Délimitation et étiquetage des morphèmes en coréen par ressources linguistiques

Hyun Gue Huh

Université de Marne-La-Vallée

THÈSE

pour obtenir le grade de

Docteur de l'Université de Marne-La-Vallée

Spécialité : Informatique Linguistique

présentée et soutenue publiquement par
Hyun Gue HUH

le 28 juin 2005

Délimitation et étiquetage des morphèmes en coréen par ressources linguistiques

Resource-based Delimitation and Annotation of Korean morphemes

Directeur de thèse
Éric LAPORTE

Jury : Jee-Sun NAM (rapporteur)
Franz GUENTHNER (rapporteur)
Anne ABEILLÈ
Agata SAVARY
Éric LAPORTE

Ma femme HyeMyoung et ma fille MinJi

Remerciements

Je voudrais remercier tout d'abord à Maurice Gross qui m'a donné la chance de découvrir le domain d l'informatique linguistique.

Je remercie Dominique Perrin, Jee-sun Nam, Franz Guenthener, Anne Abeillé, Agata Savory de 'avoir acceptés de composer le jury de cette thèse.

Je remercie Institut d'électronique et d'informatique Gaspard-Monge de m'avoir accueilli au sein de leurs équipes.

Je remercie l'équipe d'informatique linguistique : Madame Annie Meunier, Madame Nathalie Bely, Monsieur Christian Leclère, Anastasia Yannacopoulou, Claud Martineau, Du-eun Eum, Elsa Sklavounou, Eun-jin Jung, Guénaëlle Mercier, Javier Sastre, Joon-Seo Lim, Krit Kosawat, Matthieu Constant, Olivier Blanc, Rania Voskaki, Stravoriula Voyatzi, Takuya Nakamura, Tita Kyriacopoulou, Mavina Pantazara, Sébastien Paumier, Teresa Gomez-Diaz de m'avoir mis à disposition tous les moyens nécessaires pour accomplir ce travail.

Je remercie Madame Bernadette Matin de me donner son temps pour l'écriture de ma thèse.

Enfin, Je remercie Eric Laporte pour avoir suivi mes travaux et beaucoup aidé durant toutes ces années en France.

Résumé

Nous présentons un système de délimitation morphologique des textes coréens par automates à états finis. Le Coréen est une langue agglutinante et notre système peut probablement être adapté aux autres langues agglutinantes à suffixes (hongrois, finnois, turc).

Les textes coréens s'écrivent principalement avec l'alphabet Hangul qui est un ensemble de caractères syllabiques. Il est possible de les mélanger avec des idéogrammes et des caractères de l'alphabet latin. Nous utilisons le système de codage de caractères UNICODE dans lequel les syllabes coréennes sont rangées par ordre alphabétique. Pour certains traitements sur les syllabes coréennes, nous décomposons chaque syllabe en plusieurs caractères alphabétiques coréens.

Les mots coréens reçoivent des affixes. Pour le nom, un mot peut avoir plusieurs suffixes sans compter les suffixes dérivés, le nombre maximal de combinaisons étant d'environ 1600. Notre première étape pour l'analyse des textes coréens est la description des morphèmes d'un mot pour le segmenter à l'aide des séparateurs : blanc et symboles. Et on segmente encore les segments en morphèmes. Pour pouvoir analyser les segments, nous construisons des dictionnaires de racines et de séquences de suffixes. Nous utilisons les transducteurs pour représenter les compatibilités entre des morphèmes : racines et suffixes avec l'interface graphique de UNITEX. Ils sont conçus de manière à être construits et maintenus manuellement. Notre méthode est fondée sur des ressources linguistiques alors que la plupart des systèmes d'analyse morphologique sont fondés sur des données statistiques.

Nous intégrons automatiquement les dictionnaires de racine et les transducteurs des suffixes en un transducteur unique, qui remplit la fonction d'un dictionnaire. Le résultat de l'analyse d'un texte se présente sous la forme d'un automate pour rendre compte de l'ambiguïté du découpage en morphèmes. Les transitions sont étiquetées par des morphèmes annotés d'informations linguistiques (forme canonique, forme fléchie et informations linguistiques).

Mots-clefs : Texte coréen, transducteur morphologique, transcodage de syllabe, lexique-grammaire.

Abstract

We introduce a system of morphological delimitation and annotation of Korean texts by means of finite state automata. Korean language is an agglutinate language. Our system may fit in other agglutinate languages with suffixes.

Korean texts are written with a set of syllable characters HANGUL (mainly), ideograms and Latin alphabet. We use the encoding system UNICODE where Korean syllables are ranked in alphabetical order. For some applications on the Korean syllables, each syllable is composed of several Korean alphabetic characters.

Korean words contain affixes. A noun can have several suffixes without any derived suffix; the maximum number of combinations is about 1,600. Our first step for the analysis of Korean texts is the description of the morphemes of a word. In order to analyze each word, we construct dictionaries of roots and sequences of suffixes with linguistic resource-based annotation. We describe sets of sequences of suffixes by transducers. Transducers represent compatibilities between morphemes and are constructed with the graphic interface of UNITEX. They are conceived in such a way that they can be manually constructed and maintained by Korean morphology specialists. The other systems of Korean morphology analyzer use rule-based models.

We integrate dictionaries of roots and sequences of suffixes into a unique transducer with linguistic information. The result of the analysis of a text is in the form of an automaton representing the ambiguity of the morphological delimitation. The transitions are tagged with morphemes annotated with linguistic information.

Keyword

Korean text, finite state automata, transducer, syllables, grammar lexicon, resource-based morphological annotation, morphological delimitation.

Table des matières

INTRODUCTION

Notre travail porte sur le traitement automatique de texte coréen, la construction des données linguistiques et la création des outils informatiques pour les traiter. Nous introduisons une méthode de description morphologique pour les langues agglutinantes. Nous avons adapté la méthode déjà élaborée depuis longtemps pour le français par Maurice Gross au LADL[1] avec les outils informatiques INTEX [SIL 1999], UNITEX [PAU 2003] et nous utilisons les ressources linguistiques, créées par les travaux sur le lexique coréen et la description syntaxique des morphèmes.

Le traitement automatique des langues par lexiques et grammaires a été introduit par Maurice Gross [GRO 1984] [GRO 1997], et développé au LADL. L'équipe d'Informatique linguistique de l'IGM[2] en poursuit l'étude. Les grammaires locales sont bien adaptées à la description de la syntaxe des mots [SIL 1997]. Nous l'adaptons à la description des séquences de morphèmes des mots d'une langue agglutinante. UNITEX supporte les Réseaux de Transitions Récursifs (RTN) [WOO 1970] [MOH 1997]. Les RTN permettent la description et la représentation des phénomènes de la langue. Les RTN sont un outil commun entre l'informatique et la linguistique : description et application.

La langue coréenne est agglutinante et la caractéristique des lettres de la structure phonétique est d'être syllabique ; de plus, dans la phrase, l'ordre des syntagmes est plus libre que dans d'autres langues comme les langues indo-européennes.

Pour la description de la langue française, les lexèmes sont recensés sous forme de liste. Dans le texte, chaque lexème est reconnu comme une unité de texte par le dictionnaire de formes fléchies.

Un mot coréen, défini par les séparateurs : espacement et symboles, comporte une unité lexicale et éventuellement un ou plusieurs morphèmes grammaticaux. Par exemple, un nom peut avoir environ 1600 séquences de suffixes nominaux sans compter les suffixes dérivés [BAE 2002]. Dans ce travail, nous traitons les mots dans le texte coréen de la forme suivante :

$$\text{Un mot} = [\text{une racine}] + [\text{suffixe}]^{*[3]} = [\text{morphème}]^{+[4]}$$

[1] Laboratoire d'Automatique Documentaire et Linguistique, PARIS 7, CNRS.
[2] L'Institut d'électronique et d'informatique Gaspard-Monge, Université de Marne-la-Vallée.
[3] Le symbole '*' signifie une répétition zéro, une ou plusieurs fois.

Une racine est un lexème. Certaines racines peuvent apparaître sans suffixe. Un mot est une séquence de morphèmes. À chaque morphème doivent être attachées des informations linguistiques notamment des informations de compatibilité entre morphèmes.

On commence l'analyse automatique de texte coréen par l'étape de l'analyse morphologique. La plupart des systèmes de traitement automatique de textes sont en deux étapes [SPR 1992]. La première étape est la segmentation d'un mot en toutes les séquences possibles de paires (morphème, code grammatical). La deuxième étape sélectionne des séquences correctes à l'aide de données acquises par apprentissage automatique à partir de données fréquentielles issues de corpus étiquetés et d'un modèle statistique comme des modèles de Markov cachés (HMM) ou l'apprentissage par essais et erreurs [CHA 1998] [LEEgb 1997] [LEEgb 2002] [CHOI 1999] [LEE 2000] [KAN 1993]. Une variante récente [HANch 2005] change l'ordre des deux étapes : dans la première on attribue une séquence d'étiquettes à chaque mot puis on fait la segmentation morphologique avec le lexique de morphèmes. Le modèle à deux niveaux a également été utilisé [KIM 1994] [HANch 2002].

Le codage des caractères coréens amène des difficultés supplémentaires. On traite habituellement les caractères syllabiques coréens comme des lettres de l'alphabet latin, ou comme des idéogrammes chinois. Mais un caractère syllabique coréen se compose phonétiquement de lettres de l'alphabet coréen. La plupart des langues n'ont qu'un ou deux codages de leur alphabet, mais le coréen a plusieurs codages de caractères syllabiques et plusieurs codages alphabétiques. Quant on fait la description des morphèmes, si la délimitation de deux par morphèmes coupe dans un caractère syllabique, la description par syllabes ne permet pas de décrire exactement les morphèmes. Nous travaillons au niveau de phonèmes pour qu'il n'y ait aucune restriction dans la description du lexique coréen. Nous permettons que l'utilisateur fasse la description en mélangeant des syllabes et des lettres alphabétiques coréennes. La sauvegarde et l'affichage des textes coréens sont fait en syllabes.

Nous décrivons pour chaque morphème la compatibilité avec les morphèmes voisins. Dans le traitement automatique de texte, un automate peut être utilisé pour la reconnaissance de chaînes de caractères, et un transducteur pour la conversion de chaînes de caractères en d'autres chaînes de caractères. Nous utilisons des transducteurs comme outil de description de

[4] Le symbole '+' signifie une répétition une ou plusieurs fois.

la syntaxe des morphèmes d'un mot coréen. Nous l'utilisons aussi directement pour segmenter des mots en séquence de morphèmes.

L'éditeur graphique d'UNITEX permet de construire et de lire des transducteurs. La description en graphe assure une grande lisibilité et une grande facilité de manipulation sans nécessité de connaître un langage formel de description.

Nous proposons une méthode de description morphologique, syntaxique et sémantique.

Dans la première nous exposons les caractéristiques générales de la langue coréenne, l'écriture, la grammaire et les études précédentes. Dans la deuxième partie nous classons les morphèmes selon la morphologie, la sémantique et la syntaxe dans le mot et la phrase. Dans la troisième partie, plus informatique, nous montrons les traitements des caractères syllabiques, la construction des dictionnaires comprimés de séquences de morphèmes. Dans la quatrième partie, nous présentons la consultation des dictionnaires de séquences de morphèmes dans le traitement automatique du texte coréen avec UNITEX, la construction des automates de phrases, obtenus par la consultation des dictionnaires, qui analysent le texte en morphèmes et informations linguistiques, et l'utilisation des automates de phrases pour l'analyse automatique de texte.

Convention de présentation

Si on transcrit le mot 한글 : [han_gəl] « *hangul* », on a une ambiguïté avec la forme phonétique [haŋ _ə l] qui ne comporte pas les consonnes [n] et [g]. Pour éviter cette situation, nous intercalons entre les caractères le symbole « _ » qui représente la limite de syllabes. Nous représentons les mots coréens transcrits en italiques. Pour les formes canoniques des morphèmes, on utilise les crochets et le gras. Spécialement pour la présentation de la forme canonique des verbes et des adjectifs, nous ajoutons le suffixe infinitif «-다 : -*_da* ».

Tableau 0-1. Transcriptions

L'écriture coréenne	한글은 한국어의 이름이다.
La transcription	*han_geul_eun han_gug_ô_eui i_leum_i_da.*
Formes canoniques	**[han_geul][neun] [han_gug_ô][eui] [i_leum][i_da][da].**

Au début de la chaîne de caractères coréens, le symbole « - » indique le début d'un suffixe pour le distinguer d'une racine qui est en début de mot. Le mot 한글은 est représenté par la séquence de transcription « *han_geul*+-*_eun* » avec le symbole de délimiteur de morphème '+' .

Nous transcrivons selon les tableaux de voyelles et de consonnes dans l'annexe 1 et avec la marque de syllabe.

Abréviations

Catégorie grammaticale

N : nom
V : verbe
A : adjectif
ADV : adverbe
Det : déterminant
Pron : pronom

Les suffixes verbaux et adjectivaux
St Suffixe Terminal
Sd Suffixe déterminatif
Sc Suffixe conjonctif
Sc.cond : Suffixe conjonctif conditionnel
Sc.conj : Suffixe conjonctif de conjonction
St.déc : suffixe terminal déclaratif

Mpp : Morphème de temps au plus-que-parfait ou passé antérieur
Mtc : Morphème de temps au présent
Mtp : Morphème de temps au passé
Mfur : Morphème de temps au futur
Mhon : Morphème honorifique de sujet de la phrase
Sncomp : Suffixe complément

Post : postposition
Post.accu : postposition d'accusatif
Post.nmtp : postposition de nominatif
Post.dat : postposition de datif
Post.gén : postposition de génitif
Suf : Suffixe
Suf.plur : suffixe pluriel
Suf.hon : suffixe honorifique

PREMIERE PARTIE

1. Généralités

1.1 Caractéristiques du coréen

Dans cette partie, nous introduisons les principales informations dont le locuteur qui n'est pas familier du coréen a besoin.

1.1.1 Histoire de l'écriture coréenne

D'après une épitaphe de l'antiquité, on suppose en général que les Coréens ont commencé à écrire environ deux siècles après J.C avec les idéogrammes chinois. Les Coréens anciens utilisaient les idéogrammes chinois de deux façons. La première est l'écriture de la langue chinoise: la phrase coréenne est traduite en phrase chinoise et est écrite en idéogrammes chinois. La deuxième est l'emprunt des idéogrammes pour noter les mots de la langue coréenne par leur forme sonore. On appelle ces caractères 이두 : *i_du*[5]. Ces caractères étaient utilisés dans la société en général avant la dernière dynastie *ChoSun*. A cause de l'invention de la planche de métal gravée en relief[6], on peut imaginer la quantité des publications. Mais la dernière dynastie a interdit de publier des textes en *i_du*. Les sujets de la plupart des livres en *i_du* étaient désapprouvés par cette dynastie. Ces livres furent brûlés. Par la suite, la classe dirigeante n'utilisait l'écriture que de la première façon. Le reste de la société n'avait alors pas de moyen d'écriture.

L'invention du Hangul[7] visait les couches sociales qui n'avaient pas d'instruction. On peut écrire un mot selon sa prononciation syllabe par syllabe. Une syllabe[8] coréenne se compose de lettres alphabétiques coréennes. Les éléments de l'alphabet coréen ont un aspect figuratif du mode d'articulation. On écrit un caractère syllabe en utilisant les consonnes initiales, les voyelles et les consonnes finales. Nous allons expliquer la structure des syllabes coréennes dans la section 1.1.2.

---

[5] www.hangeulmuseum.org

[6] Le livre « *Jig_ji_sim_kyeung* » de 1377 est le plus ancien livre en métal gravé en relief 200 ans avant l'invention de Gutenberg

[7] Nom des caractères syllabiques coréens et de l'alphabet coréen

[8] Dans la suite, nous appelons syllabe la syllabe graphique, qui ne correspond pas toujours exactement à la syllabe phonétique.

Tableau 1-1. Différence de vocabulaire

	Ecriture	Ecriture correspondante coréenne	Transcription dans l'alphabet latin	Sens
Alphabet coréen	하늘	-	_ha_neul	ciel
Idéogramme chinois	天	천	_chôn	ciel

하늘 « ciel » : deux caractères syllabiques avec cinq éléments de l'alphabet coréen

Ciel : quatre caractères avec quatre éléments de l'alphabet latin

天 : un idéogramme chinois

Cette façon d'écrire est très différente de l'écriture latine qui aligne les éléments de l'alphabet. Et aussi, le sens de l'écriture était vertical, de haut en bas, de droite à gauche. De nos jours, on utilise officiellement l'écriture horizontale de gauche à droite.

Le but de l'invention du Hangul était d'enseigner l'écriture aux personnes qui n'avaient pas la possibilité d'une éducation par les idéogrammes. Malgré l'existence du Hangul, la société dirigeante ne l'utilisait pas souvent. Toutes les publications officielles s'écrivaient en texte chinois en idéogrammes chinois. Au début du vingtième siècle, le peuple a pu bénéficier de l'enseignement généralisé en Hangul dans l'école moderne.

Ecriture

La longue période, environ 1 500 ans, de l'utilisation de l'écriture chinoise provoqua l'emprunt d'un riche vocabulaire chinois dans la langue coréenne. Aussi la langue coréenne est-elle formée à la fois de mots sino-coréens et coréens.

On peut écrire tous les mots avec seulement des syllabes coréennes. (1a). Pour rendre clair le sens des mots sino-coréens[9], on les écrivait directement en idéogrammes chinois traditionnels, non simplifiés. (1b; 1d). Ce mode d'écriture exista jusqu'en 1990 dans tous les journaux écrits en idéogrammes et en Hangul. L'écriture verticale était aussi utilisée avant l'usage des publications électroniques. Le gouvernement recommandait d'utiliser des idéogrammes juxtaposés entre parenthèses après le mot sino-coréen écrit avec les syllabes coréennes (1c). Actuellement, dans les journaux, on peut voir les idéogrammes pour les noms

de personnes ou pour les noms de lieux asiatiques. Du fait de la mondialisation culturelle, on utilise aussi des mots étrangers occidentaux. On écrit directement les mots étrangers avec leurs propres caractères (2).

(1)

a. 서울은 한국의 수도이다.
sô_ul_neun han_gug_eui su_do_i_da.
b. 서울은 韓國의 首都이다.
sô_ul_neun han_gug_eui su_do_i_da.
c.서울은 한국(韓國)의 수도(首都)이다.
sô_ul_neun han_gug(han_gug)_eui su_do(su_do)_i_da.
d.서울은 韓國(한국)의 首都(수도)이다.
sô_ul_neun han_gug(han_gug)_eui su_do(su_do)_i_da.
Séoul est la capitale de la Corée du Sud.

(2)

a. 파리는 프랑스의 수도이다.
pa_li_neun feu_lang_seu_eui su_do_i_da.
b. paris 는 france 의 수도이다.
pa_li_neun feu_lang_seu_eui su_do_i_da.
Paris est la capitale de la France.

Vocabulaire

Avant le vingtième siècle la société était hiérarchisée politiquement. L'élite était instruite avec des textes chinois en idéogrammes, utilisait les mots sino-coréens dans les conversations, mais le reste de la société ne connaissait que les mots purement coréens. Chaque camp gardait son propre vocabulaire. Selon les linguistes, le coréen a deux vocabulaires.

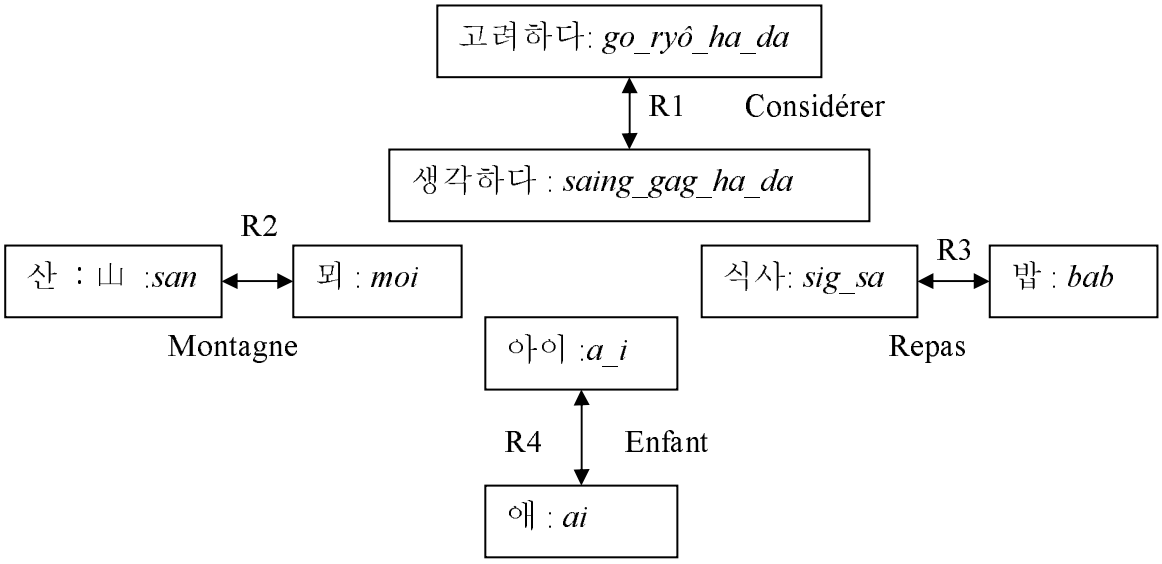
Le coréen possède aussi un important lexique de formes honorifiques. Il existe des mots pour exprimer son respect envers une personne âgée, à une personne qui a une situation sociale élevée ou envers des auditeurs. Le mot 짐 : *jim* « moi » est un pronom à la première personne. Ce mot n'est utilisé que par le roi. Aujourd'hui, on s'en sert pour les phrases métaphoriques ou dans les pièces de théâtre ancien. Parmi les suffixes, il existe les suffixes qui expriment le respect.

[9] Plusieurs idéogrammes chinois distincts, avec des sens distincts, peuvent avoir la même prononciation en

Tableau 1-2 Les variantes pour la représentation de l'honorification envers l'interlocuteur

	très honorifique	honorifique	banal	signification
Nom	수라 : *su_la*	식사 : *sik_sa*	밥 : *bab*	Le repas, de riz
Pronom	전하 : *jôn_ha* 짐 :*jim*	당신:*dang_sin* 저 : *jô*	너 : *nô* 나 : *na*	Toi Moi
Suffixe	-께옵서 : -*_gge_ob_sô*	-께서 : -*_gge_sô*	-가/이 :-*_ga/_i*	Le suffixe nominatif du sujet
Verbe	?	드시다 : *deu_si_da*	먹다 : *môg_da*	Manger
Suffixe	-옵니다 : -*_ob_ni_da.*	-십니다 :-*_sib_ni_da*	-다 :-*da*	Le suffixe verbal et adjectival terminal déclaratif

Nous pouvons considérer quatre relations entre termes synonymes : les synonymes purs (R1), les synonymes avec différence d'origine : purement coréenne ou sino-coréenne (R2), avec différence de niveau de respect (R3), la variation phonétique (R4).



Nous représentons la relation R4 entre forme canonique et variante, la relation R3 par étiquette de niveau de langue sur les suffixes, et nous étudierons plus tard les autres relations.

coréen.

Traitement des caractères en informatique

Les syllabes qui constituent un mot sont des éléments de l'alphabet coréen. Chaque caractère représente phonétiquement une syllabe au contraire d'un idéogramme qui révèle le sens. Outre les éléments anciens de l'alphabet coréen, il existe 11 172 caractères syllabiques coréens en informatique, grâce à la combinaison de trois parties : 21 consonnes initiales, 19 voyelles, 28 consonnes finales. De plus, on utilise les idéogrammes chinois. D'après le dictionnaire des idéogrammes sino-coréens [DONG 2002], il existe environ cinquante mille idéogrammes utilisés en coréen. La différence des idéogrammes entre chinois et coréens réside dans la différence des polices de caractères. Les idéogrammes coréens correspondent à la forme des idéogrammes chinois anciens. Il existe aussi des idéogrammes qui ont été créés pour le coréen. En informatique, avec le système de codage WANSUNG, on utilise 4888 idéogrammes. Dans le système de codage de UNICODE, il en existe environ quarante mille.

Dans l'écriture coréenne, il n'est pas possible d'utiliser seulement une consonne ou une voyelle pour présenter un mot ou une syllabe contrairement à la préposition française « à » qui se compose d'une seule voyelle. Dans le texte coréen, des lettres de l'alphabet coréen qui n'existent pas dans les syllabes et sont isolées, sont citées en tant que symboles. Si elles sont employées dans un mot, ce mot est un nom propre. Il est aussi possible que le mot soit écrit avec des lettres de l'alphabet coréen pour décomposer les morphèmes : 간다 *gan_da* «aller+temps présent »➔ 가-ㄴ다 :*ga-n_da*.

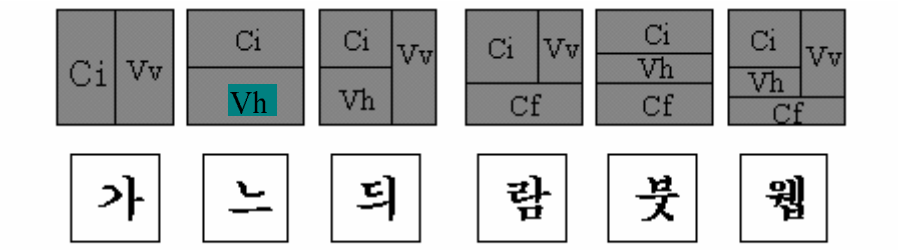
En informatique, les syllabes coréennes et les idéogrammes sont traités comme des lettres. Pour les langues écrites avec l'alphabet latin, dans le texte, on peut utiliser les codes identifiés pour les caractères de l'alphabet latin, les symboles, les chiffres. Il est possible de numériser tous les caractères dans moins d'une centaine de codes. Pour le coréen, on a besoin d'au moins soixante mille codes pour les syllabes, les idéogrammes, les lettres de l'alphabet coréen, les symboles de l'écriture coréenne.

Il existe trois systèmes basiques de codage : WANSUNG, JOHAB, UNICODE désignés comme standard par le gouvernement. Nous les expliquerons en détail dans la troisième partie.

