal.archives-ouvertes.fr/hal-00625884/fr/

Piton O., Përnaska R., Processing of Morphological Variations and Grammar of Noun Phrases for Disambiguation in Albanian, NooJ 2011, Dubrovnik

Piton O., Përnaska R., Constitution de Dictionnaires Electroniques pour l'Albanais, et Grammaire du Groupe Nominal avec Nooj, Belgrade (2006)

Piton O., Përnaska R., Etude de l'Albanais en Vue de Construire des Outils pour son Traitement Automatique, Journées NooJ Besançon (2005)

Roux M., Les Albanais de Yougoslavie. Minorité nationale, Territoire et Développement, Paris, Ed. Maison des Sciences de l'homme, 1992. (Thèse à Toulouse le Mirail en 1990)

Samara M., Les Rapports Réciproques entre la Langue Littéraire Albanaise et les Dialectes dans le Domaine du Lexique, in P.U.L.A. Nos ¾, CORTI 90, Actes du Colloque International des Langues Polynomiques, Université de Corse, 17-22 septembre 1990, J. Chiorboli Ed. 323-327

Silberstein M., 2005 NooJ's dictionaries, Proceedings of LTC (2005), Poznan University.S

Souvay G., Pierrel J.M., 2009, LGeRM, Lemmatisation des Mots en Moyen Français, TAL. Volume 50 – n° 2/2009, 149-172

Trommer, J. Kallulli.D., A Morphological Tagger for Standard Albanian. In Proceedings of LREC 2004.

Textes

Kush e solli Doruntinën (Qui a ramené Doruntine), 1979, d'Ismail Kadaré, Ed. Naïm Frashëri, Tirana

Deklarata e përgjithshme mbi të drejtat e njeriut (la Déclaration universelle des droits de l'homme), ca 1995

Rreth botës për 80 ditë (Le tour du monde en 80 jours) de *Zhyl Vern* (Jules Vernes), 2002, trad Mikail Sterjo, Ed. Redona, Tirana

Autobusi blu (L'autobus bleu), 2003, de Gérard Alba trad. Nasi Lera, Ed. Ora, Tirana

Fjalor i teologjisë biblike, (Vocabulaire de théologie biblique), 2005, de Xavier Léon Dufour, trad. Odette Marquet, Ed. Chiriaco, Naples

Apokalips (Apocalypse), 2006, d'Enver Kushi, Ed. Globus R., Tirana

Camera obscura : Origjina e botës, (La chambre obscure : l'origine du monde) 2007, de Luan Rama Ed. Globus R. , Tirana