1.1.2 Caractères coréens

Les caractères coréens sont alphabétiques et syllabiques. Un caractère coréen dans le texte correspond à une syllabe qui se compose de consonnes et de voyelles. La figure (1-1) montre les six configurations graphiques d'une syllabe coréenne selon la forme graphique des voyelles et l'existence d'une consonne finale. Nous appelons la consonne finale non pas la dernière consonne d'un mot mais la dernière consonne d'une syllabe. La consonne initiale se situe à gauche avec les voyelles verticales : ㅏ, ㅓ, ㅣ, en haut avec les voyelles horizontales : ㅗ, ㅜ, ㅡ, à gauche en haut avec les voyelles composées d'une voyelle verticale et d'une voyelle horizontale : ㅘ, ㅝ, ㅚ. La consonne finale se situe sous les voyelles. Le sens d'écriture d'une syllabe est de gauche à droite et de haut en bas. Phonétiquement, une syllabe coréenne se présente sous l'une des quatre manières : Voyelles, Consonnes + Voyelles, Voyelles + Consonnes, Consonnes + Voyelles + Consonnes.

Le livre *Hun_min_jung_eum hai_lai_bon*[10] explique le principe de la création de l'alphabet coréen et la construction des syllabes coréennes. Les voyelles de l'alphabet coréen commencent par les trois éléments fondamentaux «•», «ㅡ», « ㅣ ».



Ci : Consonnes initiales, Cf : Consonnes finales, Vh : Voyelles horizontales Vv : voyelles verticales

Figure 1-1 Configuration d'une syllabe coréenne

Le tiret vertical et horizontal signifie la figure de la langue dans la bouche. Le tiret vertical est la forme de la langue debout. Le tiret horizontal est la forme de la langue à plat. Le point donne la position de l'articulation et la résonance dans la bouche. Le point devient le tiret court à côté du tiret vertical : ㅓ, ㅏ ou en haut et en bas du tiret horizontal: ㅗ, ㅜ.

Le tableau de l'annexe A montre les voyelles coréennes modernes. Les 7 voyelles ; ㅏ : ɑ , ㅓ : ʌ , ㅗ : *o*, ㅜ : *u*, ㅡ : ɨ , ㅣ : i et • : *a* sont les voyelles de base. La voyelle «•» qui s'appelle 아래 아 : *_a_lai a* « a inférieure » a été utilisée avant le vingtième siècle. Elle a été remplacée par la voyelle ㅏ. La figure de la voyelle ㅏ se compose de deux éléments : un tiret long vertical ' ㅣ ' et un tiret court horizontal «-». Elle représente la position de l'articulation derrière la langue debout au fond de la bouche. Le long tiret représente la forme de la langue et le tiret court représente la position de l'articulation. Le tiret long horizontal représente la langue plate. La voyelle ㅏ identifie la prononciation de la voyelle 'a', la voyelle ㅣ identifie la voyelle [i]. La voyelle 'ㅡ :*eu*' est similaire à la voyelle du mot « le ».

Les voyelles ㅑ : *ya*, ㅕ : *yô*, ㅛ : *yo*, ㅠ: *yu* sont dérivées des voyelles de base ㅏ, ㅓ, ㅗ et ㅜ en ajoutant le tiret court qui a la valeur de la semi-voyelle 'y'. Par exemple, le mot 가시었다 : *ga_si_ôss_da* « être allé (honorifique) » a la variante 가셨다 : *ga_syôss_da*. Les deux voyelles '*i*' et '*ô*' forment une voyelle composée '*yô*' ; « ㅣ [i] + ㅓ [ô] = ㅕ [*yô*] ». La voyelle 'i' devient la semi-voyelle 'y' avant les voyelles ' ㅓ ' et aussi les autres voyelles a, o et u. Les autres voyelles composées sont des combinaisons des voyelles de base. Le mot 아이 : *_a_i* «enfant» a la variante « 애 : ɛ » à cause de la contraction des voyelles des syllabes : « ㅏ[a] + ㅣ[i] = ㅐ[ɛ] » qu'on peut précisément rencontrer dans les transcriptions de l'alphabet latin sous la forme « ai » prononcée [ɛ] comme en français.

Les signes utilisés pour les consonnes figurent la position de l'articulation. Les consonnes basiques sont ㄱ :*g*, ㄴ:*n*, ㅁ :*m*, ㅅ :*s* et ㅇ : ɦ [11]. La consonne ㄱ:*g* est dite vélaire. Sa forme figure le dessus de la langue touchant l'antérieur du palais. La forme de la consonne ㄴ désigne le bout de la langue touchant les dents. La consonne ㅁ :*m* représente la bouche fermée. L'idéogramme chinois « 口 » figure la bouche. La consonne 'ㅅ' figure la langue touchant les dents, la consonne 'ㅇ' désigne la glotte en position consonantique. La consonne ㄹ : *l* est différente des autres ; elle ne figure pas la forme du lieu de l'articulation.

---

[10]훈민정음 : 訓民正音 : *hun_min_jung_eum* « renseigner peuple exact son », désigne HANGUL en sino-coréen, 해례본 : 解禮本 : *hai_lai_bon* « expliquer exemple référence ».

[11] Une consonne finale.

Les autres consonnes sont dérivées des consonnes de base. Le tableau (1-3) montre les symboles des consonnes dérivées. Les consonnes de la deuxième colonne sont graphiquement proches des consonnes de base. Les consonnes doubles ㄲ, ㄸ, ㅆ, ㅃ et ㅉ renforcent, avec la tension, la prononciation des consonnes fondamentales ㄱ, ㄷ, ㅂ et ㅈ. La position de l'articulation de ces consonnes est proche de celle de la consonne fondamentale. La troisième colonne montre les consonnes aspirées. Les consonnes ㅋ, ㅌ, ㅍ et ㅊ sont dérivées, avec pour signification que la source de l'articulation [^{h}]est la forme aspirée des consonnes ㄱ, ㄷ, ㅂ et ㅈ.

Tableau 1-3 Etapes de la dérivation de la figure des consonnes en consonnes fondamentales

Consonnes fondamentales		fort	Avec la consonante aspirée
ㄱ[g,kˀ ,k]		ㄲ[kˀ]	ㅋ[k^{h}]
ㄴ[n,ɲ]			
	ㄷ[dˀ ,tˀ ,t]	ㄸ[tˀ]	ㅌ[t^{h}]
ㄹ [r, l]			
ㅁ [m]			
	ㅂ[b]	ㅃ[pˀ]	ㅍ[p^{h}]
ㅅ[s,ʃ]		ㅆ[ʃˀ]	
	ㅈ[ʧ]	ㅉ[ʧ^{h}]	ㅊ[ʧ^{h}]
ㅇ[ŋ]	ㆁ[12] [ʔ]		ㅎ[ɦ]

Outre la différence d'écriture entre les consonnes fondamentales et les consonnes aspirées, on ajoute un tiret court au-dessus du signe de la consonne fondamentale. La consonne aspirée ㅎ: h est semblable à la consonne anglaise 'h'. La consonne finale ㅎ est muette comme la consonne française 'h'.

Le mot 자모 : JAMO[13] qui est un nom de l'alphabet coréen signifie « consonne-voyelle ». D'après *Hun_min_jung_eum*, une consonne ne peut pas constituer un son sans voyelle. Par contre, une voyelle peut être prononcée seule. Mais quand on écrit une syllabe

[12] Cette consonne n'est plus utilisée.

formée d'une voyelle seule, on la fait précéder de la consonne initiale 'ㅇ' pour conserver le schéma minimal consonne+voyelle. Par exemple, pour la voyelle 'ㅏ', on écrit une syllabe 아, la consonne initiale 'ㅇ' ne note aucun son; par contre, la consonne finale 'ㅇ' a pour valeur le son nasal comme '*ng*' en français après les voyelles. Par exemple, la syllabe 앙 : *ang* se prononce [aŋ].

Phonétiquement, chaque syllabe dans une séquence de caractères est prononcée à son tour. Mais on enchaîne la consonne finale à la syllabe suivante entre les syllabes d'un mot lorsque la syllabe suivante commence par la consonne initiale 'ㅇ'. Par exemple le mot 작업 *jag_ôb* (Travail) se prononce [자겁 **:ja_gôb**](avec enchaînement) et aussi [**jag_ôb**] (sans enchaînement). La valeur de son de la consonne finale ㄱ :*g* est enchaînée à la syllabe suivante qui n'a phonétiquement pas de valeur de la consonne initiale 'ㅇ' et cette consonne finale devient la consonne initiale de la syllabe suivante.

Le tableau de l'annexe B montre les consonnes modernes dans la situation des syllabes isolées. Les valeurs des consonnes initiales sont les valeurs identiques. Mais les sons des consonnes finales ont des valeurs différentes selon les consonnes finales. Quand on prononce une syllabe isolée qui contient une consonne finale, sans enchaînement, cette consonne finale de cette syllabe se prononce selon une des sept prononciations [ㄱ :**g**, ㄴ :**n**, ㄹ :**l**, ㅁ :**m**, ㅅ :**s**, ㅇ :**ng**, ㅂ :**b**].

Les consonnes composées comportent deux lettres. Sans enchaînement, le son des consonnes finales composées est une de deux soit la première soit la deuxième (닮다 :*dalm_da* [**dam_da**] « être semblable »). Avec l'enchaînement, une des deux consonnes finales est enchaînée dans certaines conditions à la consonne initiale d'une syllabe suivante et l'autre reste (닮았다 :*dalm_ass_da* [**dal_mad_da**] « a été semblable »).

Le tableau de l'annexe C montre l'enchaînement des consonnes finales et l'influence mutuelle entre consonne finale et consonne initiale. La consonne finale 'ㅎ : *h*' est muette. Mais quand elle précède les consonnes initiales ㄱ :*g*, ㄷ :*d*, ㅂ :*b* et ㅈ :*j* de la syllabe

---

[13] 자모 : **JAMO** est un sigle constitué des premières syllabes des mots : 자음(子音 : fils-son) *JA_EUM* « consonne », 모음(母音 : mère-son) *MO_EUM* « voyelle ».

suivante, celles-ci deviennent ㅋ :*k*, ㅌ :*t*, ㅍ :*p* et ㅊ :*ch*. (놓다 : *noh_da* [**no_ta**] «mettre, poser »)Au contraire, quand la valeur des consonnes finales d'une syllabe avant la syllabe qui commence par la consonne initiale ㅎ est celle des ㄱ :*g*, ㄷ :*d*, ㅂ :*b* et ㅈ :*j*, la consonne initiale ㅎ :*h* devient ㅋ :*k*, ㅌ :*t*, ㅍ :*p* et ㅊ :*ch* (땋다 : *dda_ta* [**dda_ta**] «tresser »).

En coréen, il n'y a pas de liaison entre les mots comme dans la langue française.

Nous expliquons les systèmes de codage de caractères coréens dans la section 3.1.

1.1.3 Phrase et proposition

La structure basique de la phrase coréenne est dans l'ordre, sujet, compléments, verbe ou sujet, compléments, adjectif (prédicatif). En coréen, les adjectifs se conjuguent comme les verbes et aussi le suffixe adjectival *-i_da* « être ».

La langue coréenne est une langue à cas. Les cas sont marqués par des postpositions qui sont des suffixes nominaux. Les substantifs avec leurs suffixes se placent librement avant les divers verbes et adjectifs. L'exemple (1) montre les phrases terminées par un verbe et un adjectif. L'exemple (1b) montre un cas avec un suffixe adjectival –이다 : *-_i_da* « être » qui est soudé avec les substantifs. Ce suffixe se conjugue aussi comme les verbes et les adjectifs.

(1)
a. 막스는 방금 집으로 갔다.
mag_seu_neun bang_geum jib_eu_lo gass_da
[mag_seu][neun] [bang_geum] [jib][eu_lo] [ga_da][ôss][da].
Max+Post.nmtp à l'instant maison+Post. « vers » aller+Mtp+St.déc.
Max vient à l'instant de partir vers chez lui.

b. 그는 학생이다.
geu_neun hag_saing_i_da.
[geu][neun] [hag_saing][i_da][da].
Lui.PRO:3p+Post.nmtp étudiant+être+St.déc.
Il est étudiant.

c. 그 꽃은 아름답다.
geu ggoch_eun a_leum_dab_da.
[geu] [ggoch][neun] [a_leum_dab_da][da].
Ce.Dét fleur+Post.nmtp « être beau »+St.déc
Cette fleur est belle.

Les exemples (2a1), (2a2), (2a3) et (2a4) montrent que l'ordre des éléments dans la phrase est plus libre qu'en français.

(2)
a1. 레아는 연필을 막스에게 주었다.
Le_a_neun yôn_pil_eul Max_e_ge ju_ôss_da.
[le_a][neun] [yôn_pil][leul] [Max][e_ge] [ju_da][ôss][da].
Léa+Post.nmtp stylo+Post.accu Max+Post.datif donner+Mtp+St.déc.
Léa a donné un stylo à Max.
=a2 연필을 레아는 막스에게 주었다.

stylo+Post.accu Léa+Post.nmtp Max+Post.datif donner+Mtp+St.déc.
=a3 연필을 막스에게 레아는 주었다.
stylo+Post.accu Max+Post.datif Léa+Post.nmtp donner+Mtp+St.déc.
=a4 막스에게 레아는 연필을 주었다.
Léa+Post.nmtp Max+Post.datif stylo+Post.accu donner+Mtp+St.déc.

Mais les phrases suivantes (3a) (3b1) montrent que l'ordre des mots est important. L'interversion du sujet et du complément en présence du verbe 되다 : *doi_da* « devenir » change le sens de la phrase. Dans cette phrase, le sujet et le complément possèdent la même postposition nominative. Dans la phrase (3b2) l'ordre est libre en raison de la présence de la postposition -으로 : -*_eu_lo* « vers, de ».

(3)
a. 얼음이 물이 되었다.
ôl_eum_i mul_i doi_oss_da.
[ôl_eum][i] [mul][i] [doi_da][oss][da]
la glace+Post.nmtp l'eau+Post.accu devenir+Mtp+St.déc.
La glace devint de l'eau.
b1. 물이 얼음이 되었다.
mul_i ôl_eum_i doi_oss_da.
[mul][i] [ôl_eum][i] [doi_da][oss][da].
eau+Post.nmtp la glace+Post.accu devenir+Mtp+St.déc.
L'eau devint de la glace.
물은 얼음으로 되었다.
mul_eun ôl_eum_eu_lo doi_oss_da.
[mul][neun] [ôl_eum][eu_lo] [doi_da][oss][da]. eau+Post.spc glace+Post.« en » devenir+Mtp+St.déc.L'eau devint de la glace.
=얼음으로 물은 되었다.

Pour la phrase simple coréenne qui se termine par une racine verbale ou adjectivale, il y a un suffixe terminal verbal ou adjectival à la fin de la phrase. On appelle phrase simple une phrase qui comporte une seule proposition, et phrase complexe une phrase qui en comporte plusieurs.

Pour construire une phrase complexe à partir de phrases simples et d'une conjonction, on remplace la conjonction et le suffixe terminal de la première phrase par un suffixe conjonctif (Sc). Les phrases (4a), (4b) montrent la phrase obtenue en remplaçant la conjonction et le suffixe terminal par le suffixe conjonctif : -으나 : -*_eu_na*.

(4)

a. 그는 정시에 도착했다. 그러나, 기차는 안 도착했다.

geu_neun jông_si_e do_chag_haiss_da. geu_lô_na, gi_cha_neun an do_chag_haiss_da.

[geu][neun] [jông_si][e] [do_chag_ha_da][iss][da].[geu_lô_na], [gi_cha][neun] [an] [do_chag_ha_da][ôss][da].

Lui+Post.nmtp heure+Post.lieu arriver+Mtp+St.déc, mais train+Post.nmtp « ne pas » arriver+Mtp+St.déc

Il est arrivé à l'heure. Mais le train n'est pas arrivé.

=b. 그는 정시에 도착했으나 기차는 안 도착했다.

*geu_neun jông_si_e do_chag_haiss_**eu_na**, gi_cha_neun an do_chag_haiss_da.*

[geu][neun] [jông_si][e] [do_chag_ha_da][ôss][eu_na], [gi_cha][neun] [an] [do_chag_ha_da][ôss][da].

Lui+Post.nmtp heure+Post.lieu arriver+Mtp+Sc train+Post.nmtp « ne pas » arriver+Mtp+St.déc

Il est arrivé à l'heure. Mais le train n'est pas arrivé.

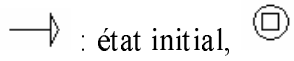
1.2 Automates finis

Un automate fini est une machine abstraite, utilisée dans l'étude du calcul et dans diverses applications telles que le traitement des langues, et qui possède une quantité finie et constante de mémoire (les états). Une telle machine peut être conceptualisée comme un graphe orienté. Il y a un nombre fini d'états, et chaque état a des transitions vers zéro, un ou plusieurs états. La ou les transitions qui peuvent être suivies sont déterminées par la valeur donnée en entrée. Les automates finis ou machines à états finis sont étudiés dans la théorie des automates, un sous-domaine de l'informatique théorique.

On définit un automate fini A par la donnée d'un quintuplet d'ensembles (B, Q, I, T, F) où

B est un alphabet fini
Q est un ensemble fini d'états
I est une partie de Q appelée ensembles d'états initiaux
T est une partie de Q appelée ensembles d'états terminaux
F est un ensemble fini de triplets (p, a, q), où p et q sont des états et a un symbole de l'alphabet B ; ces triplets sont les transitions de l'automate. 

En traitement des langues naturelles, les automates finis sont souvent représentés sous une forme dite « graphe » par Max Silberztein [SIL 1993]. Sous cette forme, les états sont représentés par les symboles graphiques :

 : état initial, : état terminal.

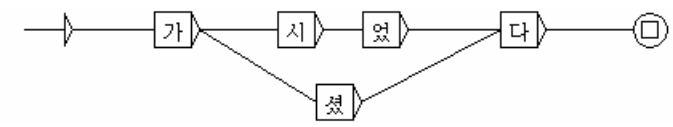


Figure 1-2 Automate fini (forme "graphe")

Dans le domaine de la théorie des langages formels, un automate fini représente un langage reconnaissable. Le graphe de la figure (1-2) reconnaît les chaînes de caractères :

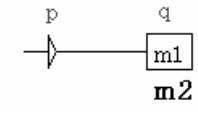
가시었다: *ga_si_ôss_da*, 가셨다 :*ga_syôss_da* . En traitement des langues naturelles, les automates finis ont été utilisés pour rechercher les lexiques [REV 1991].

Les automates finis montrent bien aussi les structures syntaxiques élémentaires ; ils sont alors appelés grammaires locales.

Les transducteurs sont des automates avec des caractères de sortie sur les transitions.

Un transducteur est un 6-uplet T = (Q, B1, B2, I, T, F)

Q est un ensemble fini d'états
I est une partie de Q appelée ensembles d'états initiaux
T est une partie de Q appelée ensembles d'états terminaux
B1 est un alphabet fini de symboles d'entrée
B2 est un alphabet fini de symboles de sortie
F est un ensemble de quadruplets (p, m1, m2, q) où p et q sont des états. m1 est un mot sur l'alphabet B1 et m2 un mot sur l'alphabet B2, qui définissent une transition de l'état p à l'état q.



Dans la forme « graphe », la transition est exprimée par un trait qui part à droite de l'état p et qui se termine à gauche de l'état q de gauche à droite.

Les transducteurs ont été utilisés pour reconnaître un mot sur le premier alphabet B1 et émettre un mot sur le deuxième alphabet B2 comme le dictionnaire DELAP[14] [LAP 1988] ou le transducteur de flexion d'INTEX [SIL 1999].

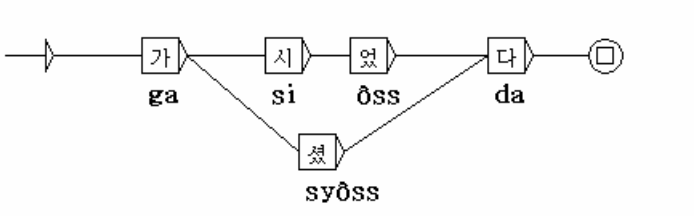


Figure 1-3 Transducteur de transcription

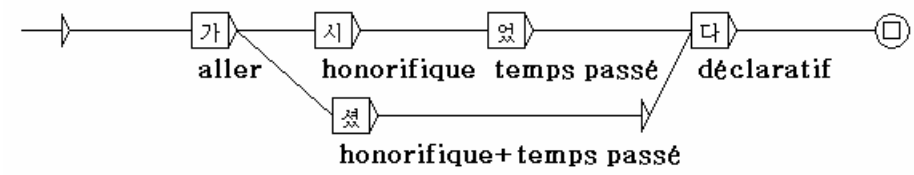


Figure 1-4 Transducteur d'étiquetage

Les transducteurs ci-dessus reconnaissent les mêmes chaînes de caractères que la figure (1-2), le transducteur figure (1-3) a pour sorties les séquences : *gasiôssda* et *gasyôssda* et le transducteur figure (1-4) a pour sorties les séquences des informations « aller honorifique temps passé déclaratif » et « aller honorifique+temps passé déclaratif ».

Dans le domaine du traitement automatique du texte, on utilise les automates pour la reconnaissance des mots. On utilise les transducteurs pour reconnaître des mots et pour produire des informations sur les mots.

Nous allons montrer dans la troisième partie une méthode de description des séquences des suffixes d'un mot coréen sous forme de transducteur. Nous allons expliquer la description des séquences de morphèmes d'un mot sous la forme d'un transducteur dans la section 3.2.

[14] Dictionnaire Electronique du LADL pour la Phonémique

1.3 Etudes précédentes sur le traitement automatique des textes coréens

Un mot coréen est une séquence d'éléments morphologiques ou morphèmes. Les linguistes étudient chaque morphème. Mais à cause de la complexité de la combinaison des morphèmes, il n'existe pas de description formelle complète de toutes les combinaisons des séquences de suffixes. Dans les dictionnaires éditoriaux coréens, les entrées décrivent les racines et les suffixes. Elles donnent la forme canonique avec très peu d'informations sur la compatibilité avec les morphèmes voisins. Pour le traitement automatique du coréen, aucun système n'utilise une description complète des propriétés des morphèmes des mots : la plupart des systèmes utilisent des ressources linguistiques en deux parties : les racines et les suffixes. Des informations sur la compatibilité entre la racine et la séquence de suffixes sont incluses dans chacune de ces deux parties et des règles de compatibilité des racines et des suffixes sont placées dans un autre fichier ou dans le programme. La partie des suffixes consiste en un dictionnaire de suffixes simples ou de séquences de suffixes.

1.3.1 Systèmes avec les suffixes simples

On décrit les entrées de suffixes avec la forme canonique et on a besoin de règles pour obtenir les formes fléchies de chaque entrée.

Pour reconnaître les morphèmes d'un mot avec ces dictionnaires électroniques, il existe en gros les deux étapes suivantes, la première étape est d'abord la délimitation des racines à l'aide des dictionnaires de racines, on applique le dictionnaire des suffixes simples à plusieurs reprises jusqu'à la fin du mot. Ce traitement peut produire des séquences incorrectes. La deuxième étape est la sélection des séquences de morphèmes en appliquant des méthodes statistiques [BRILL 1995] [LEEgb 1997] [LEEgb 2002] [HANch 2005] [SIN 1995] sur les morphèmes obtenus et leurs étiquettes grammaticales. Ces opérations sont appelées l'analyse morphologique [HONG 1996] [KANG 1993]. Quand on ajoute de nouvelles entrées, on doit ajouter des informations linguistiques dans les dictionnaires et aussi des règles de compatibilité entre racines et suffixes.

Dans la figure (1-5), le KOMA (Korean Morphological Analyzer)[15] montre un exemple de système de traitement automatique de texte coréen avec analyse morphologique. Le système garde séparés le dictionnaire et les connexions entre morphèmes [LEEec 1993].

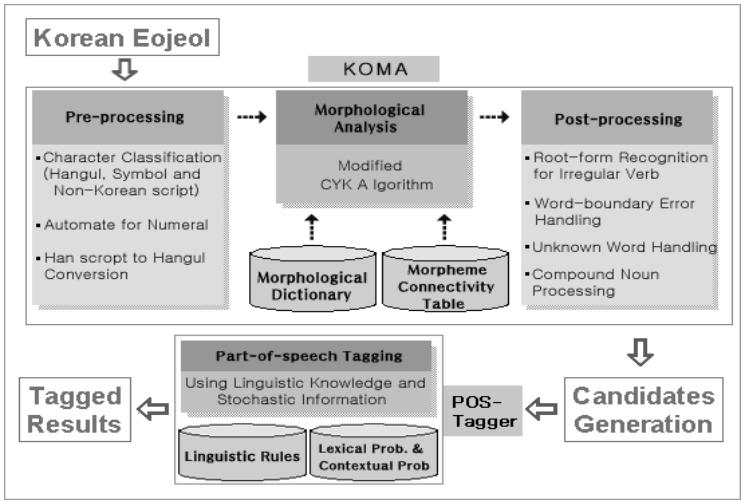


Figure 1-5 KOMA

L'autre système de l'analyse morphologique HAM[16] (Hangul Analysis Module) comporte les résultats de l'analyse morphologique comme le tableau (1-4) [KANG 1993]. Chaque mot est montré par la séquence de paires de forme canonique ou fléchies[17] et de codes grammaticaux[18].

Tableau 1-4 Résultat d'analyse morphologique de HAM

Les mots coréens	Les résultats avec la racine
도우면 :*dob_u_myon*	(V "돕")<Tb:24> + (e "면")<13>
돕게 : *dob_ge*	(V "돕")<Tb:20> + (e "게")
도와 : *do_oa_da*	(V "돕")<Tb:24> + (e "어")<6>

Certains suffixes peuvent avoir différentes interprétations selon la position : le suffixe 어 :-*_ô* peut avoir les propriétés : le conjonctif, l'impératif ; l'interrogatif. Pour l'analyse

[15] KLE(Knowledge & Language Engineering Lab.), POHANG University of SCIENCE AND TECHNOLOGY.
[16] http://nlp.kookmin.ac.kr/HAM/kor/ham-intr.html
[17] La racine 돕_*dob* « aider » a les variantes 도우 :*do_u*, 돕 :*dob*, 도오 :*do_o*, le suffixe conjonctif conditionnel 면 *_myon* est la variante de 으면 : *_eu_myon*. Le suffixe adverbial 게 :*_ge* et le suffixe déclaratif 다 :*_da* n'ont pas de variante.
[18] POS (Part-Of-Speech) V : verbe, e : suffixe final verbal et adjectival, avec les codes de variation de forme : <Tb :24>,<Tb : 20>, <13> et <6>.

syntaxique, nous avons de plus besoin des informations grammaticales et syntaxiques. [VOU 1995] [LEEsz 2000].

1.3.2 Systèmes avec séquences de suffixes et étiquetage de la séquence

Il existe deux méthodes de description de séquences des suffixes : on étiquette des séquences entières :« 아씨다 \PA1 .TmDec .Conj» [19] [NAM 1996] ou on étiquette chaque suffixe par sa forme canonique et son code : « 돕 : *_dob*/VV+었 : *_ôss*/EPF+다 : *_da*/EFN » [20] (Na-Rae HAN[21]).

Dans cette section nous étudions les systèmes où on étiquette les séquences complètes. Dans les travaux de Jee-Sun NAM [NAM 1996], les dictionnaires de suffixes contiennent toutes les séquences des suffixes verbaux et adjectivaux sous forme fléchie avec des étiquettes de compatibilité phonétique, mais ne contiennent pas les formes canoniques. Ils contiennent certaines séquences qui commencent par la partie variante de la racine pour les racines variables. Chaque séquence est étiquetée sans délimiteur entre les suffixes. Elle est seulement étiquetée selon le dernier élément de la séquence de suffixes.

Avec ces dictionnaires, CHOI [CHOI 1999] et LEE [LEE 1997] utilisaient le système de codage de la première version WANSUNG qui contient un ensemble de syllabes sélectionnées à partir d'un ensemble de textes coréens (section 3.1).

Les programmes de Chang-Yeol LEE décomposent les mots en racines, suffixes et préfixes dans un mot. Les résultats de son analyse morphologique sont les séquences possibles de morphèmes en tenant compte seulement de la compatibilité phonétique. Parmi les séquences, il existe des séquences incorrectes. Par exemple, le mot 말하다 : *mal_ha_da* « dire, parler, constater » est analysé 말 : *mal*.N + 하다 :*ha_da*.adj.faire+다 : *da*.St et 말 : *mal*.N + 하다 :*ha_da*.V.faire + 다 : *da*.St. Le mot *mal_ha_da* est un verbe mais

---

[19] La séquence 아씨: *ass* représente une syllabe incomplète sans consonne initiale. C'est une variante du suffixe de temps passé -었 : -*_ôss* après les racines adjectivales qui se terminent par les voyelles 아 **[a]**, 오 **[o]** et 와 ***[oa]***. Le suffixe -다: -*_da* peut être soit un suffixe terminal déclaratif (TmDec) ou soit une variante du suffixe conjonctif -다가 : -*_da_ga* (Conj).

[20] La syllabe 돕 : *_dob* est la forme de base de 돕다 : *_dob_da* « aider ». cette séquence contient aussi le suffixe de temps passé -었 : -*_ôss* et le suffixe terminal déclaratif -다.

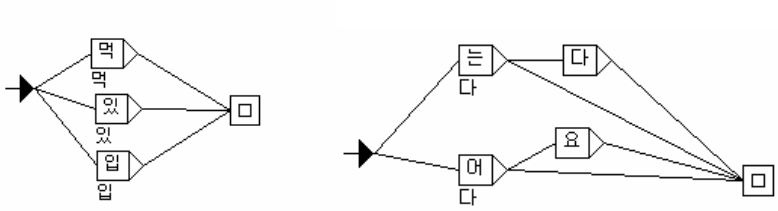
[21] http://www.ldc.upenn.edu/Catalog/docs/LDC2004T03/readme.txt

le résultat montre aussi le suffixe adjectival *ha_da* parce que le dictionnaire utilisé ne représente pas complètement les compatibilités entre éléments morphologiques.

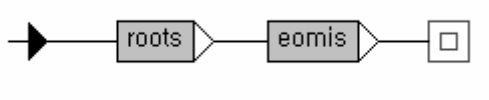
Sung-Woo CHOI a construit un dictionnaire comprimé de racines de mots dont les éléments de base sont les éléments de la liste de syllabes coréennes. Le format de ce dictionnaire est le même que celui utilisé par INTEX pour les langues écrites avec l'alphabet latin. La liste des syllabes remplace la liste des lettres de l'alphabet latin. Il construit plusieurs dictionnaires selon les catégories grammaticales. Il a utilisé une matrice qui représente les combinaisons entre racines et séquences de suffixes. Pour pouvoir traiter de nouvelles combinaisons, on doit ajouter des entrées dans les dictionnaires et changer la matrice incluse dans le programme. La mise à jour n'est pas facile parce que les ressources linguistiques ne sont pas lisibles ou comportent une redondance importante.

Le programme de DECOTEX [BER 2004] avec le dictionnaire de Jee-Sun NAM donne la description de la compatibilité entre les racines et les séquences de suffixes sous forme de transducteur. Par exemple ci-dessous, l'expression de la compatibilité entre quelques racines verbales terminées par les consonnes (roots.grfs) et quelques séquences de suffixes soudés à ces racines (emis.grfs) est représentée par une série de transducteurs de racines et de suffixes (figure 1-6c). Les sorties dessous la boite expriment les informations linguistiques de chaque morphèmes. Nous donnons le tableau à gauche pour les éléments analysés avec leurs codes.

Description
먹 :_*mog* « manger »
있 :_*iss* « exister »
입 :_*ib* « habiller »
는 :,_*neun* « Sd »
는다 :_*neun_da* « Mtc+St »
다 :_*da* « St »
어 :_*ô* « St »
어요 :_*ô_yo* « St+Mhon »



a) Roots.grfs b) Eomis.grfs



c) graphe de liaison entre racines et suffixes

Figure 1-6 Description de deux parties : racines et suffixes

DECOTEX n'utilise les caractères coréens que sous la forme de syllabes. Dans le cas des mots qui contiennent les syllabes où le début de la syllabe peut faire partie d'un morphème, et la fin de la syllabe faire partie d'un autre. On doit ajouter ces mots au dictionnaire pour les entrées qui contiennent une délimitation dans la syllabe comme « 게,것.N :이/JO » : le mot 게 : *_gôi* qui est une contraction de 것이 : *_gôs_i* se compose du nom incomplet 것 : *gôs* (ce que) et de la postposition de nominatif 이 :*_i*.

1.3.3 Systèmes avec les séquences de suffixes et étiquetage de chaque suffixe

Klex (Finite-state Lexical Transducer for Korean)[22] par Na-Rae HAN représente les séquences des morphèmes avec XFST[23] [KAR 1992] [HANch 2002]. Elle a utilisé les séquences de morphèmes qui sont construites par l'université de YONSEI. XFST décrit la séquence des morphèmes d'un mot à l'aide d'expressions régulières avec des paires (entrée/sortie) qui correspondent à deux niveaux : description lexicale et forme fléchie [KAR 1996]. L'entrée correspond à la forme canonique et aux informations linguistiques. La sortie est la forme fléchie. XFST produit un mot fléchi tel que les mots d'un texte.

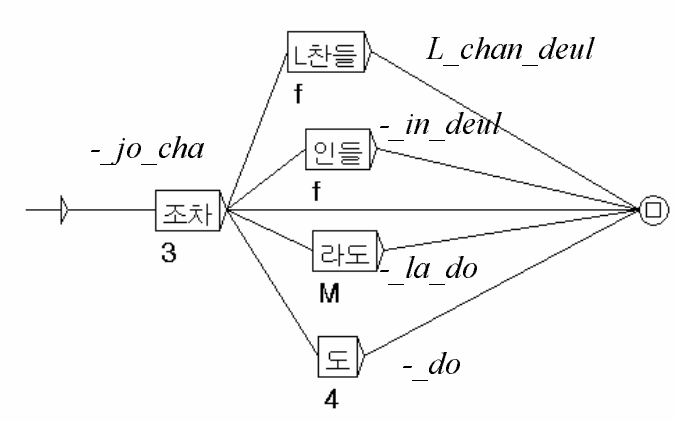
(1)

fly/VV+s/ECS	돕/VV+었/EPF+다/EFN	Niveau de la description lexicale
|	|	
flies	도왔다	Niveau des formes fléchies

Klex utilise ce transducteur en sens inverse, à partir d'un mot sous forme fléchie, il obtient la séquence des informations sur les morphèmes. L'exemple (1) montre les cas du mot anglais *flies* et du mot coréen 도왔다 : *_do_oass_da* « avoir aidé » avec la séquence 돕다 :*_dob_da* « aider » + -었 : -*_ôss* « temps passé » + -다 :-*_da* « Suffixe terminal » sous forme canonique.

[22] http://www.cis.uppen.edu/~nrh/klex.html
[23] Xerox Finite State Transducer

Sun-Mee BAE a construit avec le système INTEX [SIL 1999] les séquences de postpositions, qui sont des suffixes nominaux, pour reconnaître les noms composés. Pour cela, elle utilise des transducteurs qui décrivent les séquences de suffixes sous forme fléchie [BAE 2002]. Cette méthode marche bien pour les séquences de postpositions. Les formes canoniques sont représentées par l'intermédiaire de codes et non incluses de façon lisible dans les graphes. Par exemple, la figure (1-7) décrit les séquences qui commencent par une postposition –조차 :-*_jo_cha* « même, non plus » [BAE 2002]. La séquence de postpositions -조차인들 : -*jo_cha_in_deul*, peut être analysée -조차:-*_jo_cha* + -이다. -*_i_da* « être »+ -은들 : -*_eun_deul* « encore que » (formes canonique). Le suffixe conjonctif -은들 : -*_eun_deul* « encore que » a une variante -ㄴ들 : -*n_deul* après syllabe ouverte. Le suffixe adjectival –*i_da* « être » a une variante facultative « <E> » après les morphèmes terminés par syllabes ouvertes. Alors la séquence -조차인들 : -*jo_cha_in_deul* possède la variante libre -조찬들 : -*jo_chan_deul*. Pour décrire -조찬들 : -*jo_chan_deul*, BAE utilise la séquence « 조차 + L 찬들 ». Le symbole « L » indique qu'il faut reculer d'une syllabe et ajouter la chaîne de caractères qui suit [SIL 1999]. Dans cet exemple, la forme canonique -*i_da* « être » n'est pas représentée dans la figure. Ces transducteurs permettent de reconnaître des noms composés mais n'ont pas été intégrés à un dictionnaire comprimé. Le système INTEX utilise aussi les syllabes sur la description. Les informations linguistiques dans le tableau ci-dessous sont présentées de façon séparée.



Codes de sortie	Signification : Forme canonique, Informations linguistiques
f	*ndeul*, Aux
M	*lado*, Aux
3	*jocha*, Aux
4	*do*, Aux

Figure 1-7 Séquences de postpositions

1.3.4 Principaux points de notre méthode

Notre méthode d'étiquetage morphologique diffère des méthodes précédentes en plusieurs points : l'inclusion des informations de compatibilité entre morphèmes dans le dictionnaire, l'utilisation des lettres en plus de syllabes, le recours aux graphes pour visualiser les transducteurs.

Nous considérons un mot dans le texte comme une séquence de morphèmes : une racine et des suffixes. Chaque morphème est représenté par une entrée et des informations linguistiques, notamment sur la compatibilité entre les morphèmes. La totalité de la description de la compatibilité des morphèmes est incluse dans le dictionnaire, on n'a pas besoin d'effectuer des opérations d'analyse morphologique au niveau des morphèmes d'un mot à l'aide de règles de levée d'ambiguïtés.

Nous proposons un mode de représentation de la compatibilité entre les morphèmes qui facilite la construction et la maintenance manuelles du lexique. Les informations de compatibilité sont associées aux autres informations linguistiques sur le morphème. En cas d'erreur dans les résultats de l'analyse par dictionnaire, on peut corriger l'entrée en modifiant ou en construisant la description de la compatibilité du morphème dans une ressource linguistique qui décrit les séquences de morphèmes.

Notre méthode de représentation autorise un mélange entre les lettres de l'alphabet coréen et les syllabes coréennes. Alors que les lettres de l'alphabet coréen qui apparaissent dans le texte coréen sont considérées comme des citations de symboles, les lettres de l'alphabet coréen dans la description sont interprétées normalement.

La figure (1-8) montre notre méthode de description de la séquences « 조차인들 : *-jo_cha_in_deul* » (Comparer avec fig. 1-7). Les formes fléchies sont écrites en caractères syllabiques, et en lettre de l'alphabet coréen lorsque c'est nécessaire (exemple : *n_deul* dans cette figure). Les étiquettes fournies en sortie comprennent les formes canoniques (exemple : *_eun_deul* dans cette exemple) et des informations grammaticales (exemple : Post, Sc).

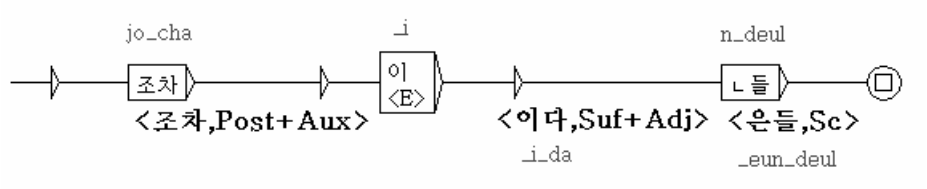


Figure 1-8 Séquences 조차인들 : *-jo_cha_in_deul.*

Nous utilisons les transducteurs comme outil pour la description des séquences de morphèmes (figure 1-9). Chaque chemin d'un transducteur décrit une séquence de morphèmes autorisés. Dans ce transducteur, une transition signifie la compatibilité entre les morphèmes et une sortie représente les informations linguistiques sur l'entrée, qui est représentée sous la même forme que dans un texte. Les informations linguistiques contiennent la forme canonique. Pour distinguer la forme de la sortie par rapport aux autres formes de transducteur d'UNITEX, nous mettons les informations linguistiques entre les symboles « <> ».

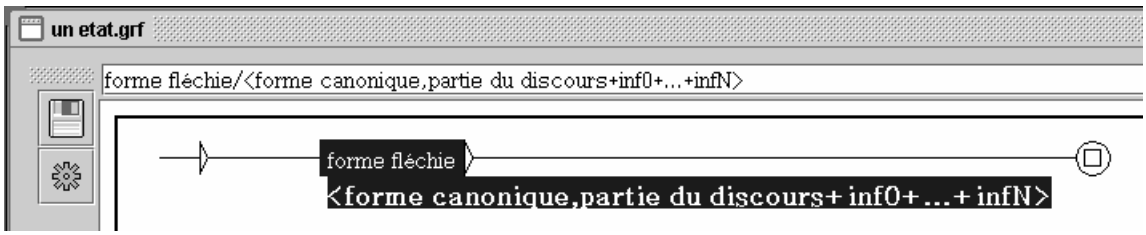


Figure 1-9 Un état d'un transducteur

La figure (1-10) montre un transducteur décrivant les mots avec la racine 돕다 :*dob_da* « aider » et les suffixes -으면 : *-_eu_myôn* « conditionnel », -었 : *-_ôss* « passé » et -게 : *-_ge* « adverbial ou impératif ». Dans ce graphe, la racine du verbe apparaît sous plusieurs formes différentes : 도우 *do_u*, 도오 : *_do_o* et 돕 : *_dob*.

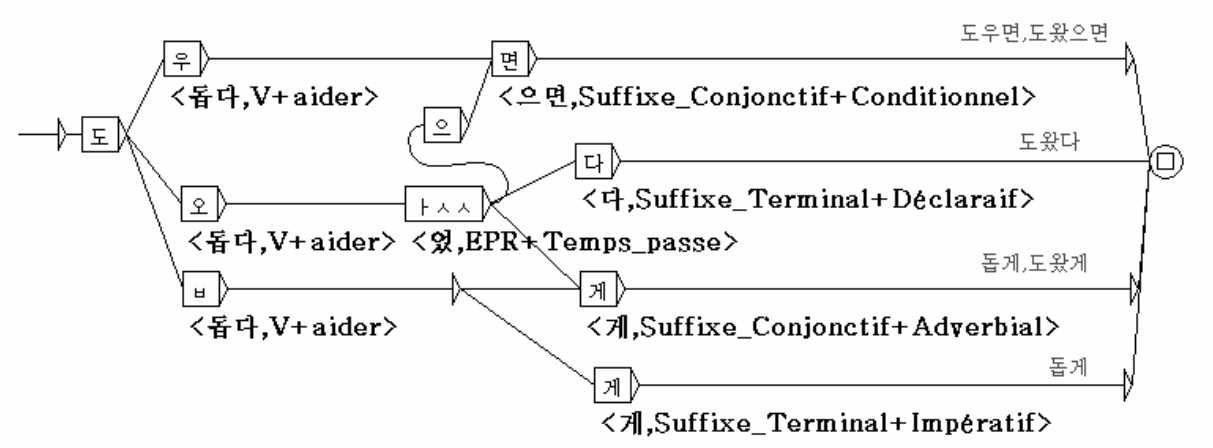


Figure 1-10 Description de séquences de morphèmes avec un transducteur

Notre méthode de description est équivalente aux expressions régulières de KLEX mais la représentation par graphe rend la description plus lisible.

Nous avons adapté au coréen le format informatique d'une entrée du DELAF, [*Forme fléchie*, *forme canonique, informations linguistiques*], en ajoutant le champ pour la compatibilité, qui indique le nom du transducteur qui décrit sous forme de graphes l'ensemble des séquences de suffixes pour l'entrée décrite. Nous exposons la construction du dictionnaire dans les sections 3.2 et 3.3.

DEUXIEME PARTIE

2. Classification des racines et des suffixes

Dans cette section, nous allons classifier les racines et les suffixes qui composent un mot. Nous allons commencer par la définition d'un mot comme une séquence de morphèmes. Dans la séquence de morphèmes d'un mot, chaque morphème est ordonné et il a la fonction sémantique ou syntaxique ou sémantique et syntaxique. Pour décrire les relations entre les morphèmes d'un mot, les morphèmes doivent être ordonnés et posséder des informations linguistiques. Nous divisons les racines en 3 groupes selon leur fonction et le type de suffixes qu'ils peuvent prendre.

Avec suffixes verbaux et adjectivaux : verbe, adjectif.
Avec postpositions : nom, pronom, adverbe.
Sans suffixe : déterminant, épithète, mot invariable.

Nous utilisons le mot « postposition » pour les suffixes nominaux.

Nous travaillons sur la base des dictionnaires réalisés sous la direction de Jee-Sun NAM à l'IGM: les adjectifs, les verbes, les noms simples [NAM 1994], [NAM 1997]. Notre classification des morphèmes se fonde sur le livre *Gug_ô_Mun_Bob* [SUH 1996] et celle des suffixes sur le dictionnaire *Dai Sa Jeon* . Pour les variantes des racines verbales et adjectivales, nous avons aussi consulté le dictionnaire *Keun Sa Jeun*.

2.1 Classification des racines verbales et adjectivales

Dans cette section, nous allons classifier les racines verbales et adjectivales et les suffixes qui se soudent à elles. Les mots composés de racines verbales ou adjectivales et de suffixes verbaux et adjectivaux se situent à la fin de la phrase ou de la proposition. Pour décrire les relations entre les morphèmes : ordre et forme des morphèmes, nous distinguons la compatibilité morphologique qui regroupe les contraintes qui portent sur les formes canoniques des éléments morphologiques et la compatibilité phonétique qui regroupe les contraintes qui mettent en jeu les formes effectives de ces éléments.

Dans l'exemple (1), selon les racines, le suffixe terminal –어 : -_*ô* peut être un suffixe impératif ou un suffixe exclamatif (1a, 1b) et peut aussi avoir la variante phonétique 아 :-_*a* (1a, 1b).

(1)
a. 잡아 !
jab_a !
[jab_da][ô] !
rattraper+St.impératif
Rattrape le !
b. 높아 !
nop_a !
[nop_da][ô] !
« être haut(e) »+St.exclamatif »
Que c'est haut !
Quelle hauteur !
c. 열어 !
yôl_ô !
[yôl_da][ô] !
ouvrir+ St.impératif !
Ouvre !

2.1.1 Classification morphologique

Nous commençons par classifier les racines en quatre catégories de compatibilité morphologique selon les séquences de suffixes verbaux et adjectivaux qui peuvent se combiner avec elles : les verbes, les adjectifs (cas général), les adjectifs en *iss_da* « exister, être » et leurs variantes, le suffixe -*_i_da* « être » qui se soude aux racines nominales et les adjectifs en -*_i_da*.

Tableau 2-1 Classification des verbes et des adjectifs

Catégorie		Exemples
V	Tous les verbes	날다 : *nal_da* « voler ».
A	Adjectifs sauf ISS, IDA	이쁘다 : *i_bbeu_da* « être beau »
ISS	Adjectif 있다 : *iss_da* « exister, avoir, être » et mots dérivés par l'ajout du suffixe -*iss_da* 계시다 : *gyôi_si_da* est honorifique de *iss_da* 없다 : *ôbs_da* « ne pas exister » et mots dérivés par l'ajout du suffixe -ôbs_da	맛있다 : *mas_iss_da* « être délicieux » 맛없다 : *mas_ôbs_da* « ne pas avoir bon goût, être fade »
IDA	Adjectifs dérivés par l'ajout du suffixe -이다 :-*_i_da* « être » 아니다 :*a_ni_da* « ne pas être » et adjectifs dérivés par l'ajout du suffixe -*_a_ni_da*	*geu_gôs_eun X_**i_da***. C'est X. *geu_gôs_**eun** X_**ga** a_ni_da.* Ce n'est pas X.

En général, il existe des critères importants de classification des racines entre le verbe et l'adjectif selon la compatibilité de certains suffixes. Le premier est la compatibilité avec les suffixes impératifs. L'adjectif n'a pas de suffixes impératifs. Nous allons appliquer aux

racines adjectivales utilisées pour les souhaits ou les vœux[24], des suffixes identiques aux formes du suffixe impératif. Dans les exemples (2), on applique le suffixe –어 : -_*ô* aux racines adjectivales. Selon les racines adjectivales, le suffixe prend une valeur exclamative (2a, 2b), ou propositive sur la racine *hang_bog_ha_da* (2b).

(2)
a. 아름다워!
a_leu_da_uô !
[a_leu_dab_da][ô] !
« être beau »+St.exclamatif
Quelle beauté!
b. 행복해 !
haing_bog_hai !
[haing_bog_ha_da][ô] !
« être heureux »+St.exclamatif ou être heureux+St.propositif
Quel bonheur ! ou je te souhaite du bonheur

Le deuxième critère est la compatibilité avec le suffixe présent –는/ㄴ : -_*neun*/-*n*. Ce suffixe se soude aux racines verbales (3a).

(3)
a. 그 나라는 발전한다.
geu na_la_neun bal_jôn_han_da.
[geu] [na_la][neun] [bal_jôn_ha_da][neun][da].
Ce pays+Post.nmtp développer+Mtc+St.déc.
Ce pays se développe ou Ce pays est en train de se développer.

b. *그는 행복한다.
**geu_neun haing_bog_han_da.*
[geu][neun] [haing_bog_ha_da][neun][da].
*Lui+Post.nmtp « être bonheur »+Mtc+St.

Grammaticalement, *iss_da* est intermédiaire entre un verbe et un adjectif. Le mot *iss_da* peut être complété par des séquences de suffixes verbaux. Il peut avoir le suffixe impératif (4b) et présent (4c). Nous avons fait la distinction entre la racine *iss_da* et les autres racines adjectivales.

(4)
a. 그는 집에 있다.
geu_neun jib_e iss_da.

[24] Les racines adjectivales dérivées des sino-coréennes avec le suffixe –*ha_da* : 건강 : *gôn_gang* (santé), 행복 *haing_bog* (bonheur) 편안 :*pyông_an* (l'aise), 장수 : *jang_su* (vie longue).

[geu][neun] [jib][e] [iss_da][da].
Lui+Post.adjectif maison+Post.lieu être+St.déc.
Il est chez lui.

b. 집에 있어!
jib_e iss_ô!
[jib][e] [iss_da][ô].
Maison+Post.lieu exister+St.imp.
Reste à la maison !

c. 그는 집에 계속 있는다.
geu_neun jib_e iss_neun_da.
[geu][neun] [jib][e] [iss_da][neun][da].
Lui+Post.adjectif maison+Post.lieu être+Mtc+St.déc.
Il est chez lui ou Il est en train d'être chez lui.

Le suffixe -*_i_da* a est compatible avec certains suffixes qui ne s'unissent pas aux autres catégories. Par exemple, le suffixe verbal et adjectival 구나 : -*_gu_na* a une fonction qui représente la pensée du locuteur. Le suffixe 로 . -*_lo* dans la phrase (5b) est utilisé par les personnes âgées pour s'exclamer et par politesse. Le suffixe -*_lo* existe seulement dans la séquence des suffixes qui suivent le suffixe -*_i_da*.

Nous distinguons aussi le suffixe -*i_da* et les autres racines adjectivales.

(5)

a.그는 집에 있겠구나.
geu_neun jib_e iss_gess_gu_na.
[geu][neun] [jib][e] [iss_da][gess][gu_na].
lui+Post.adjectif maison+Post.lieu exister+ Mfut+St.déc.
(je pense que) il est à la maison.

b. 그것은 꽃이로구나.
geu_gôs_eun ggoch_i_lo_gu_na.
[geu_gôs][neun] [ggoch][i_da][lo][gu_na].
Ce.Pron+Post.adjectif fleur+être+ Masp+St.déc.
(Je pense que) c'est la fleur.

Pour la négation, on utilise l'adverbe 아니 :*_a_ni* « ne pas » avant le verbe ou l'adjectif (6b1) (6b2). La variante de *_a_ni* est 안 :*_an*. Le verbe *_ iss_da* et le suffixe adjectival -*_i_da* ont leur propre mot négatif : *_ obs_da* « ne pas exister » (6b3), *_a_ni_da* « ne pas être » (6b4).

(6)

a) N0+post.nmtp V0+St.

1) 그는 집에 간다.
geu_neun jib_e gan_da.
[geu][neun] [jib][e] [ga_da][neun][da].
lui+Post.nmtp maison+Post.lieu aller+Mtp+St.
Il va à la maison.

2) 저 꽃은 아름답겠다.
Jô ggoch_eun a_leum_dab_gess_da.
[jô] [ggoch][neun] [a_leum_dab_da][gess][da].
Ce fleur+Post.nmtp « être beau »+Mfut+St.déc.
Cette fleur-là sera belle.

3) 그는 집에 있다.
geu_neun jib_e iss_da.
[geu][neun] [jib][e] [iss_da][da]
lui+Post.nmtp maison+Post.lieu être+St.déc.
Il est à la maison.

4) 십 년 전 그는 학생이었다.
sib nyôn jôn geu_neun hag_saing_i_ôss_ôss_da.
[sib] [nyôn] [jôn] [geu][neun] [hag_saing][i_da][ôss_ôss][da]
Dix ans avant.N lui+Post.nmtp étudiant+être+Mtpp+St.
Il y a dix ans, il était étudiant.

b) N0+post.nmtp ***a_ni/an*** V0+St.

1) 그는 집에 안 간다.
geu_neun jib_e an gan_da.
Il ne va pas à la maison.

2) 저꽃은 안 아름답겠다.
jô ggoch_eun an a_leum_dab_gess_da.
Cette fleur-là ne sera pas belle.

3) 그는 집에 없다.
geu_neun jib_e ôbs_da.
Il n'est pas à la maison.

3.1) *그는 집에 안 있다.

**geu_neun jib_e an iss_da.*

4) 십 년 전 그는 학생이 아니었다.
sib nyôn jôn geu_neun hag_saing_eu a_ni_ôss_da.
Il y a dix ans, il n'était pas étudiant.

2.1.2 Classification phonétique

Les racines des verbes et des adjectifs sont citées dans les dictionnaires conventionnels sous une forme qui se termine par le suffixe –다 :-*_da*. Nous reprenons cette tradition en utilisant des formes canoniques en -*_da*. Mais les formes effectives des racines auxquelles se soudent les suffixes sont des formes sans -*_da*.

La limite entre deux morphèmes ne coïncide pas toujours avec une limite de syllabe au sens de l'écriture coréenne. Quand on découpe le mot entre racine et séquence de suffixes, les parties découpées sont différentes selon le jeu de caractères coréens utilisés. Les exemples suivants montrent les racines et leurs variantes avec le même suffixe impératif 어라 : -*_ô_la*. Le mot 잡아라 : *jab_a_la* est analysé en une racine verbale invariable *_jab_da* et un suffixe -*_a_la* variante de –*_ô_la* lorsqu'on utilise les syllabes ou l'alphabet coréen (7a). Le mot 해라 : *hai_la* est représenté par la séquence d'une racine 해 : *hai* variante de *ha_da* et d'un suffixe -*_la* variante de -*_ô_la* lorsqu'on utilise les syllabes, mais par la séquence d'une racine 하 : *_ha* de *ha_da* et un suffixe -*i_la* variante de -*_ô_la*(7b) avec l'alphabet coréen. Dans le cas du mot 주워라 : *ju_uô_la*, le découpage en syllabes ne permet pas de découper ce mot en racine et suffixe(7c).

(7) Delimitation des mots coréens

	Mots	Transcription	Découpage en syllabes	Découpage en lettres d'alphabet	Formes canoniques	Sens des racines
a)	잡아라	*jab_a_la*	**jab | _a_la**	**jab | _a_la**	**[jab_da][ô_la]**	« Saisir »
b)	해라	*hai_la*	**hai | _la**	**ha | i_la**	**[ha_da][ô_la]**	« Faire »
c)	주워라	*ju_uô_la*	**ju_uô | _la ou ju | _uô_la**	**ju_u | ô_la**	**[jub_da][ô_la]**	« Ramasser »

Nous analysons les mots au niveau de l'alphabet coréen pour découper précisément. Dans le cas (7c) sur le mot *ju_uô_la* nous le découpons en la racine *ju_u* variante de *jub_da* et le suffixe - ㅝ 라 :–*ô_la* variante de 어라 : -*_ô_la*.

Selon la variation de la forme canonique avec les suffixes suivants, nous classifions les racines verbales et adjectivales en trois catégories : invariable, variable régulière, variable irrégulière.

Nous proposons des codes structurés en trois parties :

RAC_VC_JY

La première partie du code (RAC) indique les variantes de la forme de la racine. La deuxième (VC) indique si la racine est ouverte phonétiquement. En coréen, les voyelles sont distribuées en 2 classes Yin et Yang selon la caractéristique de la voyelle de la dernière syllabe. La troisième (JY) indique l'accord de la caractéristique de voyelle entre celle de racine et celle de suffixe (en détail voir le tableau 2-13).

Tableau 2-2 Classification des voyelles en Yin et Yang

Caractéristique	Les voyelles de la dernière syllabe
Yin	ㅓ *ô*, ㅕ : *yô* , ㅜ : *u*, ㅠ :*yu*, ㅔ : *ôi*, ㅝ : *uô*, ㅟ :*ui*, ㅖ :*yôi*, ㅞ :*uôi* :, ㅇㅓ : *_ô* , ㅣ : *i*, ㅡㅣ:*eui*
Yang	ㅏ :*a*, ㅑ :*ya*,ㅗ :*o*, ㅛ :*yo*, ㅐ :*ai*, ㅘ :*oa*, ㅚ :*oi*, ㅒ :*yai*, ㅙ :*oai*

On met les séquences de syllabes avec les trois parties : Consonne initiale, voyelle et consonne finale. Une racine verbale ou adjectivale soit $C_i0Y0C_f0\ldots.C_in\text{-}1Yn\text{-}1C_fn\text{-}1C_inYnC_fn$ sans suffixe ㄷㅏ :*_da*[25].

On peut définir la caractéristique de racine verbale et adjectivale comme suit ;

Si *Yn* est Yin.
    La caractéristique de racine est Yin
Si *Yn* est Yang.
    La caractéristique de racine est Yang
Si *Yn* est '— : *eu*'.
    Si n== 0 # la racine n'a qu'une syllabe.
        la caractéristique de racine est Yin
    Si n != 0 # la racine a plusieurs syllabes.
        Yn-1 est Yin
            La caractéristique de racine est Yin
        Autrement
            La caractéristique de racine est Yang

[25] La forme canonique de la racine verbale ou adjectivale soit $C_i0Y0C_f0\ldots.C_in\text{-}1Yn\text{-}1C_fn\text{-}1C_inYnC_fnC_in{+}1Yn{+}1C_fn{+}1$ ($C_in{+}1$ = ㄷ $Yn{+}1$ = ㅏ $C_fn{+}1$ = ø)

Racines invariables

Ce sont les racines qui ont une seule forme. Pour ce type, nous étiquetons 'SANS' à la place de « RAC ». Le dernier élément de racine conditionne la forme du suffixe suivant. Nous distinguons les racines invariables selon l'existence ou non d'une consonne finale. Nous étiquetons 'C' pour l'existence d'une consonne finale, sinon nous mettons 'V'. De plus, pour les suffixes qui commencent par la voyelle ㅓ : *ô*, lorsque les racines sont terminées par certaines voyelles, il y a contraction de deux syllabes en une. L'exemple (8) montre la combinaison de racine et suffixe et la variation de racines et de suffixes. La première étape est en forme canonique et la deuxième montre les formes fléchies.

(8)

a. 가 +-어 ➔ 가

ga.aller +-_***ô***.St.imp ➔ *ga !*

Vas-y !

b.세 +어 ➔세어 | 세

sôi. « être fort » +-_***ô***.St.excl ➔ *sôi_ô | sôi !*

Que c'est fort !

c. 오 +-어 ➔와 !

_o.venir +-_*ô*.St.imp ➔ *_oa !*

Viens !

Dans le cas (8a) la racine ga a la caractéristique de Yang, le suffixe '-_*ô*' en forme canonique devient la variante '-_*a*' et la syllabe de voyelle '-_*a*' est disparue par l'élisions de voyelle répétée.

Tableau 2-3 Classification des racines par variation phonétique

CODE	Dernier élément	Exemples
SANS_C_[J/Y]	Consonne	넘다 : *nôm_da*(franchir, gravir et sauter l'obstacle)
SANS_V_[J/Y]	Voyelle	뛰다 : *ddui_da*(sauter)
SANS_V_E	Une voyelle ô' ou 'a'	가다 *: ga_da*(aller)

SANS_V_EJ	Une voyelle composée avec les voyelles ô' ou 'a'	세다 : *se_da*(être fort)
SANS_V_C[J/Y]	Une voyelle 'o' ou 'u'	오다 : *o_da*(venir)

Dans ces cas, nous considérons toutes les variations sur la partie de suffixes, la partie de racine n'est pas changée.

Racines variables régulières

Toutes les racines qui se terminent par les consonnes 'l' et 'h' ont des variantes sans ces consonnes lorsqu'elle sont suivies de certains éléments initiaux des morphèmes suivants.

Tableau 2-4 Classification des racines variables régulières

CODE	Dernier élément	Conditions	Action	Exemples
LXX	ㄹ : l	Avant les voyelles et consonnes 's', 'n'	disparu	난: *nan* [날다 : *nal_da*][*eun*] voler+Sd.pass qui a volé
HXX	ㅎ : h	Avant la voyelle du suffixe Toutes les racines adjectivales *	disparu	파란 : *pa_lan* [파랗다 : pa_lah_da][*eun*] « être bleu »+Sd bleu

*. Il existe une racine exceptionnelle de HXX : le mot '좋다 : *joh_da* (être bon).

Racines variables irrégulières

Les racines qui se terminent par certains éléments peuvent avoir les variantes avec certains éléments initiaux de morphèmes suivants. Le tableau (2-5) montre les racines qui se terminent par la consonne 'd'. la racine 믿다 : *mid_da* se termine par la consonne 'd' mais elle n'a pas de variante contrairement à la racine 묻다 : *mud_da* (poser une question). l'autre homonyme 묻다 : *mud_da* (enterrer) n'a pas de variante.

Tableau 2-5 Variantes des racines terminées par la consonne ‘ㄷ : d'

		sens	Suffixe		code
			으면 : *eu_myôn*	었다 : *ôss_da*	
			Sc.cond	Mtp+St.déc	
Racine	믿다 *mid_da*	croire	믿으면 : mid_eu_myôn	믿었다 : *mid_ôss_da*	SANVARC
	묻다 *mud_da*	enterrer	묻으면 : *mud_eu_miôn*	묻었다 : *mud_ôss_da*	SANVARC
	묻다 *mud_da*	demander, poser une question	물으면 : *mul_î_myôn*	물었다 : *mul_ôss_da*	DACVARC

Les types des racines variables irrégulières qui se terminent par certains éléments sont dans le tableau ci-dessous.

Tableau 2-6 Classification des racines variables irrégulières

CODE	Dernier élément	Condition	Modification des derniers éléments	Exemples
BCU	ㅂ :b	Devant tous sauf les consonnes ‘g’, ‘d’, ‘n’	Remplacé par une syllabe ‘_u’	추운 :chu_un [춥다 : chub_da][n] «avoir froid »+Sd froid
LEU	르 : *leu*	Devant les voyelles ‘ô’ et ‘a’	Remplacé par « leu_l »	푸르러 갔다:*pu_leu_lô gass_da* [푸르다 :pu_leu_da][_ô] [ga_da][ôss][da] « être bleu »+Sc aller+Mtp+St a rendu de plus en plus bleu
LLE	르 : *leu*	Devant les voyelles ‘ô’ et ‘a’	Remplacé par « l_l »	흘러 : 갔다 :*heul_lô gass_da*. [흐르다 :heu_leu_da][_ô] [ga_da][ôss][da] « courir, couler »+Sc aller+Mtp+St a passé en courant
VVX	一 : *eu*	Devant les voyelles ‘ô’ et ‘a’	Eliminé -*eu*	들러 갔다 : *deul_lô gass_da* [들르다 :deul_leu_da][ô] [ga_da][ôss][da] « rendre visite »+Sc aller+Mtp+St a rendu visite et est parti
DVC	ㄷ: d	Devant les voyelles ‘ô’ et ‘a’ et les consonnes m, b	Remplacé par « l »	물었다 : *mul_ôss_da*. [묻다 :mud_da][ôss][da] poser la question+Mtp+St a posé la question.

SXV	ㅅ: s,	Devant les voyelles 'ô' et 'a' et les consonnes m, b	Eliminé -*s*	부었다 :*bu_ôss_da* [붓다 :bus_da][ôss][da] verser des liquides+Mtp+St a versé
HADA	하 : ha	Devant les consonnes 'ㄱ :*g*', 'ㄷ :*d*', 'ㅈ :*j*'	Eliminé une voyelle '*a*'	발전타 : *bal_jon_ta*. [발전하다 :bal_jon_ha_da][da] développer

Chaque ligne du tableau (2-6) s'applique à un groupe de racines, mais il existe aussi quelques racines variables irrégulières qui ont une correspondance spécifique (tableau 2-7)

Tableau 2-7 Racines irrégulières exceptionnelles

racine	Sens	Avec le suffixe *eu_myon*	Avce les suffixes *ôss_da*	code
푸다 *pu_da*	puiser	푸면 : *pu_myôn*	푸었다 : *pu_ôss_da* 폈다 *poss_ da*	PUDVAR
놓다 *noh_da*	mettre, laisser, poser, placer	놓으면 : *noh_eu_myôn*	놓았다 : *noh_ass_da* 놨다 : *noass_da*	NOHVAR

Nous mettons les codes à chaque racine de forme canonique, dans le tableau (2-15) de la section 2.2.2, les trois colonnes à gauche montrent les codes de racine verbale, le tableau (2-16) pour les racines adjectivales.

Nous obtenons 18 classes de racines verbales et 18 classes de racines adjectivales, 1 pour la racine *iss_da* et 2 pour le suffixe *–i_da* selon la dernière syllabe des morphèmes précédents : syllabe ouverte ou fermée.

Nous expliquons la méthode de la description de la compatibilité entre racines et séquences de suffixes et la construction du dictionnaire électronique dans la troisième partie.

2.2 Classification des suffixes verbaux et adjectivaux

Les suffixes verbaux et adjectivaux ont les fonctions indiquant le mode, un lien entre un mot et une proposition, un lien entre les propositions ou entre les mots, le temps et le niveau de langue. Nous classifions ces suffixes selon la caractéristique syntaxique, sémantique et morphologique.

2.2.1 Classification syntaxique et sémantique

Dans la séquence des suffixes verbaux et adjectivaux, chaque suffixe a une fonction par rapport à la syntaxe de la phrase. Nous utilisons cette fonction pour classer les suffixes.

La séquence des suffixes verbaux et adjectivaux après la racine est :

(Honorification de sujet) + (temps et aspect) + (modalité liés au locuteur) + suffixe final.

Nous donnons les codes St, Sc, Sd, Scomp aux suffixes finaux selon leurs fonctions dans la phrase.

Nous donnons le code « Morph » aux autres suffixes.

2.2.1.1 Honorification

Pour exprimer l'honorification, on applique certains morphèmes.

Le premier cas est les exemples (9) et montre les combinaisons sur le verbe *ga_da*(aller). La phrase (9a) est appliquée par le locuteur qui est plus âgé que l'auditeur et que le sujet de la phrase, la phrase (9b) est appliquée par le locuteur aux auditeurs pour lesquels il n'y a pas de respect, mais le suffixe :-*_eu_si* montre le respect envers le sujet. Le locuteur de la phrase (9c) respecte le sujet de la phrase en appliquant le suffixe :-*_eu_si* et des auditeurs avec le suffixe final honorifique -*_seub_ni_da* au lieu du suffixe -*da*.

(9)

a. 너의 애비는 집에 갔다.

nô_eui ai_bi_neun jib_e gass_da.
[nô][eui] [a_bi][neun] [jib][e] [ga_da][ôss][da]
toi+Post.gen Père+Post.nmtp maison+Post.à aller+ Mhor.suj+Mtp+St.déc.
Ton père est allé à la maison.

b. 나의 아버님은 집에 가시었다.
jô_eui a_bô_nim_i jib_e ga_si_ôss_da.
[jô][eui] [a_bô_nim][ga] [jib][e] [ga_da][eu_si][ôss][da]
moi+Post.gen Père.respect+Post.nmtp maison+Post.à aller+ Mhor.suj+Mtp+St.déc.
Mon père est allé à la maison.

c. 저의 아버님은 집에 가시었습니다.
jô_eui a_bô_nim_i jib_e ga_si_ôss_seub_ni_da.
[jô][eui] [a_bô_nim][ga] [jib][e] [ga_da][eu_si][ôss][seub_ni_da]
moi+Post.gen Père.respect+Post.nmtp maison+Post.à aller+ Mhor.suj+Mtp+St.déc.hso.
Mon père est allé à la maison.

Le suffixe -으시 :-*_eu_si* exprime le respect envers le sujet de la phrase. Il a la variante -시 :-*_si* après les racines ouvertes. Il se situe toujours après les racines verbales et adjectivales.

Le deuxième cas est comme le suffixe -*_seub_ni_da,* les suffixes finaux expriment le niveau de langue et de l'attitude d'honorification adaptée envers l'interlocuteur. Le tableau (2-8) montre six groupes d'honorification avec les suffixes finaux [26].

Tableau 2-8 Code pour les suffixes honorifiques

Nom coréen du niveau	Exemples	Niveau d'honorification	Code
합쇼 : *hab_syo :***[ha_da][eub_syo]**	-습니다 : -*_seub_ni_da*	très élevé	Hso
하오 : *ha_o:***[ha_da][o]**	-오 : -*_o*	élevé	Hao
하게 : *ha_ge:***[ha_da][ge]**	-네 : -*_ne*	bas	Hge
해라 : *hai_la:***[ha_da][ô_la]**	-어라 : -ô*_la*	très bas	Hla
해요 : *hai_yo:***[ha_da][ô_yo]**	-지요 : -*_ji_yo*	*élevé familial*	(Hyo)
해 : *hai:***[ha_da][ô]**	-어 : -*_ô*	*bas familial*	Hai

[26] [SUH] pp. 1014-1015

Nous n'avons pas utilisé la catégorie du code « Hyo ». Les suffixes Hso, Hao sont très rarement appliqués dans la parole moderne.

Dans le cas de la racine de l'honorification, les racines verbales se présentent elles-mêmes au niveau de respect du sujet parce que leur forme canonique contient le suffixe -으시 :-*_eu_si*. Dans l'exemple (10), le verbe '*manger*' a trois lexèmes.

(10)
a. **먹**다 : *mog_da* (manger sans respect)
b. **드시**다 : *deu_si_da* ;**잡수시**다 : *jab_su_si_da* (manger avec respect)

Ses séquences de suffixes n'ont pas le suffixe '*-_eu_si*' qui se situe juste après la racine.

2.2.1.2 Suffixe de modalité liée au locuteur

Dans les phrases (11), le suffixe -더- : -*_dô*- représente le point de vue du locuteur. Ce suffixe est incompatible avec le suffixe déclaratif -다 : -*da*, mais il est compatible avec le suffixe déclaratif 구나 : -*_gu_na*. le suffixe -더- : -*_dô*- signifie l'opinion du locuteur comme dans « je pense que ». Il a la variante -군 : -*_gun*.

(11)
a. 아버님이 집에 가시었군.
a_bô_nim_i jib_e ga_syôss_gun.
[a_bô_nim][i] [jib][e] [ga_da][eu_si][ôss][gu_na]
Père.respect+Post.nmtp maison+Post.à aller+ Mhor.suj+Mtp+St.déc.
Je pense que le père est allé à la maison. (Le fait est assuré par le locuteur)
b. 아버님이 집에 가시었더군.
a_bô_nim_i jib_e ga_si_gôss_dô_gun.
[a_bô_nim][i] [jib][e] [ga_da][eu_si][ôss][dô][gu_na]
Père.respect+Post.nmtp maison+Post.à aller+ Mhor.suj+Mtp+ Masp+St.déc.
Il semble que le père est allé à la maison. (Le fait n'est pas assuré par le locuteur)

2.2.1.3 Suffixes de temps et d'aspect

Nous représentons les suffixes de temps dans le tableau (2-9).

Tableau 2-9 Codes pour les suffixes de temps

Codes	Suffixes en forme canonique	Temps
tp+pre	는 : *neun*	Présent
tp+pass	었 : *ôss*	Passé ou Imparfait
tp+pp	었었 : *ôss_ôss*	Imparfait ou plus-que-parfait
tp+future	겠 : *gess*	Futur ou conditionnel
tp+pf	었겠 : *ôss_gess*	Conditionnel passé
tp+accomp	었 : *ôss*	Accompli

2.2.1.4 Catégories de suffixes finaux

Les suffixes finaux se situent à la fin de la séquence de suffixes verbaux et adjectivaux. Les verbes et les adjectifs se situent à la fin de la phrase ou de la proposition, Les exemples (12) suivants montrent les cas de suffixe terminal déclaratif -*_da* (12a), de suffixe conjonctif -*_go* (12b), de suffixe terminal déterminatif -*_eun* (12c), et de suffixe terminal nominalisateur -*_eum* (12d).

(12)

a. 그가 그녀에게 꽃을 주었다.
geu_ga geu_nyô_e_ge ggoch_eul ju_ôss_da.
[geu][ga] [geu_niô][e_ge] [ggoch][leul [ju_da][ôss][da].
Lui+Post.nmtp elle+Post.datif fleur+Post.accu donner+Mtpass+St.déclatif
Il a donné la fleur à elle.

b. 그가 그녀에게서 꽃을 받고 책을 주었다.
geu_ga geu_nyô_e_ge_sô ggoch_eul bad_go ju_ôss_da.
[geu][ga] [geu_nyô][e_ge_sô] [ggoch_eul] [bad_go] [chaig][leul] [ju_da][ôss][da].
Lui+Post.nmtp elle+Post.de fleur+Post.accu recevoir+Sc livre+Post.accu donner+Mtpass+St.déc.
Il a reçu la fleur d'elle et il lui a donné un livre.

c. 그가 그녀에게서 받은 꽃은 아름답다.
geu_ga geu_nyô_e_ge_sô bad_eun ggoch_eun a_leum_dab_da.
[geu][ga] [geu_nyô][e_ge_sô] [bad_da][neun] [ggoch][neun] [a_leum_dab_da][da].
Lui+Post.nmtp elle+Post.src recevoir+Sd fleur+Post.nmtp « être beau »+St.déc.
La fleur qu'il a reçue d'elle est belle.

La fleur qu'elle lui a donnée est belle.

d. (나는) 그가 그녀에게서 꽃을 받았음을 알았다.
(*na_neun*) *geu_ga geu_nyô_e_ge_sô ggoch_eul bad_ass_eum_eul al_ôss_da.*
([na][neun]) [geu][ga] [geu_nyô][e_ge_sô] [ggoch][leul] [bad_da][ôss][eum][leul] [al_da][ôss][da].
moi+Post.nmtp Lui+Post.ntmp elle+Post.de fleur+Post.accu recevoir+Mtp+Scomp+Post.accu savoir+Mtp+St.déc.
J'ai su le fait qu'il a reçu une fleur qu'elle lui a donnée.

Nous donnons les codes pour des suffixes finaux selon leur fonction, ce sont les suivants :

Tableau 2-10 Catégorie de suffixes finaux

Code		fonction
St	Suffixe terminal	Terminaison d'une phrase
Sc	Suffixe conjonctif	Terminaison d'une proposition et Conjonction entre constituants
Sd	Suffixe déterminatif	Marque d'une proposition relative
Sncomp	Suffixe complémentaire	Conjonctive d'une proposition complétive

Les suffixes finaux terminaux expriment les modes de la phrase (tableau 2-11).

Tableau 2-11 Codes des modes correspondant aux suffixes finaux

Code	Mode	Exemples avec 빨리 *bbal_li* : vite 뛰다 : *ddui_da* : courir
déc	Déclaratif	너는 빨리 뛴다 **[ddui_da][neun][da]** Tu+Post .nmtp vite courir+Mtpresent+St.déc. Tu cours vite.
int	Interrogatif	너는 빨리 뛰니 ? **[ddui_da][eu_ni].** Tu+Post .nmtp vite courir+Mtpresent+St.int ? Cours-tu vite ?
Excl	Exclamatif	너는 빨리 뛰네 **[ddui_da][ne].** *nô_neun bbal_li dduôn_da.* Tu cours vite !
Imp	Impératif	빨리 뛰어라 [ddui_da][ô_la] Vite courir+St.imp. Cours vite.
Prop	Propositif	빨리 뛰자 [ddui_da][ja] Vite courir+St.prop.. C'est mieux que nous allions courir vite.
inf	Infinitif	그는 빨리 뛰다. lui+Post .nmtp vite courir+St.inf. Courir vite.

Les suffixes conjonctifs se divisent en trois groupes selon la fonction conjonctive [LEE 2000] : conjonction coordonnée (13a), conjonction subordonnée (13b) et conjonction auxiliaire (13c).

(13)
a. 그는 식당에서 밥을 먹고 집으로 왔다.
gue_neun sig_dang_e_sô bab_eul mog_go jib_eu_lo oass_da.
[geu][neun] [mog_da][go] [jib][eu_lo] [o_da][ôss][da].
lui+Post.nmtp restaurant+Post.lieu « riz cuit »+Post.accu maison+Post.vers venir+Mtp+déc.
Il a mangé au restaurant et il est venu à la maison.
B 해가 뜨면 그는 일을 시작했다.
hai_ga ddeu_myôn geu_neun il_eul si_jag_haiss_da.
[hai][ga] [ddeu_da][eu_myôn] [geu][neun] [il][leul] [si_jag_ha_da][ôss][da].
Soleil+Post.nmtp « se lever »+Sc.cond lui+Post.nmtp travail+Post.accu commencer+Mtp+St.déc.
Si le soleil se lève, il a commencera le travail.
c. 그는 집에서 밥을 먹고 있다.
geu_neun jib_e_sô bab_eul môg_gô iss_da.
[geu][neun] [jib][e_sô] [bab][leul] [môg_da][gô] [iss_da][da].
lui+Post.nmtp maison+Post.lieu « riz cuit »+Post.accu manger+Sc exister+St.déc.
Il est en train de manger chez lui.

Nous mettons les codes (tableau 2-12) pour les suffixes.

Tableau 2-12 Codes de la fonction des suffixes conjonctifs

Code	Fonction
Coor	Coordination
Subor	Subordination

Nous ne pouvons pas distinguer la fonction par la forme de suffixes. Dans les exemples (14) avec le suffixe conjonctif -*_go*, les phrases (14a) qui expriment seulement les deux descriptions individuelles sont conjointes dans une phrase. Dans la phrase (14b). Le suffixe exprime les actions ordonnées. L'adjectif auxiliaire 싶다 :*sib_da* « vouloir » accompagne toujours des suffixes conjonctifs, avec le suffixe -*_go* il construit une locution du vouloir(14c). Et aussi, avec la racine *_iss_da*, le suffixe -*_go* construit une locution du temps progressif (13c).

(14)
a. 산은 높고 물이 맑다.
san_eun nop_go mul_i malg_da.

[san][neun] [nop_da][go] [mul][ga] [malg_da][da].
Montagne+Post.nmtp « être haut »+Sc.coor eau+Post.nmtp « être claire ».
La montagne est haute et les eaux sont claires.
b 그는 가방을 열고 책을 꺼냈다.
geu_neun ga_bang_eul yôl_go chaig_eul ggô_naiss_da.
[geu][neun] [ga_bang][leul] [yôl_da][go] [chaig][leul] [ggô_nai_da][ôss][da].
lui+Post.nmtp sac+Post.accu ouvert+Sc.subor livre+Post.accu sortir+Mtp+St.déc.
Il a ouvert le sac et il est sorti le libre.
c. 나는 집으로 가고 싶다.
na_neun jib_eu_lo ga_go sip_da.
[na][neun] [jib][eu_lo] [ga_da][go] [sip_da][da].
moi+Post.nmtp maison+Post.vers aller+Sc vouloir+St.déc.
Je veux aller à la maison.

La figure (2-1) montre la description des séquences avec le suffixe –고 :-*_go* après les racines verbales et adjectivales. Nous avons mis les codes Coor, Subor. Nous traiterons les séquences avec le suffixe conjonctif auxiliaire comme mots composés.

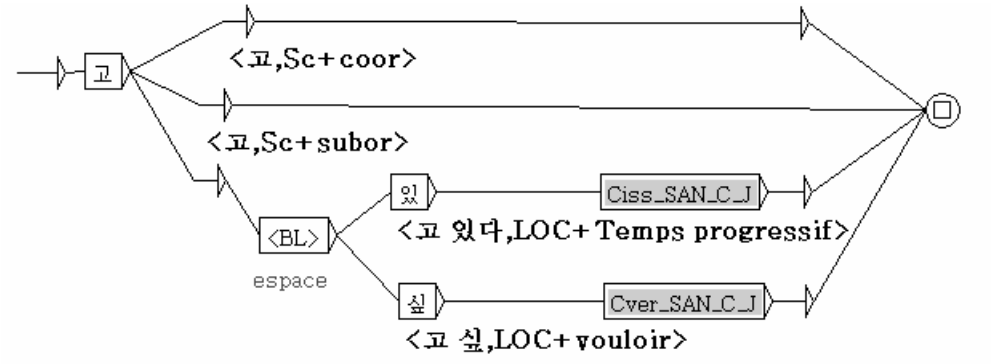


Figure 2-1 Suffixe -*_go*

2.2.2 Classification morphologique

La forme des variantes phonétiques des suffixes verbaux et adjectivaux a trois modes : invariables, variables par élément euphonique, variables avec la forme canonique 어 :*ô*.

La variation par élément euphonique supprime un contact de consonnes, un des deux est le dernier élément d'un morphème et l'autre est le premier élément du suivant.

A cause de l'harmonisation de voyelle avec Yin et Yang, les suffixes verbaux et adjectivaux dont la forme canonique commence par la syllabe '어 : _ô ' ont les variantes suivantes : '아 : _a', '어 : _ô', ' ㅕ : a', ' ㅏ : ô', '<E>', ' ㅣ :i'.

De plus, nous classifions les suffixes verbaux et adjectivaux en 9 groupes avec deux critères:

-. Les variations que subissent ces suffixes

-. Les variations que subit le contexte de ces suffixes (racines et autres suffixes) en fonction d'eux.

Nous allons montrer les variations entre suffixes et racines avec les racines d'exemple suivantes :

파랗다 : *pa_lah_da* (être bleu)
가다 : *ga_da* (aller)
잡다 : *jab_da* (saisir)
길다 : *gil_da* (être long)

pour les suffixes qui ont des variantes, nous prenons comme formes canoniques la forme qui contient l'élément euphonique ou avec la forme qui commence par la voyelle '어 :*_ô*'.

Invariable

Groupe 1

Tous les suffixes commençant par une syllabe qui a pour consonne initiale une des

consonnes ‘ㄱ : g’, ’ㄷ: d’,’ㅈ: j ’. Ils n’ont pas de variante.

Le suffixe -지 : _ji :

a. Un suffixe interrogatif non honorifique (n’est-ce pas ?)

b. Le suffixe terminal affirmatif

파랗지 *pa_lah_ji* (être bleu)
가지 *ga_ji* (aller)
잡지 *jab_ji* (saisir)
길지 *gil_ji* (être long)

La consonne ㅎ[h] et les consonnes ㄱ[g], ㄷ[d], ㅈ[j] deviennent les consonnes ㅋ[k], ㅌ[t], ㅊ[ch], pour certains verbes et adjectifs qui se terminent par le suffixe « -ha_da ». Par exemple, le mot 발전하게 : *bal_jôn_ha_ge* [27]est aussi écrit en 발전케 : *bal_jôn_ke*.

(15)
a. –하 :_*:ha* + 다*:_da* ➔ -ㅎ다 – *_hda* ➔ -타 : -*_ta* 다: St.déc
b. –하 :_*ha* + 지*:_ji* ➔ -ㅎ지 : -*_hji* ➔ -치 :-*_chi* 지: St.déc

D’après la description en syllabes, il est difficile de délimiter entre racine et suffixe dans une consonne. Avec la description en lettres alphabétiques, nous remplaçons les consonnes initiales ‘ㅋ : k’, ‘ㅌ : t’, ‘ㅊ : ch’ par les séquences « ㅎㄱ », « ㅎㄷ », « ㅎㅈ ». Nous les traitons comme des consonnes composées. La racine avec le suffixe -*_ha_da* a une variante «-*_h* » sans voyelle. La figure (2-2) montre la description de cette élisions.

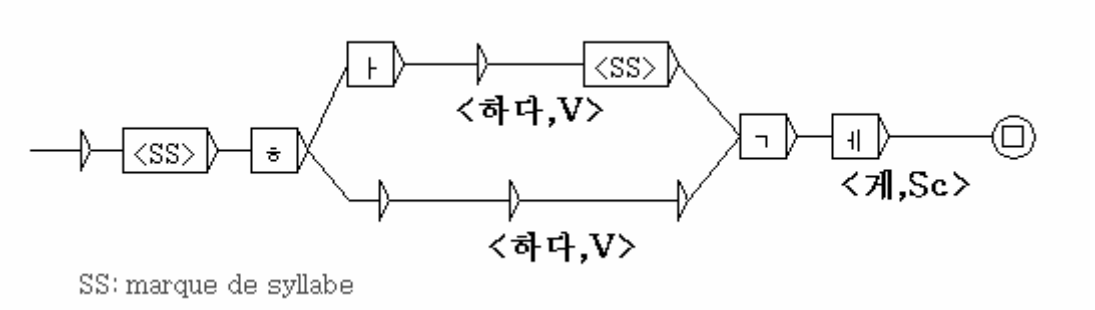


Figure 2-2 Variation des mots ha_da et suffixe -*_ge*

Le mot 용하다 : *yong_ha_da* « être habile » a la variation. Mais le mot 가하다 : *ga_ha_da* « ajouter, s’aggraver, joindre, être plus » n’a pas cette variation. Et le verbe *ha_da*

[27]발전하게 **[bal_jôn_ha_da][ge]** développer+Sc.adv, En développement

« faire » n'est pas de variation. Nous distinguons les racines avec le suffixe -*_ha_da* avec variation et les racines sans variation.

Groupe 2

Ils n'ont pas de variante. Mais ils affectent les racines qui se terminent par la consonne 'l' et par la consonne 'h'.

파라네 : *pa_la_ne* [pa_lah_da][ne] « être bleu »
가네 : *ga_ne* [ga_da][ne] « aller »
잡네 : *jab_la_ne* [jab_da][ne] « saisir »
기네 : *gi_ne* [jil_da][ne] « être long »

Groupe 3

Tous les suffixes commençant par les caractères '사: *_sa-*', '소: *_so_-*'. Ils n'ont pas de variante. Ils affectent les racines qui se terminent par la consonne '르 : l'. La consonne finale 'l' disparaît avant ces suffixes.

-사옵니까 *:_sa_ob_ni_da* : suffixe terminal très honorifique interrogatif
파랗사옵니까 : *pa_lah_sa_ob_ni_da*
가사옵니까 : *ga_sa_ob_ni_da*
잡사옵니까 : *jab_sa_ob_ni_da*
기사옵니까 : *gi_sa_ob_ni_da* [gil_da][sa_ob_ni_da]
Ce groupe de suffixes s'emploie très rarement.

Variables par élément euphonique

Groupe 4

Tous les suffixes commençant par les caractères '는 : *neun*' / 'ㄴ : –n', '습 : '*seub*'/'ㅂ : -*b*'. Ils affectent les racines qui se terminent par la consonne finale '르 : l'.

-습니다 :-*_seub_ni_d*a : le suffixe terminal déclaratif honorifique envers l'auditeur
파랗습니다 *pa_lah_seub_ni_da*
갑니다 *gab_ni_da*
잡습니다 *jab_seub_ni_da*
깁니다 *gib_ni_da* [gil_da][seub_ni_da]

Groupe 5

Tous les suffixes commençant par les caractères '-으르 :-_*eu_l-*' /'-르 :-_l-', '-음 :_*eum*-/ㅁ :-_m-'. Ils affectent les racines qui se terminent par les consonnes 's', 'h','d' et 'b'. Ils ont une variante courte après les voyelles et une seule consonne 'l'.

-으렵니다 : *_eu_lyôb _ni_da* : suffixe déclaratif honorifique volitif
*파라렵니다 : *pa_eu_liôb_ni_da*
가렵니다 *ga_yiôb_ni_da*
잡으렵니다 *jab_eu_lyôb_ni_da*
*길으렵니다 *gil_eu_lyôb_ni_da*

Le suffixe *_eu_liôb_ni_da* peut se combiner avec les racines des verbes. Il correspond à la forme de la contraction du suffixe honorifique volitif conjonctif *_eu_lyô_go* et du mot *_ha_da* (faire).

(16)
a) 그는 집에 가려고 하다.
geu_neun jib_e ga_lyô_go ha_da.
lui+Post.ntmf maison+Post.à aller+Sc faire+suff.terminal.
Il veut aller chez lui.
b) 그는 집에 가려다.
geu_neun jib_e ga_lyô_da.
lui+Post.ntmf maison+Post.à aller+Sc+suff.terminal.
Il veut aller chez lui.

-음 : *eum* :suffixe terminal nominalisateur
파람 *pa_lam* l'état d'être bleu
감 *gam* l'état de départ
잡음 *jab_eum* l'état de saisir
긺 *gilm* l'état de la longueur

Groupe 6

Tous les suffixes commençant par les caractères '은/ㄴ *:-_eun-/-n*', '을/ㄹ *:-_eul_-/-l_-*', '읍/ㅂ*:-_eub-/-_b-*', '으사/사 *:-_eu_sa-/-_sa*', '으오/오 *:-_eu_o-/-_o-*', '으우/우 *:-_eu_u-/-_u_-*'. Les suffixes affectent les racines qui se terminent par une des consonnes 's',

‘h’, ‘d’ et ‘b’.

-은/ㄴ : le suffixe déterminatif pour les racines adjectivales et déterminatif passé pour les racine verbales
파란 *_pa_lan* : qui est bleu
간 *_gan* : qui est allé
잡은 *_jab_eun* : qui a saisi
긴 *_gin* : qui est long
좁은 *_job_eun* **[job_da][eun]** « être étroit »+Sd : qui est étroit
L’autre suffixe est le suffixe déterminatif futur ‘-을/ㄹ’: ‘-_eul/l’ :

파랄 *_pa_lal* : d’aller être bleu
갈 *_gal* : d’aller aller
잡을 *_jab_eul* : d’aller saisir
길 *_gil* : d’aller être long

Groupe 7

Les suffixes ‘으니/니 : *_eu _ni’/’_ni’*, ‘으되/되 *:_eu_doi/_doi’*. Ils affectent les variantes de la racine qui se terminent par les consonnes ‘b’ et ‘s’.

Généralement, le suffixe se soude à une des variantes de la racine correspondante, mais ces suffixes sont soudés à toutes les variantes de la racine.

-_니 : suffixe terminal interrogatif
춥다 *_chub_da* (être froid)
춥니 ? *_chub_ni* : Est-ce que tu as froid ?
추우니 ? *_chu_u_ni* : Est-ce que tu as froid ? : *chu_u* variante de *chub*

Groupe 8

Les formes canoniques commencent par une voyelle '어 :*_ô*'. Le tableau (2-13) montre les variations de la syllabe '*_ô*' avec le dernier élément des racines.

Tableau 2-13 Variation des racines avec suffixes commençants par la voyelle 'ô'

Fin de racine+début de suffixe sous forme canonique	Formes observées		Phénomène
CV(eu) +_ôX		CV(ô)X	Elision (sauf C = 'l')
CV(n) +_V(n)X		CV(n)X	Réduction n := 'a' ou' ô'
CV(*zy*) +_V(*y*)X	CV(*zy*)_V(*y*)X	CV(*zy*)X	Réduction *z* = 'o', y = 'a' ou z = 'u', y = 'ô'
CV(x) +_V(y)X	CV(x)_ V(y)X	CV(xy)X	Voyelle composée
CV(i) +_V(*z*)X	CV(i)_V(*z*)X	CV(y*z*)X	Demi-voyelle, *z* = a, ô, o, u
C(l)V(eu) +_ôX	C(l)V(eu)_ 'l'ôX		Ajoute une consonne 'l'
C(h)V(a) + _V(ô)X	C(h)V(a)_V(yô)X	C(h)V(ai)X	

*. C : consonne initiale de dernière syllabe du morphème précédent
C(x) : une consonne initiale de dernière syllabe du morphème précédent avec 'x',
V : voyelle, V(x) = une voyelle 'x', X = consonne finale de première syllabe de suffixe
'_' : marque de limité de syllabe, '+' marque de limité entre racine et suffixe

Le tableau (2-14) montre les formes fléchies des racines avec le suffixe du temps passé –었 : -*_ôss* et le suffixe terminal -다 :-*_da*.

Tableau 2-14 Variante de groupe 8

Racine	Code de racine	Forme fléchie
끄다 : *ggeu_da* :étendre	VXV_X_J	껐다 :*ggôss_da*
서다 : *sô_da* :se lever, s'arrêter, se bâtir, être début	SAN_V_E	썼다 *ssôss_da*
가다 : *ga_da* :aller	SAN_V_E	갔다 :*gass_da*
개다 : *gai_da* :(le ciel) s'éclaircir	SAN_V_EJ	개었다 *gai_ôss_da* 갰다 *gaiss_da*
세다 : *se_da* :être fort, compter	SAN_V_EY	세었다 *se_da* 셌다 *sess_da*

쑤다 : *ssu_da* : faire bouillir	SAN_V_CJ	쑤었다 *ssu_ôss_da* 쒔다 *ssuôss_da*
꼬다 : *ggo_da* : tortiller, cordonner, se tordre	SAN_V_CY	꼬았다 *ggo_ass_da* 꽜다 *ggoass_da*
누르다 : *nu_leu_da* :être de couleur jaune	LLE_V_X	누르렀다 *nu_leu_lôss_da*
누르다 : *nu_leu_da* : presser en base, peser sur	VXV_X_J	눌렀다 *nul_lôss_da*
넓다 : *nôlb_da* :être large, étendu, généreux	SAN_C_J	넓었다 *nôlb_ôss_da*
하다 : *ha_da* :faire	HA0_V_H	하였다 *ha_iôss_da* 했다 *haiss_da*

La racine verbale 괴다 : *goi_da* « l'eau) s'arrêter, caler »a le code « SAN_V_CJ ». Il se soude avec la séquence du suffixe –었다 *: -_ôss_da*. On obtient les mots 괴었다 : *goi_ôss_da* et 괬다 :*goiô_da*. Normalement, la voyelle '괘 :oai' est présentée par la séquence « 'ㅗㅏㅣ :oai ». Mais pour cette contraction, nous la codons « ㅗㅣㅋ : oiô ».

Après la classification, nous avons construit les tableaux du type des racines verbales et adjectivales : tableau (2-15), tableau (2-16)

Tableau 2-15 Tableau de la classification des racines verbales

groupes de racines et caractéristiques			groupes de suffixes par la variante																				code de racines avec compatibilité des séquences de suffixes
variation	fin de la racine	harmonie vocalique	1	2	3	4		5		6		7		8									
			variante de suffixe																				
			-			S*	A*	S	A	S	A	S	A	_ô 어	ô ㅓ	_a 아	a ㅏ	<E>	i ㅣ	_yô 여		
SANS	C**	J	+	+	+	-	+	-	+	-	+	+	-	+	-	-	-	-	-	-		Cver SAN C J
		Y												-	-	+	-	-	-	-		Cver SAN C Y
	VSN	J				+	-	+	-	+	-	+	-	+	-	-	-	-	-	-		Cver SAN V J
		Y												-	-	+	-	-	-	-		Cver SAN V Y
	VXV	-												-	-	-	-	+	-	-		Cver SAN V E
	VCR	J												+	-	-	-	+	-	-		Cver SAN V EJ
	VCC	J												+	+	-	-	-	-	-		Cver SAN V CJ
		Y												-	-	+	+	-	-	-		Cver SAN V CY
HADA	a	-	+	+	+	+	-	+	-	+	-	+	-	-					+	+		Cver HA0 V H
	h			-										-								Cver HA1 H X
LLE	î	-	+	+	+	+	-	+	-	+	-	+	-	-								Cver VVX V X
LEU	C	J	-											-	+	-						Cver VVX X J
VXV		Y												-	-	-	+	-	-	-		Cver VVX X Y
LXX	l	J	+	-				-	-	+	-	-		+	-							Cver LXX L J
		Y												-	+	-						Cver LXX L Y
	V**	-	-	+	+	+	-	+	-	-	-	+	-	-								Cver LXX V X
DXC	d	-	+	+	+	-	+	-						-								Cver SXV S X
	l	J	-					-	+	-	+	-	+	+	-							Cver SXV V J
		Y												-	-	+	-	-	-	-		Cver SXV V Y
SXV	s	-	+	+	+	-	+	-						-								Cver SXV S X
	V	J	-					-	+	-	+	-	+	+	-							Cver SXV V J
		Y												-	-	+	-	-	-	-		Cver SXV V Y
BCU	b	-	+	+	+	-	+	-				+	-	-								Cver BCU B X
	V	J	-					+	-	+	-	+	-	-	+	-	-	-	-	-		Cver BCU U J
		Y	-					+	-	+	-	+	-	-								Cver BCU U X
			-											-	-	-	-	+	-	-		Cver BCU O Y
PUDA	V	-	+	+	+	+	-	+	-	+	-	+	-	-								Cver PUD V X
	p	Y	-											+	+	-	-	-	-	-		Cver PUD X CJ
NOHDA	V	-	+	+	+	+	-	+	-	+	-	+	-	-								Cver NOH C X
	p	Y	-											-	-	+	+	-	-	-		Cver NOH X CJ

*. ‘S’ signifie la variante Sans élément euphonique.

‘A’ signifie la variante Avec élément euphonique

**. V : voyelles, C : consonnes

Tableau 2-16 Tableau de classification des adjectifs

groupes de racines et caractéristiques			groupes de suffixes par la variante																			codes de racines avec compatibilité des séquences de suffixes
variation	fin de la racine	harmonie vocalique	1	2	3	4		5		6		7		8								
			variante de suffixe																			
			-			S	A	S	A	S	A	S	A	_ô 아	ô 나	_a 아	a 나	<E>	i 이	_yô 여	
SANS	C	J	+	+	+	-	+	-	+	-	+	+	+	+	-	-	-	-	-	-	-	Cadj SAN C J
		Y												-	-	+	-	-	-	-	Cadj SAN C Y
	VSN	J				+	-	+	-	+	-	+	-	+	-	-	-	-	-	-	Cadj SAN V J
		Y												-	-	+	-	-	-	-	Cadj SAN V Y
	VXV	-												-	-	-	-	+	-	-	Cadj SAN V E
	VCR	J												+	-	-	-	+	-	-	Cadj SAN V EJ
	VCC	J												+	+	-	-	-	-	-	Cadj SAN V CJ
		Y												-	-	+	+	-	-	-	Cadj SAN V CY
HADA	a	-	+	+	+	+	-	+	-	+	-	+	-	-					+	+	Cadj HA0 V H
	h			-										-							Cadj HA1 H X
LLE LEU VXV	î	-	+	+	+	+	-	+	-	+	-	+	-	-							Cadj VVX V X
	C	J	-											-	+	-					Cadj VVX X J
		Y												-	-	-	+	-	-	-	Cadj VVX X Y
HXX	h	-	+	-	+	-	+							-							Cadj HXX H X
	V		-	+	-	-	-	+	-	+	-	+	-	-					+	-	Cadj HXX V I
LXX	l	J	+	-				-	-	+	-	-		+	-						Cadj LXX L J
		Y												-	+	-					Cadj LXX L Y
	V	-	-	+	+	+	-	+		-	-	+	-	-							Cadj LXX V X
BCU	b	-	+	+	+	-	+	-				+	-	-							Cadj BCU B X
	V	J	-					+	-	+	-	+	-	-	+	-	-	-	-	-	Cadj BCU U J
		Y	-					+	-	+	-	+	-	-							Cadj BCU U X
			-											-	-	-	-	+	-	-	Cadj BCU O Y
SXV	s	-	+	+	+	-	+	-				+	-	-							Cadj SXV S X
	V	J	-					-	+	-	+	-	+	+	-						Cadj SXV V J
		Y												-	+	-					Cadj SXV V Y
VIDA**	V_i	J	+	+	+	+	-	+	-	+	-	+	-	+	+	-	-	-	-	-	Cida SAN V CJ
	V		+	+	+	+	-	+	-	+	-	+	-	-							Cida SAN X X
CIDA***	C i		+	+	+	+	-	+	-	+	-	+	-	+	-						Cida SAN V J

*. 'S' signifie la variante Sans élément euphonique.
'A' signifie la variante Avec élément euphonique
**. V : voyelles, C : consonnes

2.3 Classification des autres racines et des postpositions

Nous classifions les racines qui se soudent aux séquences des postpositions : les noms, les adverbes et les pronoms. La fonction du nom dans la phrase est indiquée par une postposition casuelle. La langue dispose également d'un jeu très riche de postpositions supplémentaires employées en combinaison avec les cas. Les postpositions de cas servent à désigner le sujet ou l'objet et les postpositions supplémentaires ont le même rôle que la préposition française.

Les racines de noms et d'adverbes n'ont pas de variante phonétique contrairement aux racines verbales et adjectivales. Certaines postpositions ont la variante selon la caractéristique phonétique des morphèmes précédents.

Les racines de nom et d'adverbe peuvent aussi exister sans postposition.

Au cas des séquences des mots de groupe de noms, seul le dernier élément se soude à la séquence de postpositions.

2.3.1 Noms autonomes

Les racines de nom peuvent être complétées par de nombreux suffixes. Par exemple, le suffixe pluriel -들 : -_*deul* peut se placer entre le nom et la séquence des postpositions. Les exemples (17) montrent les phrases avec la postposition nominative -이 : -_*i* qui désigne le nom comme sujet de la phrase. La postposition 로 : -_*lo* a la même fonction que la préposition *vers* en français (17a). Dans la phrase (17b) avec la postposition -에 :-_*e* « à », l'école est la destination. Dans la phrase (17c), le nom 버스 :_*bô_seu* « bus » est un complément circonstanciel de manière avec la postposition de manière 로 : -_*lo* « en». Dans la phrase (17d), la postposition -에 : -_*e* « à ,en » change le nom de temps en adverbe de temps.

Les racines nominales existent sans postposition, le mot 내일 :_*nai_il* « demain » est un nom, mais dans la phrase (17e), il est un adverbe sans postposition. De plus, il est possible qu'il existe avec la postposition spécifique -은 : -_*eun* qui a la fonction de la focalisation.

(17)

a. 학생들이 학교로 간다.
hag_saing_deul_i hag_gyo_lo gan_da.
[hag_saing][deul][ga] [hag_gyo][eu_lo] [gan_da][eun][da].
Etudiant+Suf.plur+Post.nmtp école+Post.« vers » aller+Mtp+St.déc.
Les étudiants vont à l'école.

b. 학생들이 학교에 간다.
hag_saing_deul_i hag_gyo_e gan_da.
[hag_saing][deul][ga] [hag_gyo][eu_lo] [gan_da][eun][da].
Etudiant+Suf.plur+Post.nmtp école+Post.lieu aller+Mtp+St.déc.
Les étudiants vont à l'école.

c. 학생들이 버스로 집에서 학교까지 간다.
hag_saing_deul_i bô_seu_lo jib_e_sô hag_gyo_gga_ji gan_da.
[hag_saing][deul][ga] [bô_seu][eu_lo] [hag_gyo][gga_ji] [gan_da][eun][da].
Etudiant+Suf.plur+Post.nmtp bus+Post.manière maison+Post.origine école+Post.destination aller+Mtp+St.déc.
Les étudiants vont en bus de chez eux à l'école.

d. 학생들이 9 시에 등교한다.
hag_saing_deul_i a_hob_si_e deung_gyo_han_da.
[hag_saing][deul][ga] [a_hob][si][e] [deung_gyo_ha_da][eun][da].
Etudiant+Suf.plur+Post.nmtp 9+heure+Post.temps « aller à l'école »+Mtp+St.déc.
A 9 heures, les étudiants vont en bus à l'école.

e. 학생들이 내일(은) 9 시에 버스로 학교에 등교한다.

hag_saing_deul_i nai_il(_eun) a_hob_si_ebô_seu_lo hag_gyô_e deung_gyo_han_da.
[hag_saing][deul][ga] [nai_il]([neun]) [a_hob][si][e] [bô_seu][eu_lo] [hag_gyo][eu_lo] [deung_gyo_ha_da][eun][da].
Etudiant+Suf.plur+Post.nmtp demain(+Post.spc) 9+heure+Post.temps école+Post.lieu « aller à l'école »+Mtp+St.déc.
Demain à 9 heures, les étudiants iront en bus à l'école.

Dans la phrase (17d), le mot 등교: 登校 :*_deung_gyo* est un mot sino-coréen, la syllabe 등 correspondant à un idéogramme 登 signifie dans ce mot «aller ou monter et arriver à la position ou au niveau », la syllabe 교 :校 signifie l'école. Dans les mots sino-coréens qui se composent de plus de deux syllabes associées aux idéogrammes, l'ordre des idéogrammes est l'ordre de la phrase chinoise SVO, le cas du mot 登校 a la structure « Verbe+Objet ». Avec le suffixe verbal -*_ha_da* « faire », il devient un prédicat 등교한다 : *deung_gyo_han_da*. La syllabe 교:*_gyo* associée à l'idéogramme 校 :*_gyo* qui signifie l'école n'est pas un mot autonome.

Les noms n'ont pas de variante mais selon le dernier élément des racines de nom, ils influent sur la forme de certaines postpositions. Nous classifions les noms en trois groupes selon l'influence phonétique : NV, NC, NL (tableau 2-17).

On applique la postposition de nominatif -이:-*_i* aux racines nominales qui se terminent par les consonnes finales, et la postposition canonique –가 :-*_ga* aux racines qui se terminent par les voyelles. Les postpositions se divisent en trois groupes : invariables, variables phonétiquement selon que la syllabe précédente est ouverte ou fermée et une postposition de manière ou d'intention -으로 :-*_eu_lo*. Le tableau montre les formes fléchies de : *–eu_lo* selon les racines nominales.

Tableau 2-17 Variantes de la postposition -*_eu_lo*

Code	Dernier élément	exemples	Mots fléchis	Forme fléchie de -*_eu_lo*	sens
NC	Consonnes sauf 'l'	산 : *san*	산으로 : *san_eu_lo*	-으로 :-*_eu_lo*	à la montagne
NV	Voyelles	바다 : *ba_da*	바다로 : *ba_da_lo*	-로 :–*_lo*	à la mer
NL	Consonne 'l'	건물 : *gun_mul*	건물로 : *gun_mul_lo*	-로 :–*_lo*	au bâtiment

La postposition spécifique -*_neun* est appliquée aux racines fermées (18a). Pour les racines ouvertes, on applique -*_eun* (18b) et aussi –*n* (18c) dans les conversations.

(18)
a. 산은 *san_eun* **[san][neun]** montagne+Post.spc
b.바다는 *ba_da_neun* **[ba_da][neun]** mer+Post.spc
c. 바단 *ba_dan* **[ba_da][neun]** mer+Post.spc

2.3.2 Noms non autonomes

Certains noms sont toujours accompagnés des modifieurs : Numéraux, Modifieurs phrastiques, de plus, des adjectifs indéfinis, des adjectifs démonstratifs. On les appelle noms incomplets [NAM 1994]

Il y a deux groupes de noms incomplets ; l'un est le groupe des classifieurs, qui entrent dans la séquence « <Nom comptable> <Numéral> <Nom incomplet>», les classifieurs sont sélectionnés par le nom comptable. Par exemple, dans la phrase (19a), le morphème 마리 : *ma_li* est une unité de compte des noms d'animal à comparer à 커피 두 잔 *kô_pi du jan* « deux tasse de café » ou 술 세 병 : *sul se byông* « trois bouteilles de vin ».

Les autres noms incomplets sont les mots grammaticaux comme le pronom *ce* de la séquence de pronoms relatifs « ce que » en français avec le morphème précédent étant le suffixe déterminatif qui a la fonction de pronom relatif « que », avec les racines verbales ou adjectivales (19b1), (19b2), (19b3), (19b22). Dans les exemples (19b), le suffixe de déterminatif passé -는 : -*_neun* termine les propositions avec les verbes et la proposition précède le nom incomplet suivant.

(19)

a. 길 위에 새 한 마리**가** 있다.
gil ui_e sai han ma_li_ga iss_da.
[gil] [ui][e] [sai] [han] [ma_li][ga] [iss_da][da].
rue dessus.N+Post.lieu oiseau un NI.animal+Post.nmtf exister+St.déc.
Dans la rue il y a un oiseau.

b1. 그는 나를 모르는 척 했다.
geu_neun na_leul mo_leu_neun chôg haiss_da.
[geu][neun] [na][leul] [mo_leu_da][neun] [chôg] [ha_da][ôss][da].
lui+Post.nmtp na+Post.accu « ne pas connaître »+Sd NI.« fait semblant » faire+Mtp+St.déc.
Il a fait comme s'il ne me connaissait pas.

b2.내가 태어난 곳은 서울이다.
nai_ga tai_ô_na_n gos_eun sô_ul_i_da.
[na][ga] [tai_ô_na_da][eun] [gos][neun] [séoul][i_da][da].
moi+Post.nmtp naître+Sd.passé NI.lieu+Post.nmtp Séoul+« être »+St.déc.
Le lieu où je suis né est Séoul.

b3. 내가 너를 만난 때가 여름이다.
nai_ga nô_leul man_nan ddai_ga yô_leum_i_da.

[na][ga] [nô][leul] [man_na_da][eun] [ddai][ga] [yô_leum][i_da][da].
moi+Post.nmtp toi+Post.accu rencontre+Sd.passé NI.temps+Post.nmtp « été »+St.déc.
Le temps où on s'est vu est l'été.

b4. 내가 너를 만난 것은 우연이다.
nai_ga nô_leul man_nan gôs_eun u_yôn_i_da.
[na][ga] [nô][leul] [man_na_da][eun] [gôs][neun] [u_yôn][i_da][da].
moi+Post.nmtp toi+Post.accu rencontre+Sd.passé NI. « fait que »+Post.nmtp « hasard »+St.déc.
C'est un hasard que je t'aie rencontré.

Le nom incomplet 때문 : *_ddai_mun* « la cause , la raison » construit une locution causale. La figure (2-3) montre les séquences de mots avec le blanc : <SP> qui précède le nom incomplet.

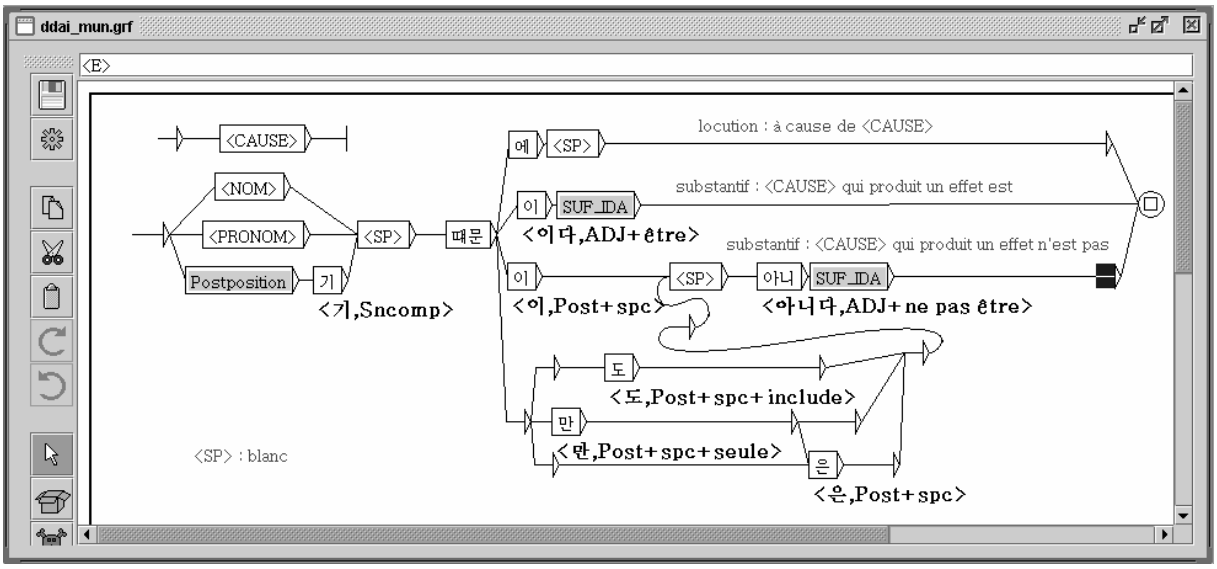


Figure 2-3 Locution causale

2.3.3 Pronoms et autres pro-formes

Le pronom est un mot qui remplace le nom ou désigne directement des personnes ou des objets non humains. Dans la langue coréenne, pour désigner la personne, on utilise fréquemment les mots appellatifs à la place des pronoms.

2.3.3.1 Appellatifs

Jusqu'au vingtième siècle, la société coréenne était hiérarchique et entre les gens, on utilisait les appellatifs pour désigner les personnes selon le niveau de respect, le niveau social dans la relation et la relation de famille. Dans la famille, le tableau montre des mots d'appellatif personnel entre les frères et les sœurs. Nous n'utilisons jamais le prénom pour appeler ou évoquer une personne plus âgée que le locuteur.

Tableau 2-18 Appellatifs des frères et sœurs

Locuteur	Homme		Femme	
Auditeur	Homme	Femme	Homme	Femme
Agé	형 : *hyông*	누나 : *nu_na*	오빠 : *o_bba*	언니 : *ôn_ni*
Agé avec respect	형님 : *hyông_nim*	누님 : *nu_nim*	오라버니 : *o_la_bô_ni* 오라버님 :*o_la_bô_nim*	언니 : *ôn_ni*
Jeune	동생 : *dong_saing*	누이 : *nu_i*	동생: *dong_saing*	동생 : *dong_saing*

De nos jours, la plupart des termes de respect anciens ne sont plus utilisés ou ont disparu. Mais l'utilisation des appellatifs existe toujours. Le mot 선생 : *sôn_saing* est un nom « le professeur » ou le nom général de celui qui instruit à l'établissement d'enseignement. Aujourd'hui, on l'applique à l'homme qui n'est pas forcement professeur.

L'inclusion des appellatifs dans la catégorie du pronom reste toujours un point litigieux.

2.3.3.2 Pronoms

La figure (2-4) montre le pronom à la première personne 나 :*na* « moi », le pronom 저 :*jô* « moi » est un morphème de respect envers les auditeurs. Normalement, la postposition de sujet -가 : -*_ga* se soude aux noms qui se terminent avec une syllabe ouverte phonétiquement. Mais avec les pronoms des première et deuxième personnes du singulier, on intercale la voyelle - ㅣ :-*i*. On peut dire que les pronoms personnels des première et deuxième personnes ont une forme variante avec une voyelle 'i' ou que la postposition de sujet '*ga*' a la forme variante - ㅣ가 :-*i _ga*. De même, la postposition de génitif -의 :-*_eui* transforme les pronoms en pronoms possessifs. Elle a la forme variante - ㅣ :-*i*. Avec les noms et avec les autres pronoms, elle n'a pas de forme variante.

Le phénomène d'ajouter une syllabe de voyelle 이 :*_i* se trouve aussi pour les prénoms personnels qui se terminent par les consonnes entre ceux-ci et les postpositions (20b).

(20)

a. 레아가 *Le_a_ga* « Léa+Post.nmtp », 레아를 *Lea_leul* « Léa+Post.accu »

b. 에릭이|에릭이가 *Eric_i* | *Eric_i_ga* « Eric+Post.nmtp », 에릭을 | 에릭이를 *Eric_eul, Eric_i_leul* « Eric+Post.accu »

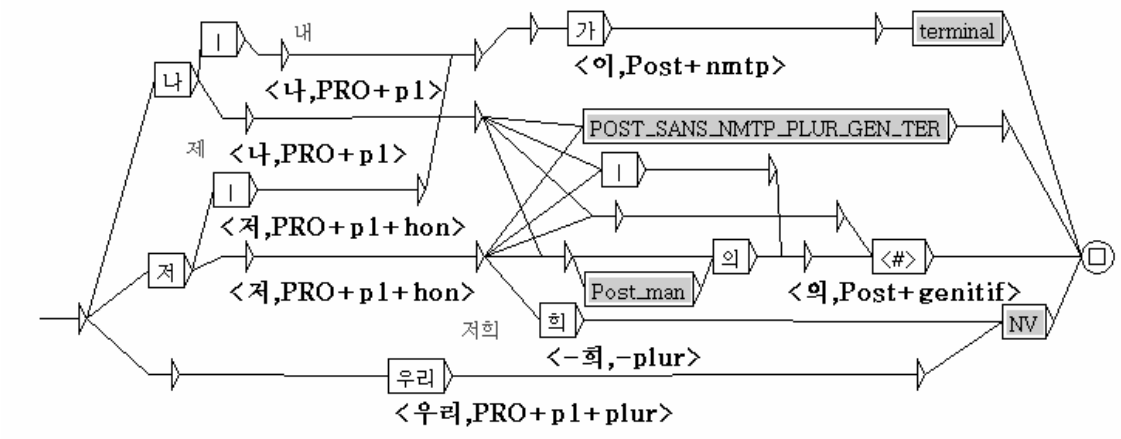


Figure 2-4 Pronoms personnels de 1ères personnes

Ce phénomène provoque une ambiguïté pour les prénoms qui sont mono-syllabique et en fermeture phonétique, comme un mot 영이가 *yông_i_ga* a les séquences

« *yông_i*.Prénom+Post.nmtp » ou « *yông*.Prénom+Post.nmtp » avec le graphe ci-dessous.

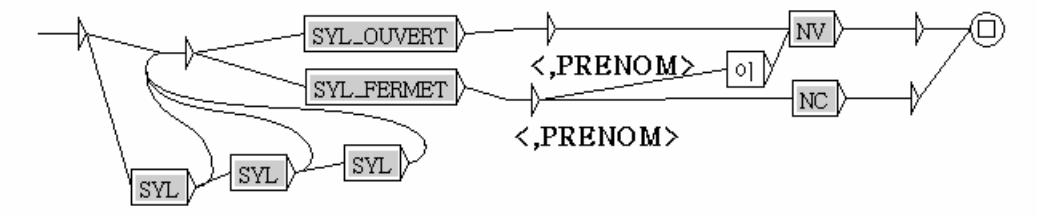


Figure 2-5 Recherche des prénoms personnels dans les mots inconnus

Les autres morphèmes qui désignent la première personne sont, 본인 :本人 :*bon_in*, 소자 :小子[28] :*so_ja*, 소인 : 小人[29] *: so_in*, 짐 :*jim*. Les morphèmes *so_ja, so_in, jim* étaient appliqués par le locuteur selon la relation ou le niveau social vers l'auditeur dans la société hiérarchique (tableau 2-19).

Tableau 2-19 Application des pronoms et relation entre locuteur et auditeur

Morphèmes	Relation	
	Locuteur, Annonceur	auditeur
so_ja	fils	parents
so_in	domestique	maître
jim	roi	quel qu'un

La figure (2-6) représente les pronoms de deuxième personne. Le pronom 너 : *nô* « toi» est la deuxième personne du singulier. Le pronom 당신 : *dang_sin* « vous» est un morphème de respect envers les auditeurs.

Les pronoms pluriels 우리 *: u_li* « nous », 저희 : *jô_heui* « nous + respect», 너희 : *nô_heui* : « vous» qui sont les pronoms pluriels peuvent se combiner aussi avec le suffixe pluriel -들 : *-_deul* :우리들 *: u_li_deul*, 저희들 : *jô_heui_deul*, 너희들 : *nô_heui_deul* comme les autres noms. Selon les grammairiens coréens, la syllabe –희 :*-_heui* est un morphème pluriel.

[28]小 : petit, 子 : fils
[29]小 : *_so* petit, 人 : *_in* humain : peuple

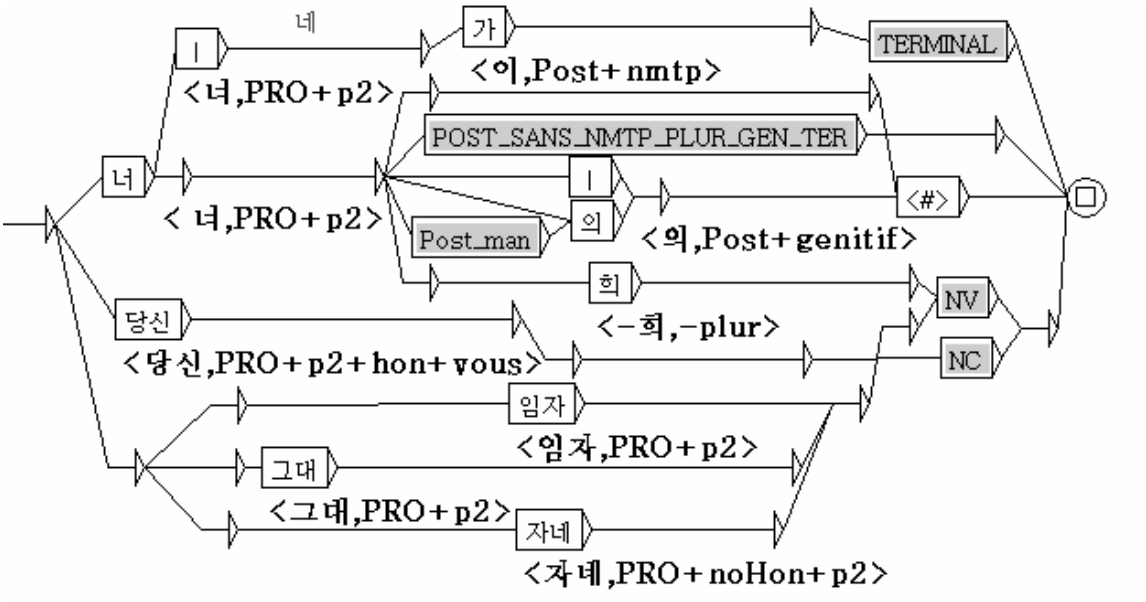


Figure 2-6 Pronoms personnels de la 2ème personne

Les pronoms de la troisième personne, sont constitués d'un des déterminants démonstratifs 이 : *_i*, 그 : *_geu*, 저 : *_jô* dans la figure (2-7) et d'un nom incomplet. Les déterminants démonstratifs se placent toujours avant un nom.

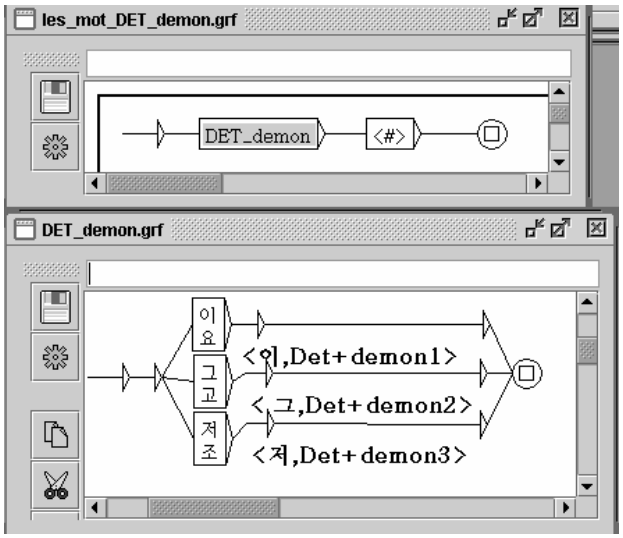


Figure 2-7 Déterminants démonstratifs

Ils ont les mêmes fonctions que les déterminants démonstratifs français, « ce…-ci, ce…-la, là-bas » pour l'espace et le temps.

Tableau 2-20 Sens des déterminants démonstratifs.

	Espace		Temps		Occasion	
이 : *_i*	Près du locuteur	이 집 : *_i _jib* cette maison	Présent, futur	이 때 : *_i _ddai* ce temps-ci	Présent, futur	이 번 :*i bôn* cette fois
그 : *_geu*	Près de l'auditeur	그 집 : *_i _jib* cette maison-ci	Passé	그 때 :*_geu _ddai* ce temps-là	-	-
저 : *jô*	Loin des auditeurs et locuteurs	저 집 : *_jô _jib* cette maison-là	-	-	Passé	저 번 : *jô bôn* l'autre fois

Ils ont des fonctions de deixis ou d'anaphore. Les conversations (21A) et (21B) montrent les applications.

(21)

A : 내가 이 생선을 가지고 그 곳으로 가겠다.

nai_ga i sang_sôn_eul ga_ji_go geu gos_eu_lo ga_gôss_da.

[na][ga] [i] [sang_sôn][leul] [ga_ji_da][go] [geu] [gos][eu_lo] [ga_da][gôss][da].

Moi+Post.nmtp ce poisson+Post.accu tenir+Sc ce lieu+Post.vers aller+ Mfut+St.déc.

Je vais emporter ce poisson là-bas.

B : 너가 이 곳에 그 생선을 가지고 오면 저 번에 못 다 한 그 이야기를 하자.

nô_ga i gos_e geu sang_sôn_eul ga_ji_go o_myôn jô bôn_e mos da han geu i_ya_gi_leul ha_ja.

[nô][ga] [i] [gos][e] [geu] [sang_sôn][leul] [ga_ji_da][go] [o_da][eu_myôn] [jô] [bôn][e] [mos] [da] [ha_da][neun] [geu] [i_ya_gi][leul] [ha_da][ja].

Toi+Post.nmtp ce.dét NI.lieu+Post.lieu ce.dét poisson+Post.accu avoir(tenir)+Sc venir+Sc.cond cela.dét fois+Post.temps « ne pas ».Adv tous faire+Sd.tpass cela.dét conversation+Post.accu faire+St.prop.

Après que tu aies apporté ici ce poisson-là, nous aurons cette conversation-là qui n'a pas été finie la dernière fois.

Les formes variantes des déterminants démonstratifs 요 : *_yo*, 고 : *_go* , 조 : *_jo* sont fréquentes à l'oral.

Les pronoms personnels de la 3^{e} personne sont obtenus par composition des déterminants démonstratifs (figure 2-8) avec le pronom 그 : *_geu* « lui ». Pour la femme, le pronom 그녀 : 그女 :*_geu_nyô* « la femme » comporte le suffixe sino-coréen 녀 : 女 : *_nyô* « la femme ».

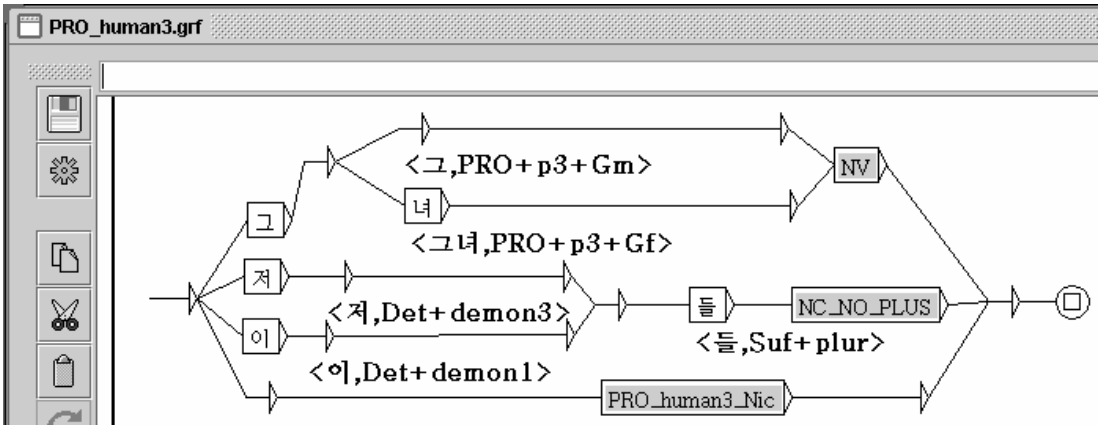


Figure 2-8 Pronoms de 3èmes personnes

Tableau 2-21 Codes pour les pronoms

	Respect		Genre	
	honorifique	Non honorifique	Masculin	Féminin
Code	hon	noHon	Gm	Gf

Dans la figure (2-9), le sous-graphe « PRO_human3_Nic » de la figure (2-8) montre les pronoms de 3e personne qui se composent avec les déterminants démonstratifs et les noms incomplets.

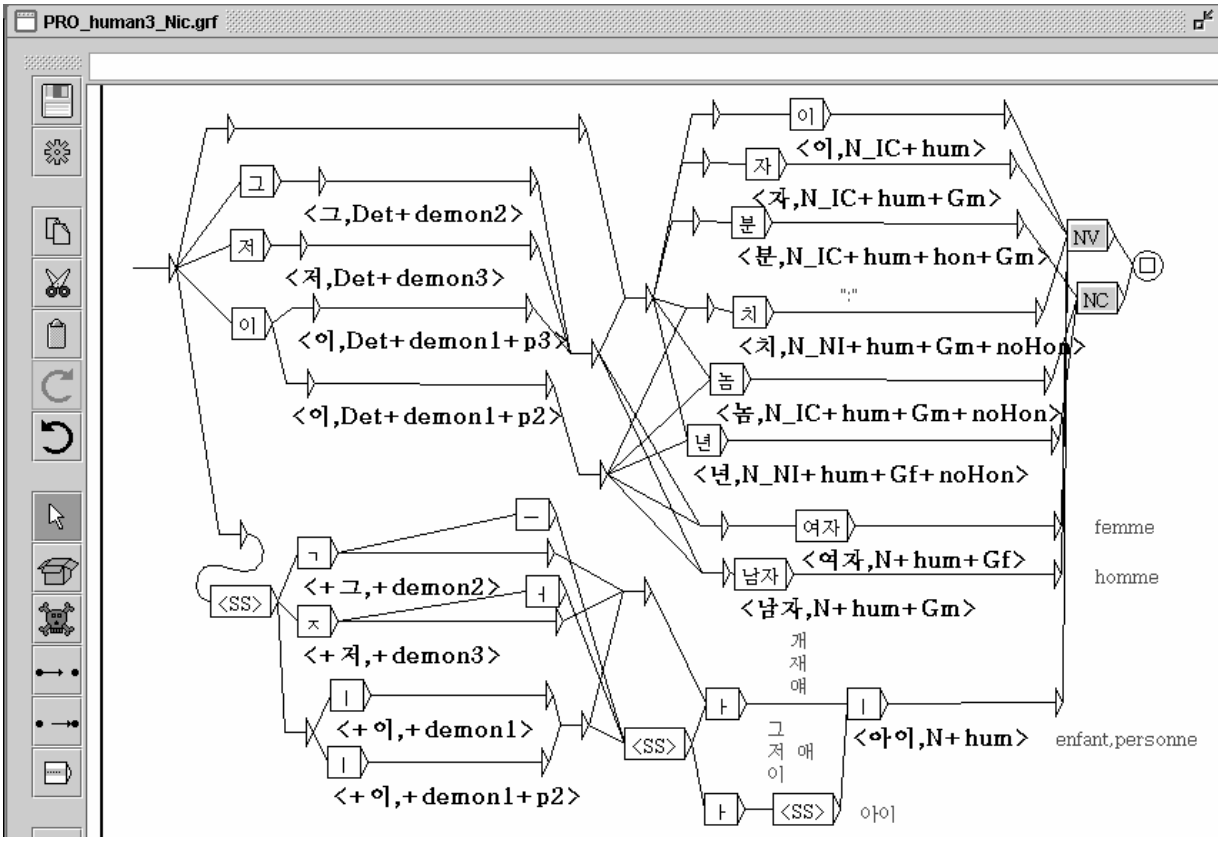


Figure 2-9 Pronoms humains avec les déterminants démonstratifs et les noms

Il existe aussi un problème de l'espacement entre les pronoms avec les déterminants démonstratifs qui sont en fait des déterminants.

Normalement, les déterminants doivent se séparer des noms suivants. En fait, ces mots sont écrits avec ou sans espacement.

Le graphe (2-10) montre les pronoms non humains. Les mots 이것 *: gi_gôs* , 그 : *geu_gôs* , 저것 : *jô_gôs* ont les formes variantes 이거 : *i_gô*, 그거 : *geu_gô*, 저거 : *jô_gô*. S'ils désignent une personne, la phrase est très péjorative.

La figure (2-10) montre les pronoms démonstratifs avec l'espacement Les morphèmes 여기 :*yô_gi* , 거기 :*gô_gi* , 저기 :*jô_gi* sont les pronoms des lieux démonstratifs (*ici, là, là-bas*). les mots 이리 : *i_li*, 그리 : *geu_li*, 저리 : *jô_li* ont un sens locatif (*par ici, par-là, par là-bas*) ou de manière.

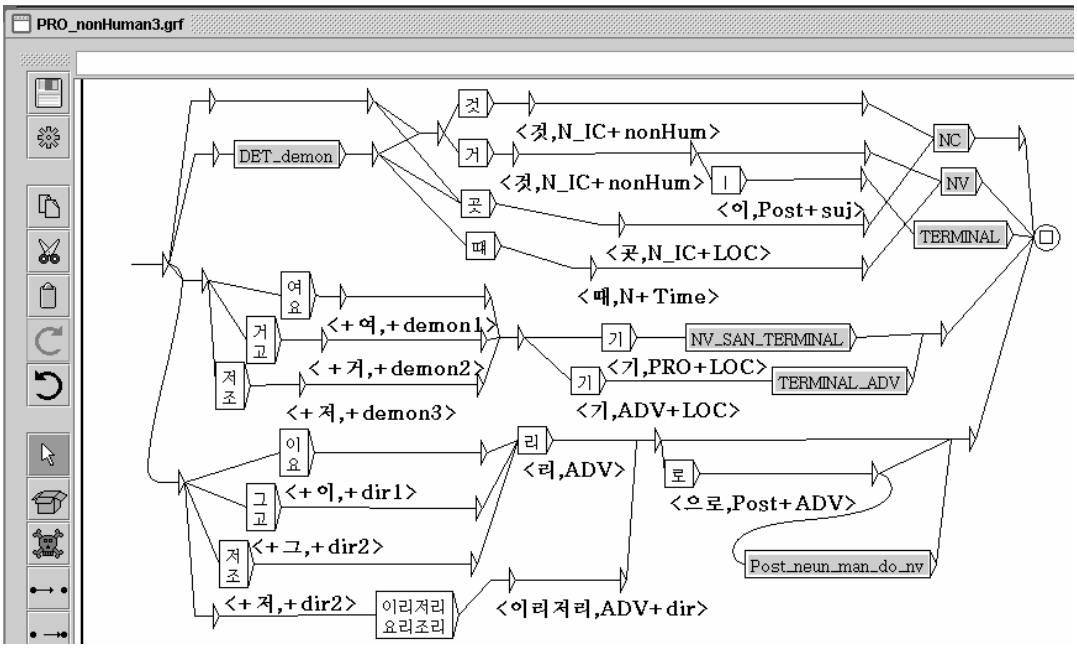


Figure 2-10 Pronoms non humains

La racine 여기 : *yô_gi* est le sujet de phrase (22a) avec la postposition nominative mais dans la phrase (22b) elle est un adverbe avec ou sans la postposition de lieu.

(22)
a.여기가 서울이다.
yô_gi_ga séoul_i_da.
[yô_gi][ga] [séoul][i_da][da].

Ceci+Post.nmtp séoul+être+St.déc.
Ici, c'est séoul.
b.그는 여기(에) 살고 있다.
geu_neun yô_gi (_e) sal_go iss_da.
[geu][neun] [yô_gi]([e]) [sal_da][go] [iss_da][da].
Lui+Post.nmtp ici vivre+Sc (être+exister)+St.déc. cf. « *–go iss_da* » : « être en train de »
Il est en train d'ici vivre.
Il habite ici.

La grammaire coréenne accepte les pronoms sous forme soudée sans espacement avec les déterminants démonstratifs et le nom incomplet d'objet non humain 것 :*_gôs* « le fait que, la chose que » comme la figure (2-11) : 이것 : *i_gôs*, 그것 : *geu_gôs*, 저것 : *jô_gôs* etc.

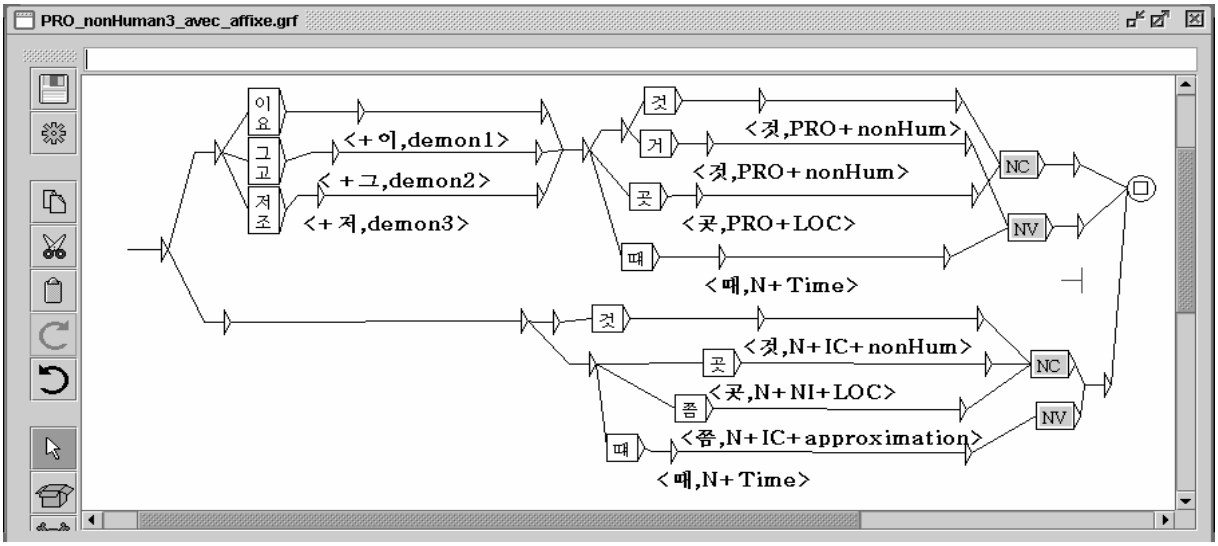


Figure 2-11 Pronoms non humains avec les affixes démonstratifs

2.3.3.3 Pronoms interrogatifs et indéfinis

Les morphèmes spécifiques qui apparaissent dans une phrase interrogative sont également utilisés pour donner une valeur d'indéfini à un constituant. La nature interrogative ou déclarative de la phrase est déterminée par le suffixe terminal.

Pour construire une phrase interrogative coréenne à partir de la phrase affirmative, on remplace les mots de la phrase (23a) par les mots indéfinis (tableau 2-22), le suffixe terminal déclaratif par la séquence de suffixes qui se termine par le suffixe terminal interrogatif et on ajoute à la fin de la phrase le symbole de la ponctuation interrogative « ? ». L'exemple suivant montre la transformation de la phrase affirmative (23a) en des phrases interrogatives avec le suffixe terminal interrogatif 니 :-*_ni* et les noms indéfinis. La phrase de réponse se compose de l'objet interrogé et du suffixe adjectival -이다 :-*_i_da* et du suffixe terminal déclaratif honorifique de locuteur -습니다 :-*_seub_ni_da*. Le nom incomplet 권 *: kuôn* accompagne la séquence 책 : livre +<numéral>+권 *: kuôn* .

Tableau 2-22 Noms indéfinis dans la phrase interrogative

Racine indéfinie	Sens	Objet indéfini
언제 : *ôn_je*	Quand	Temps
어디 : *ô_di*	Où	Lieu
누구 : *nu_gu*	Qui	Personne
무었 : *mu_ôss*	Quel	Objet

(23)

a. 어제 학교에서 학생들에게 책 2 권씩 주었다.
ô_je hag_kyo_e_sô hag_saing_deul_e_ge chaig do kuon_ssig ju_ôss_da.
[ô_je] [hag_kyo][_e_sô] [hag_saing][_deul][_e_ge] [chaig] [do] [kuon][ssig] [ju_da][ôss][da].
Hier école+Post.lieu étudiant+Suf.plur+Post.datif livre 2 N.IC.livre+Post.chaque donner+Mtp+St.déc.
Hier, à l'école, on a donné deux livres par personne aux étudiants.
b. 언제 학교에서 학생들에게 책 2 권씩 주었니 ?
Quand a-t-on donné deux livres par personne aux étudiants ?
어제입니다.
C'est hier.
c. 어제 어디(에)서 학생들에게 책 2 권씩 주었니 ?
Hier, où a-t-on donné deux livres par personne aux étudiants ?
학교(에서)입니다.
C'est à l'école.

d. 어제 학교에서 누구에게 책 2 권씩 주었니 ?
Hier, à l'école, à qui a-t-on donné deux livres par personne ?
학생들(에게)입니다.
C'est aux étudiants.
e.어제 학교에서 학생들에게 무었을 주었니 ?
Hier, à l'école, qu'est ce qu'on a donné aux étudiants ?
책입니다.
Ce sont des livres.
책 2 권씩입니다.
(Ce sont) deux livres par personne.
f. 어제 학교에서 학생들에게 책을 몇 권씩 주었니 ?
Hier, à l'école, combien de livres a-t-on donné aux étudiant par personne ?
2 권입니다.
(C'est) deux.
h. 언제 어디에서 누구에게 책을 2 권씩 주었니 ?
Quand, où et à qui a-t-on donné deux livres par personne ?
어제 학교에서 학생들에게입니다.
C'est hier, à l'école, aux étudiants.

La figure (2-12) montre d'autres racines indéfinies : le pronom indéfini d'humain inconnu 아무 : *a_mu*, le déterminant indéfini de nombre 몇 : *myôch*, un adverbe de cause 왜 : *oai* et le pronom indéfini de quantité 얼마 : *ôl_ma*.

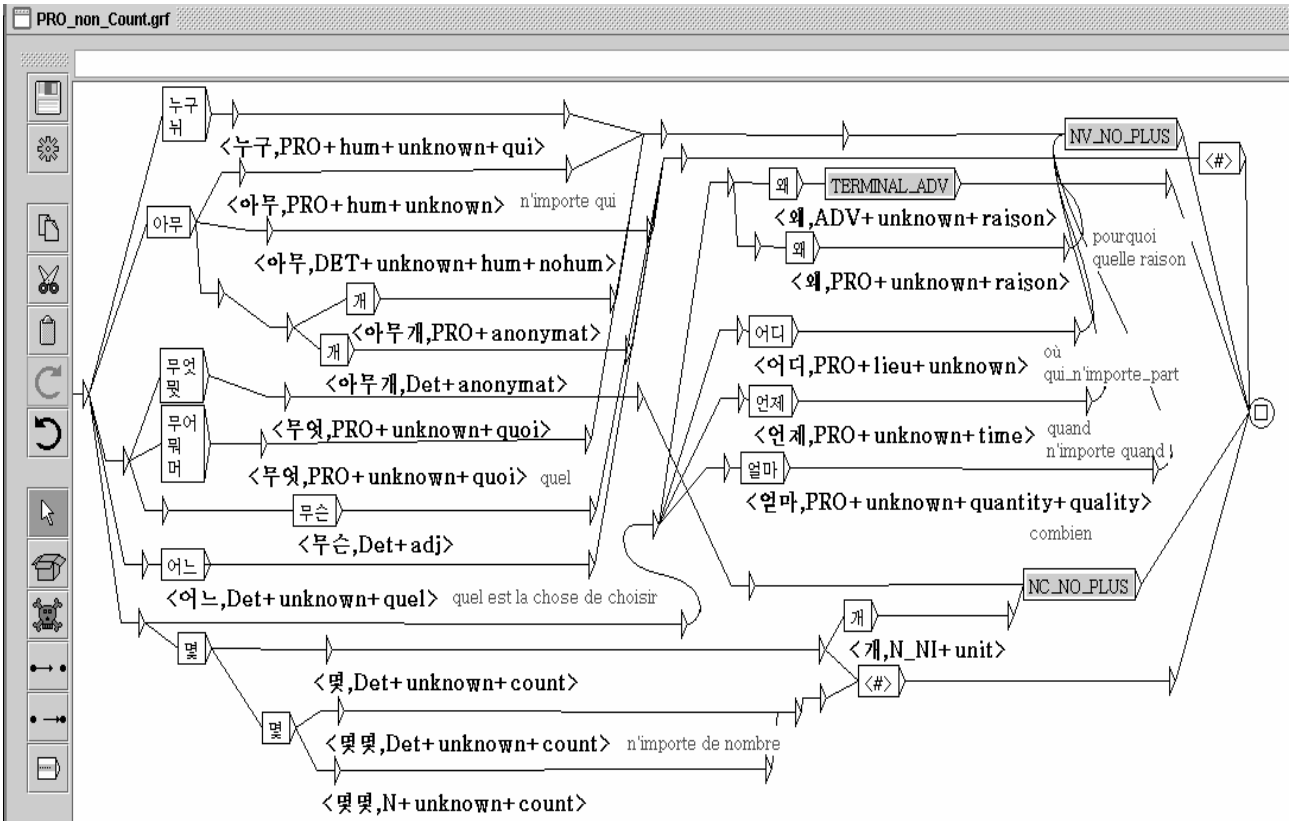


Figure 2-12 Mots indéfinis

Avec le suffixe déclaratif en fin de phrase, on peut utiliser les mots interrogatifs, mais le sens des mots interrogatifs gagne une valeur indéfinie et ils ont une fonction de remplacement comme le pronom (24).

(24) (언제+어제) (어디+학교)에서 (누구+학생들)에게 책을 2 권씩 주었다.

(à quelque temps+hier), (quelque part+à l'école), on a donné deux livres par personne à (quelqu'un+les étudiants).

2.3.3.4 Pro-adjectif, pro-verbe, pro-verbe interrogatif, pro-adjectif interrogatif

De la même façon que les pronoms remplacent les noms, en coréen, il existe aussi des pro-verbes et des pro-adjectifs dans la figure (2-13).

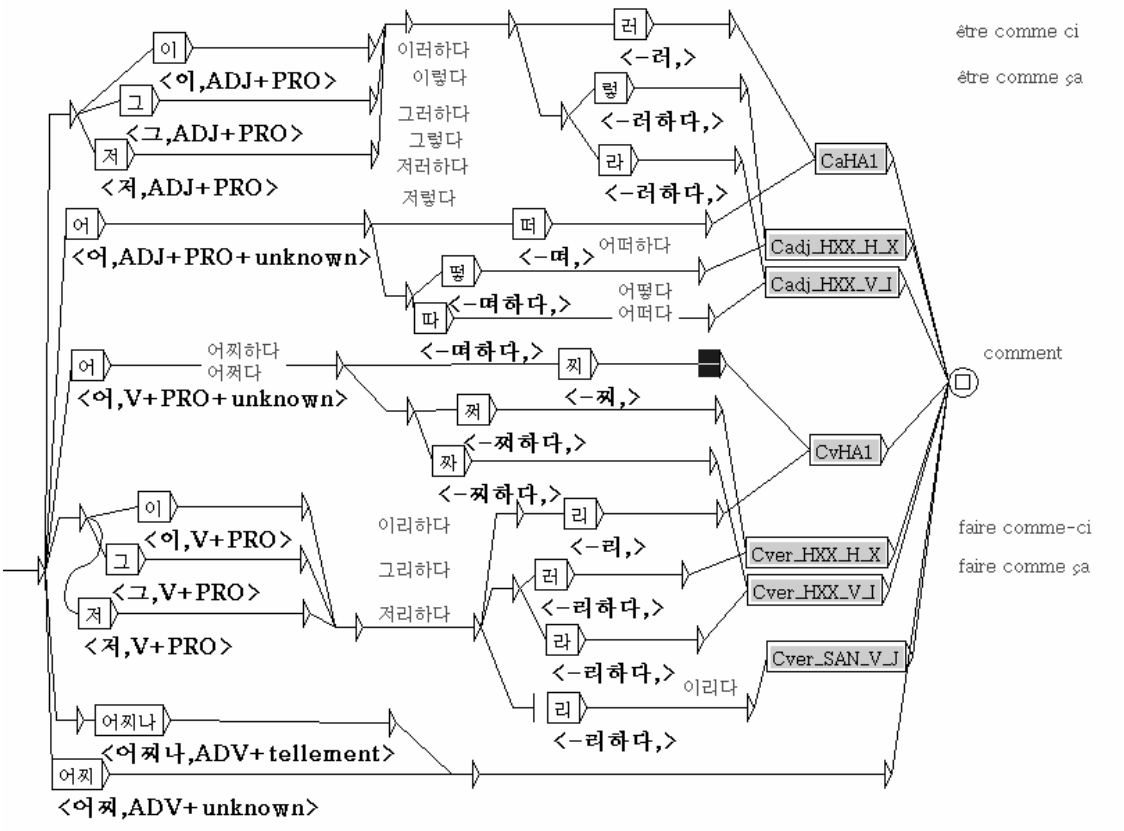


Figure 2-13 Pro-mots coréens

De nos jours, la fréquence de l'application des mots qui contient la racine 어찌하다 : *ô_jji_ha_ha* (comment que faire) est en train d'être remplacée par la séquence 어떻게 하다 : *ô_ddôh_ge ha_da* « (être comment)+Sc.adv faire+St.inf » . La racine

어떻다 : *ô_ddôh_da* est une variante de la racine 어떠하다 : *ô_ddô_ha_da* « être comment». Le graphe ci-dessus montre les pro-adjectifs et le pro-verbe, l'adjectif indéfini 어떠하다 : *ô_ddô_ha_da* « être comment », le verbe indéfini 어찌하다 : *ô_jji_ha_da* « comment que faire ».

Le pro-adjectif 그러하다 : *geu_lô_ha_da* a la variante 그렇다 : *geu_lôh_da*. Les pro-verbes (이+그+저)리하다 : *(_i+jô+geu)_li_ha_da* ont les variantes (이+그+저)리다 : *(i+jô+geu)_li_da* d'après le dictionnaire[30]. Mais nous n'avons pas trouvé ces variantes dans le texte et nous n'avons pas inclus ces variantes dans le graphe.

[30] Dictionnaire *keun_sa_jôn*

2.3.4 Adverbes

Il existe plusieurs types morphologiques d'adverbes :

-. Racines nominales avec les postpositions

-. Racines verbales et adjectivales avec les suffixes verbaux et adjectivaux

-. Racines adverbiales avec les postpositions.

Dans la phrase (25a), le groupe nominal 빠른 속도로 :*bba_leun sog_do_lo* joue le rôle d'un adverbe de manière avec une postposition –로 : -*_lo* Dans la phrase (25b) l'adverbe 빠르게 : *bba_leu_ge* comporte un adjectif 빠르다 : *bba_leu_da* (être rapide) et le suffixe conjonctif -게 : -*_ge* mais dans la phrase (25c), le mot est un adverbe 빨리 : *bba_li* (vite).

(25)

a.그는 빠른 속도로 나에게 다가왔다.
geu_neun bba_leun sog_do_lo na_e_gei da_ga_oass_da.
[geu][neun] [bba_leu_da][eun] [sog_do][eu_lo] [na][e_gei] [da_ga_o_da][ôss][da].
Lui+Post.nmtp «être rapide»+Sd vitesses+Post.manière moi+Post.datif approcher +Mtp+St.déc.
Il m'a approché à vitesse rapide.
Il m'a vite approché.

b.그는 빠르게 나에게 다가왔다.
geu_neun bba_leu_ge da_ga_oass_da.
[geu][neun] [bba_leu_da][ge] [na][e_gei] [da_ga_o_da][ôss][da].
Lui+Post.nmtp «être rapide»+Sc.conj moi+Post.datif approcher+Mtp+St.déc.
Il m'a en étant rapide approché.
Il m'a vite approché.

c.그는 나에게 빨리 다가왔다.
geu_neun bba_li da_ga_oass_da.
[geu][neun] [bba_li] [na][e_gei] [da_ga_o_da][ôss][da].
Lui+Post.nmtp moi+Post.datif vite.Adv approcher+Mtp+St.déc.
Il m'a vite approché.

Dans cette section, nous traitons des racines adverbiales et de leurs dérivés mais pas des adverbes dérivés d'autres types de racines. Les adverbes peuvent se souder aux séquences des postpositions mais ils ne se soudent pas avec les postpositions nominatives, accusatives ou génitives. Le tableau ci-dessous montre les racines adverbiales.

Tableau 2-23 Classification des adverbes

Adverbe	
Simple	잘 : *jal* (bien), 못 : *mos* (mal), 또 : *ddo* (encore), 너무 : *nô_mu* (très, trop, plus que suffisant(e)), 빨리: *bbal_li* (vite)
Onomatopéique	살짝 :*Sal_jjag*(l'attention à pas de loup et sans attirer)
Composé	잘못 :*jal_mos* (pas bien) 빨리빨리:*bbal_li_bbal_li* (vite vite)

Les adverbes dérivés sont les racines qui se composent des racines verbales, adjectivales, nominales avec les suffixes dérivés. Dans l'exemple (26), les racines adjectivales terminées par le suffixe -하다 :–*_ha_da* produisent un dérivé en remplaçant -*_ha_da* par le suffixe -히 : -*_hi*.

(26) 넉넉히
:*nôg_nôg_hi*
[nôg_nôg_ha_da][i]
« être (suffisant+abondant+généreux) »+Suf.adverbial
assez

Nous avons construit des graphes pour les adverbes qui peuvent se combiner en séquences. Un morphème « 반짝 :*ban_jjag* » qui est une onomatopée [HAN 2000], évoque une lumière clignotante, étincelante ou scintillante. Dans la figure (2-14) le graphe montre les séquences avec ce lexème. Voici quelques mots représentés dans ce graphe :

(27)
a. 반짝: *ban_jjag*: «une fois clignant »
b. 반짝반짝 : *ban_jjag_ban_jjag* : «plusieurs fois clignant ou plusieurs lieux être clignants »
c. 반짝하다 : *ban_jjag_ha_da* [-*ha_da* : suffixe verbal intransitif]
d. 반짝거리다 : *ban_jjag_gô_li_da* [*_gô_li_da*: suffixe verbal intransitif: montre la façon de]
e. 반짝이다: *ban_jjag_i_da* [-*_i_da* : être]
f. 반짝되다 : *ban_jjag_doi_da* [suffixe verbal intransitif : montre l'état de]
g. 반짝반짝하다: *ban_jjag_ban_jjag_ha_da* [renforcement du mot *ban_jjag_ha_da*]
h. 반짝반짝거리다 : *ban_jjag_ban_jjag_gô_li_da* [renforcement du mot *ban_jjag_gô_li_da*]
i. 반짝반짝이다: *ban_jjag_ban_jjag_i_da* [renforcement du mot *ban_jjag_i_da*]
j. 반짝반짝되다: *ban_jjag_ban_jjag_doi_da* [renforcement du mot *ban_jjag_doi_da*]
k. 되다: *doi_da* : arriver à un état, devenir
l. 하다 : *ha_da*. : faire

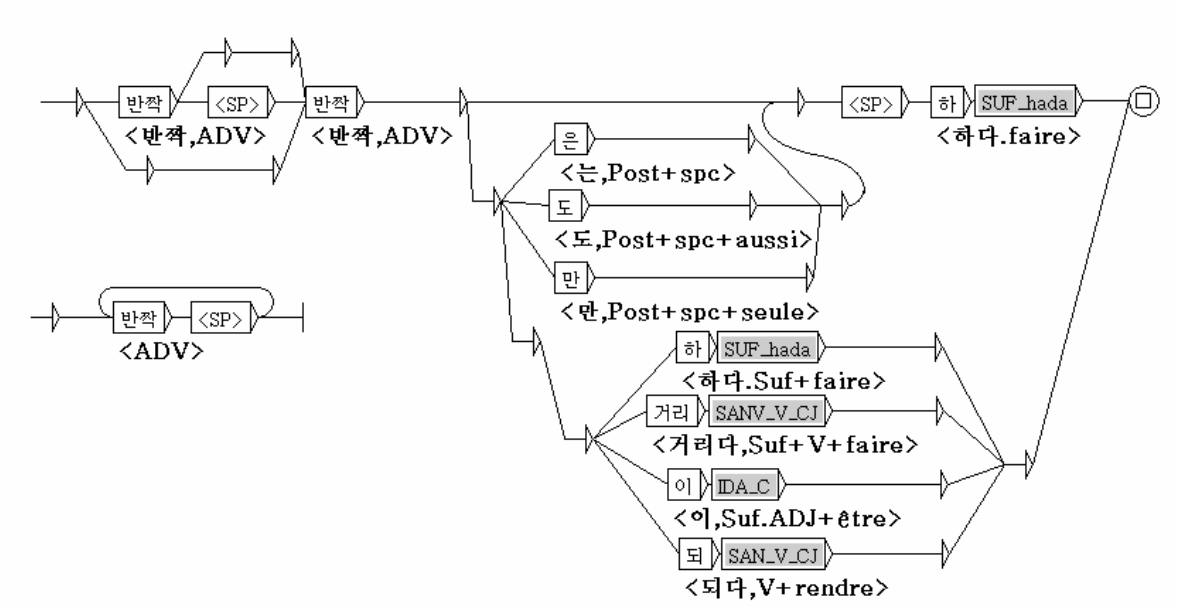


Figure 2-14 Séquences des mots avec le mot "반짝 :ban_jjag"

La figure (2-15) représente tous les mots qui ont la même caractéristique morphologique que 반짝 :*ban_jjag*.

Pour la construction du dictionnaire sous forme de liste, il existe trop de répétitions.

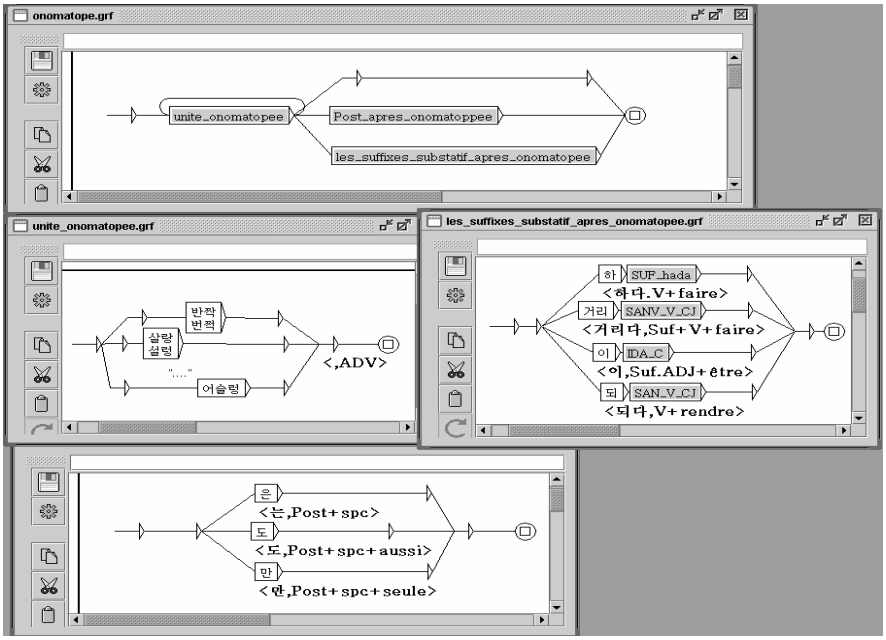


Figure 2-15 Description des mots avec les onomatopées avec sous-graphes

La description d'un mot sous forme d'automate des mots en morphèmes donne les effets suivants :

- Facilité de description d'un mot par la séquence des morphèmes.
- Diminution de la taille du dictionnaire.
- Cohérence de la description entre le mot « *ha_da* » et le suffixe « *-_ha_da* ».
- Concentration des informations autour d'un lexème.

2.3.5 Postpositions

D'après la grammaire coréenne générale, la postposition est une catégorie grammaticale. Elle se soude avec les racines nominales et adverbiales. Selon leur fonction grammaticale, les postpositions sont divisées en deux types : les postpositions casuelles et les postpositions auxiliaires ou spécifiques.

La combinaison entre les postpositions casuelles et les postpositions auxiliaires se fait de différentes façons. Dans l'exemple (28), la phrase (28a) montre les postpositions à cas avec le verbe *ju_da* « donner » et les combinaisons avec les postpositions, qui indiquent le nominatif, l'accusatif et le datif. Les autres phrases comportent une postposition auxiliaire - 만 : -*_man* « seule, seulement ». Dans les séquences de postpositions, avec les postpositions nominatives et accusatives, elle se situe à gauche (28b2, 28d2) mais avec les postpositions datives elle se situe à droite (28c2). Elle existe aussi sans postpositions casuelles (28b1, 28c1, 28d1).

(28)

a. 그가 너에게 책을 주었다.
geu_ga nô_e_ge chaig_eul ju_ôss_da.
[geu][ga] [nô][e_ge] [chaig][eul] [ju_da][ôss][da].
lui+Post.nmtp toi+Post.datif livre+Post.accu donner+Mtp+St.déc.
Il t'a donné le livre.

b1. 그**만** 너에게 책을 주었다.
geu_man nô_e_ge chaig_eul ju_ôss_da.
[geu][man] [nô][e_ge] [chaig][eul] [ju_da][ôss][da].
C'est seulement lui qui t'a donné le livre.

b2. 그**만이** 너에게 책을 주었다.
geu_man_i nô_e_ge chaig_eul ju_ôss_da.
[geu][man][ga] [nô][e_ge] [chaig][eul] [ju_da][ôss][da].
C'est seulement lui qui t'a donné le livre.
b3.*그가**만** 너에게 책을 주었다.
geu_ga_man nô_e_ge chaig_eul ju_ôss_da.
C'est seulement lui qui t'a donné le livre.

c1. 그가 너만 책을 주었다.
geu_ga nô_man chaig_eul ju_ôss_da.
[geu][ga] [nô][man] [chaig][eul] [ju_da][ôss][da].

C'est à toi seulement qu'il a donné le livre.

c2.그가 네에게**만** 책을 주었다.
geu_ga_man nô_e_ge_man chaig_eul ju_ôss_da.
C'est à toi seulement qu'il a donné le livre.

d1. 그가 너에게 책**만** 주었다.
geu_ga nô_e_ge chaig_man ju_ôss_da.
[geu][ga] [nô][e_ge] [chaig][man] [ju_da][ôss][da].
C'est seulement le livre qu'il t'a donné.

d2. 그가 너에게 책**만**을 주었다.
geu_ga nô_e_ge chaig_man_eul ju_ôss_da.
[geu][ga] [nô][e_ge] [chaig][man] [ju_da][ôss][da].
C'est seulement le livre qu'il t'a donné.

L'exemple (29) montre les séquences avec l'autre postposition du nominatif 는 :-*_neun* à la place de la postposition –*_ga* avec -*_man*. Les séquences montrent la postposition de l'accusatif dans le même ordre « nom » + « postposition auxiliaire » + « postposition casuelle ».

(29)
a. 그**는** 너에게 책을 주었다.
geu_neun nô_e_ge chaig_eul ju_ôss_da.
[geu][neun] [nô][e_ge] [chaig][eul] [ju_da][ôss][da].
lui+Post.nmtp toi+Post.datif livre+Post.accu donner+Mtp+St.déc.
Il t'a donné le livre.

b.그**만은** 너에게 책을 주었다.
geu_man_eun nô_e_ge chaig_eul ju_ôss_da.
[geu][man][eun] [nô][e_ge] [chaig][eul] [ju_da][ôss][da].
C'est lui seulement qui t'a donné le livre.

L'exemple (30) montre que la postposition -*_neun* a une parenté syntaxique avec la postposition *_man* et dans les phrases (30b), (30b2), les séquences variantes de –에는 : *_e_neun* sont –엔 -*_en*.

(30)
a. 서울에 거주자가 많다.
sô_ul_eun gô_ju_ja_ga manh_da.
[sô_ul][e] [gô_ju_ja][ga] [manh_da][da].
Séoul+Post.lieu habitant+Post.nmtp « être plein »+St.déc.
Il y a beaucoup d'habitants à Séoul.

b. 서울에는 거주자가 많다.
sô_ul_e_neun gô_ju_ja_ga manh_da.
[sô_ul][e][neun] [gô_ju_ja][ga] [manh_da][da].
Séoul+Post.lieu+Post.spc habitant+Post.nmtp « être plein »+St.déc.
À Séoul il y a beaucoup d'habitants.

b1. 서울**은** 거주자가 많다.
sô_ul_eun gô_ju_ja_ga manh_da.
[sô_ul][neun] [gô_ju_ja][ga] [manh_da][da].
Séoul+Post.spc habitant+Post.nmtp « être plein »+St.déc.
À Séoul il y a beaucoup d'habitants.

b2. 서울엔 거주자가 많다.
Sô_ul_en gô_ju_ja_ga manh_da.
[sô_ul][e][neun] [gô_ju_ja][ga] [manh_da][da].
Séoul+Post.lieu+Post.spc habitant+Post.nmtp « être plein »+St.déc.
À Séoul il y a beaucoup d'habitants.

Dans les phrases (31) construites avec le verbe 주다 : *ju_da* « donner », la postposition '-에게 : -*_e_ge*' est appliquée aux noms et pronoms humains pour indiquer le complément d'objet indirect . Si on l'applique aux mots non humains, elle prend la forme '-에 :-*_e*' (31a). La postposition -*_e* est aussi appliquée aux morphèmes nominaux pour indiquer le lieu et le temps (31e).

(31)

a. 그는 도서관에 책을 주었다.
geu_neun do_sô_goan_e chaig_eul ju_ôss_da.
[geu][neun] [do_sô_koan][e] [chaig][eul] [ju_da][ôss][da].
lui+Post.nmtp bibliothèque +Post.datif livre+Post.accu donner+Mtp+St.déc.
Il a donné le livre à la bibliothèque.

b. *그는 도서관에게 책을 주었다.

c. 그 분께서(만) 당신께(만) 책을 드렸다.
geu bun_ggô_sô dang_sin_gge(man) chaig_eul deu_lyôss_da
[geu] [bun][ggô_sô] [dang_sin][gge]([man]) [chaig][eul] [deu_li_da][ôss][da].
ce NI.homme.hon+Post.nmtp.hon toi.hon+Post.datif.hon livre+Post.accu donner.hon+Mtp+St.déc.
Il vous a donné le livre.

d. 그는 너에게 책을 주고**만** 갔다.
geu _neun nô_e_ge chaig_eul ju_go_ man gass_da.
[geu][neun] [nô][e_ge] [chaig][eul] [ju_da][go][man] [ga_da][ôss][da].
lui+Post.nmtp toi+Post.datif livre+Post.accu donner+Sconj+Post.spc aller+Mtp+St.déc.
Il t'a seulement donné le livre et est parti.

e. 그는 산에 산다.

geu_neun san_e san_da.
[geu][neun] [san][e] [sal_da][eun][da].
lui+Post.nmtp montagne+Post.lieu habituer+Mtc+St.déc.
Il habite à la montagne.

La phrase (31c) est construite avec la variante honorifique de la postposition -*_e_gei* et aussi avec le verbe honorifique de 주다 : *ju_da* « donner » : 드리다 : *deu_li_da*. Elle montre les séquences possibles avec la postposition auxiliaire '-*_man*'. Dans la phrase (31d) la postposition -*_man* se soude aussi au suffixe conjonctif '-고 : -*_go* : conjonctif'.

La figure (2-16) montre les séquences qui ont la fonction nominative après racine ouverte.

La postposition de nominatif -가 : -*_ga* a la variante -이 : -*_i* après les morphèmes qui se terminent par les syllabes qui ne contiennent pas de consonne finale. La figure montre aussi les postpositions spécifiques 만 : -*_man* « seul », 도 : -*_do* « aussi », 는 :-*_neun* combinées avec le suffixe adjectif -이다 . -*_i_da* . Le tableau (2-24) montre les formes des variantes.

Le morphème 이다 . -*_i_da* « être», qui se soude aux séquences de morphèmes, a les caractéristiques des racines verbales et adjectivales qui se soudent à des séquences de suffixes verbaux et adjectivaux et qui sont des prédicats de phrases ou de propositions. Les grammairiens coréens le comptent comme postposition de cas descriptif au point de vue de la connectivité avec les morphèmes nominaux, ou comme copule avec la compatibilité de séquences des suffixes verbaux et adjectivaux [NAM 1994]. Nous le traitons avec les racines verbales et adjectivales.

La postposition vocative -야 : -*_ya* est appliquée pour s'adresser directement à quelqu'un, à quelque chose. On l'applique dans la plupart des cas aux prénoms personnels dans la conversation pour indiquer la personne. Selon le dernier élément du morphème précédent il a la variante -아 : -*_a*.

L'autre allomorphe- ㅣ가/- ㅣ : *-i_ga/-i* de la postposition de nominatif –*_ga* avec les prénoms (figure (2-4)).

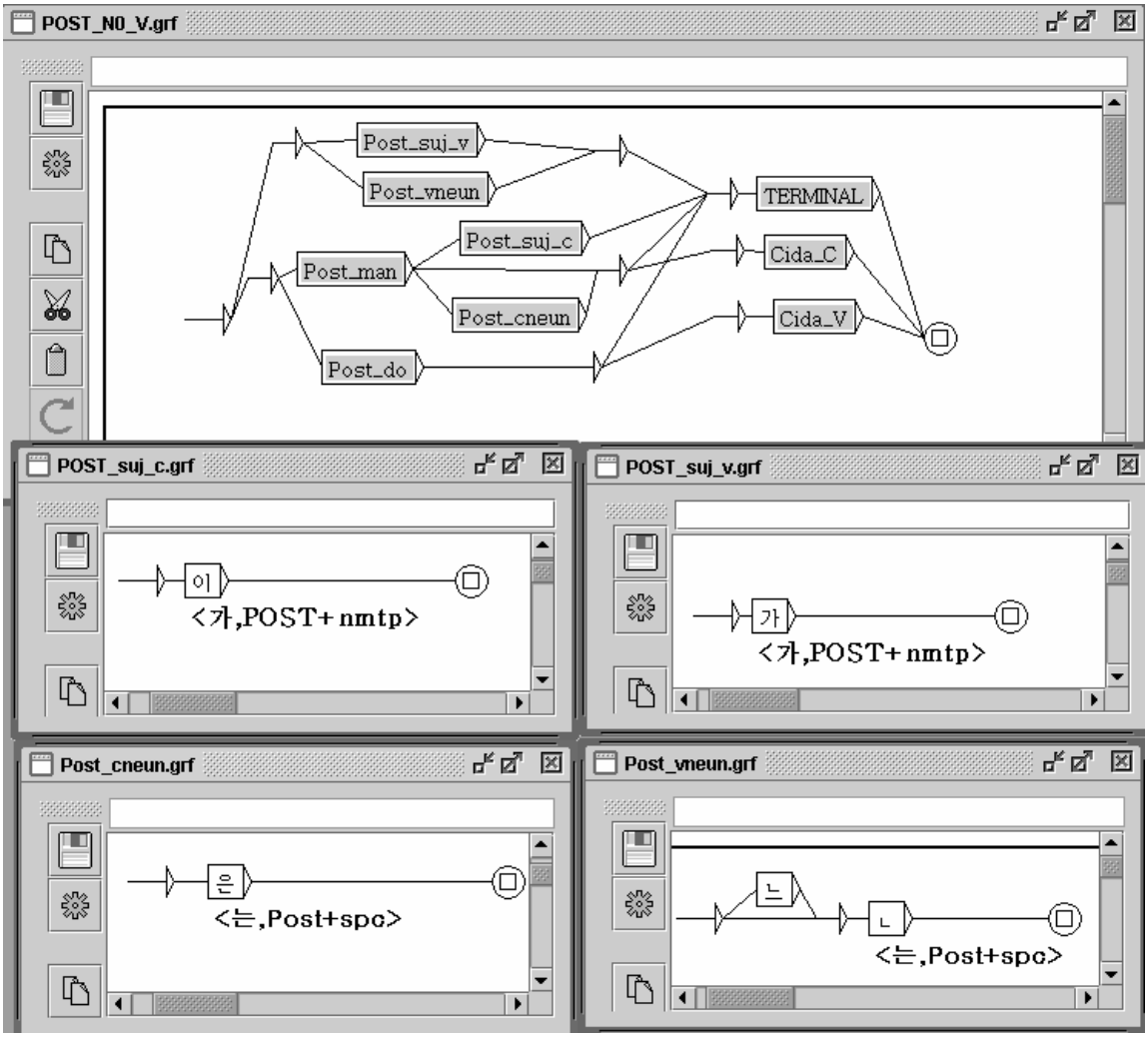


Figure 2-16 Séquences des postpositions de nominatif

Tableau 2-24 Variation après les morphèmes des noms

Forme canonique	Phonétique selon le dernier élément du morphème précédent			code
	Après Voyelle	Après la Consonne '*l*'	Après Consonnes sauf '*l*'	
가	가 : -_*ga*	이 :-_*i*	이 : -_*i*	nmtf
는	는/ㄴ : -_*neun/n*	은 : -_*eun*	은 : -_*eun*	spc
를	를/ㄹ : -_*leul/l*	을 : -_*eul*	을: -_*eul*	accu
과	와 : -_*oa*	과 : -_*goa*	과: -_*goa*	conj
야	야 : -_*ya*	아 : -_*a*	아 : -_*a*	vocatif
으로	로 : -_*lo*	로 : -_*lo*	으로 : -_*eu_lo*	dest
일랑	ㄹ : *l*	일랑/을랑 : -_*il_lang/-_eul_lang*	일랑/을랑: -_*il_lang/-_eul_lang*	conj
하고	-	-	-	conj

2.4 Mots invariables

Nous traitons ici les mots sans suffixes ou qui se combinent rarement avec les suffixes : déterminants, interjections, les mots invariables.

Déterminants

Ces mots se situent avant les noms et les autres déterminants. Dans la section 2.3.3, les trois déterminants démonstratifs sont représentatifs. Et les autres sont les déterminants adjectivaux.

Certains adjectifs épithètes n'ont pas de variantes morphologiques ni de suffixes, et ne se placent pas à la fin de la phrase ni à la fin de la proposition. Dans les exemples, la racine adjective 새롭다 : *sai_lob_da* « être nouveau » se situe avant les noms et la fin de la proposition ou la fin de la phrase (32b), le déterminant adjectif 새 : *sai* (neuf) se situe toujours avant les noms (32a).

(32)

a. 그는 새 차를 샀다.
geu_neun sai cha_leul sass_da.
[geu][neun] [sai] [cha][leul] [sa_da][ôss][da].
lui+Post.nmtp neuf.Det voiture+Post.accu acheter+Mtp+St.déc.
Il a acheté une neuve voiture.

a1. *그가 산 차는 새었다.
geu_neun san cha_leul sai_da.
lui+Post.nmtp acheter+Sd.passé nouveau+St.déc.

b. 그는 새로운 차를 샀다.
geu_neun sai_lo_un cha_leul sass_da.
[geu][neun] [sai_lob_da][eun] [cha][leul] [sa_da][ôss][da].
lui+Post.nmtp « Etre nouveau »+Sd voiture+Post.accu acheter+Mtp+St.déc.
Il a acheté une nouvelle voiture.

b1. 그가 산 차는 새롭다.
geu_ga san cha_neun sai_lob_da.
[geu][ga] [sa_da][eun] [cha][neun] [sai_lob_da][da].
lui+Post.nmtp acheter+Sd.passé voiture+Post.nmtp « être nouveau ».
La voiture qu'il a achetée est nouvelle.

Interjections

Les interjections sont des mots autonomes qui n'ont pas de fonction au sein de la

phrase ni de suffixes. Une interjection peut jouer à elle seule le rôle d'une phrase comme le mot 참 : *cham* ! « Ah, eh bien ».

Noms et verbes défectifs

Il s'agit de mots qui ont très peu de cas de combinaison avec la racine et les suffixes.

Le mot 말마따나 : *mal_ma_dda_na* « comme la parole de qqn » se compose du nom 말 : *mal* « langage, mot, langue, parole, le dire » et du suffixe -마따나 : *-_ma_dda_na* « comme » qui ne peut se souder qu'avec le nom *mal*. Ce mot suit toujours un nom au génitif comme les noms incomplets (figure 2-18).

(33) 너의 말마따나 오늘은 비가 오겠다.
nô_eui mal_ma_dda_na o_neul_eun bi_ga o_ggess_da.
[nô][eui] [mal][ma_dda_na] [o_neul][neun] [bi][ga] [o_da][gess][da].
Toi+Post.gen parole+Post.« comme » aujourd'hui+Post.spc pluie+Post.nmtp venir+Mfut+St.déc.
Comme tu l'as dit, aujourd'hui il va pleuvoir.

La figure (2-17) montre la séquence de morphèmes de ce mot. Et la figure (2-18) montre la séquence des mots dans le texte.

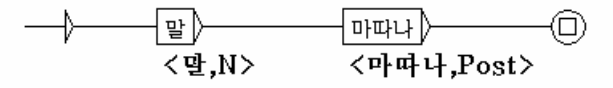


Figure 2-17 Expression du mot "*mal_ma_dda_na*"

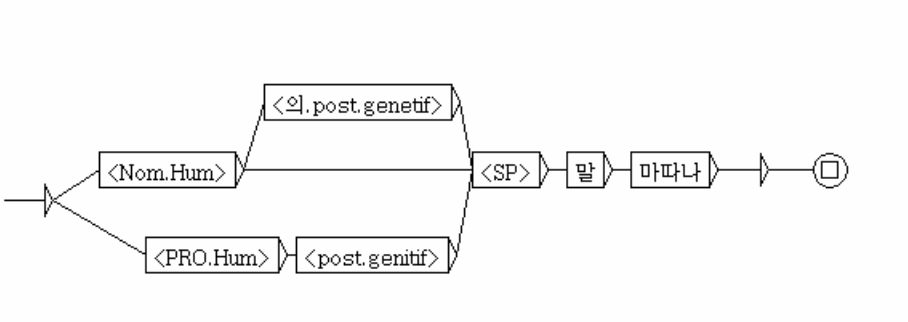


Figure 2-18 Locution *mal_ma_dda_na*

La racine verbale 관하다 : *koan_ha_da* (être en relation avec quelque chose) a peu de possibilité de combinaisons avec les suffixes verbaux : le suffixe déterminant -는 :*-_neun* (34b) et le suffixe conjonctif 어서 :*-_ô_sô* (34a). La racine verbale 대하다 : *dai_ha_da* a le sens « faire face à » mais avec la séquence de la figure (2-19), le sens de celle-ci devient le

même que le mot *koan_ha_da*.

(34)
a. 그녀는 그에 관해서 얘기했다.
geu_nyô_neun geu_e koan_hai_sô ai_geu_haiss_da.
[geu_yôn][neun] [geu][e] [kuôn_ha_da][ô_sô] [ai_geu_ha_da][ôss][da].
Elle+Post.nmtp lui+Post.« à » concerner+Sc.conj dire+Mtp+St.déc.
Elle parle à propos de lui.

b. 그녀는 그에 관한 얘기를 했다.
geu_nyô_neun geu_e koan_han ai_geu_haiss_da.
[geu_yôn][neun] [geu][e] [koan_ha_da][neun] [ai_geu][leul] [ha_da][ôss][da].
Elle+Post.nmtp lui+Post.« à » concerner+Sd parole+Post.accu faire+Mtp+St.déc.
Elle dit une parole qui le concerne

Elle parle à lui.

Nous traiterons ces séquences comme les séquences composées.

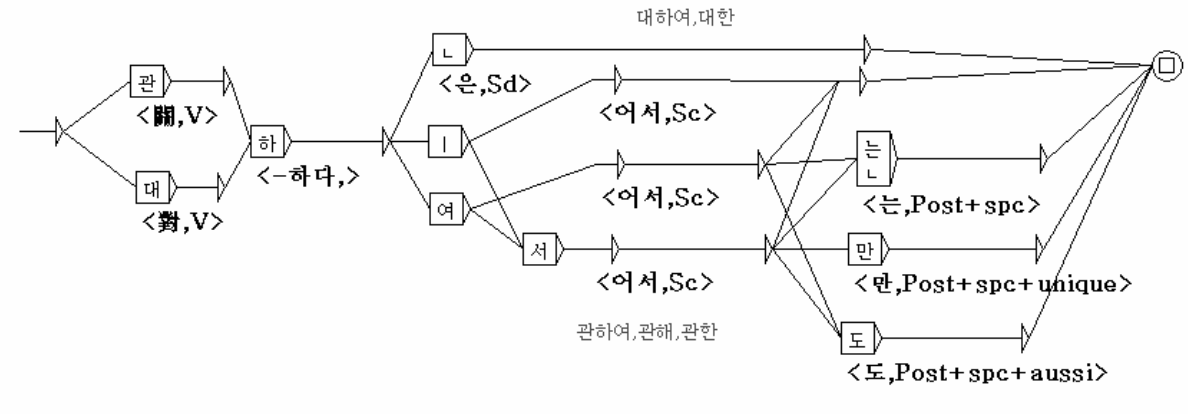


Figure 2-19 Les mots "*kuôn_ha_da*" et "*dai_h_da*"

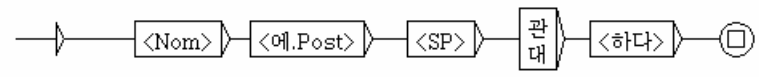


Figure 2-20 Locution de "en ce qui concerne"

TROISIEME PARTIE

3. Construction des dictionnaires des mots

En raison de certaines caractéristiques de la langue coréenne la plupart des traitements traditionnellement utilisés pour l'analyse des langues occidentales ne peuvent pas être appliqués tels quels pour l'analyse des textes coréens.

En effet le coréen est une langue agglutinante, ce qui signifie qu´un mot est formé à partir d´un radical sur lequel se fixe toute une série de particules. Les caractères sont des syllabes. Une syllabe se compose de plusieurs lettres coréennes. Comment traiter et exprimer un mot de la langue agglutinante comportant plusieurs parties ? Où définir la borne dans un mot en syllabes entre la racine et les affixes ? Comment exprimer les informations sur la partie lexicale, et la partie grammaticale ? En raison de l'existence de plusieurs systèmes de codage pour l'écriture coréenne, quel codage de système de caractères choisir ? Quelle méthode de transcodage entre syllabe et alphabet coréen, si on a besoin de traiter les mots au niveau des phonèmes coréens ?

Dans cette partie, nous présentons nos méthodes pour analyser informatiquement les mots coréens depuis le traitement des caractères coréens syllabiques jusqu'à l'analyse morphologique.

Nous présentons dans un premier temps les jeux de caractères existants et notre méthode de conversion entre syllabes et lettres alphabétiques. Ensuite, nous présentons notre méthode utilisant des descriptions formalisées dans des transducteurs à états finis pour l'analyse des séquences des morphèmes dans un mot coréen. Nous expliquerons la composition des mots en deux parties : traitement de la racine par des dictionnaires sous forme de listes et traitement des séquences des suffixes sous forme de graphes, puis production d'un dictionnaire comprimé pour le traitement automatique.

3.1 Différences entre les jeux de caractères coréens

En informatique, le jeu de caractères est l'ensemble des caractères qui sont utilisés en informatique (charset en anglais). Le jeu de caractères est la liste de caractères codés et reconnus par les systèmes. Normalement, les langues alphabétiques ont un jeu de caractères pour chaque langue. La langue coréenne a plusieurs jeux de caractères.

Pour les langues d'Europe occidentale, les textes sont généralement transcrits dans un des différents jeux de caractères ASCII étendus. Il s'agit de systèmes de codage à taille fixe d'un octet qui permettent de représenter un maximum de 256 caractères. Pour le coréen, l'unité de caractères est la syllabe qui est coupée en 3 parties : la consonne initiale, la voyelle et la consonne finale, chacune des parties étant elle-même composée de une à plusieurs lettres de l'alphabet coréen. Avec les éléments de l'alphabet coréen contemporain, on peut compter jusqu'à 11 773 syllabes[31]. Aujourd'hui, plusieurs systèmes de codage existent pour encoder les textes coréens. Le tableau 3-1 montre la valeur de caractère '가 : *ga*' dans les différents systèmes de codage.

Tableau 3-1 Comparaison des valeurs selon les jeux de caractères

	Une lettre latine 'a'	une syllabe coréenne '가'
Windows959	0x60	0xa4a1
ISO-2200-KR	0x60	0x2421
UNICODE	0x0060	0xac00
JOHAB	0x60	0x8861
KSC5601-1987	0x60	0xa4a1

Nous pouvons classifier les systèmes de codage en deux catalogues : les systèmes de codage syllabique et les systèmes de codage alphabétique selon la présentation des mots coréens dans les textes par des caractères : syllabe ou lettre de l'alphabet coréen. Les jeux de caractères syllabiques sont généralement utilisés dans les systèmes d'exploitation pour sauvegarder ou afficher les textes coréens codés syllabe par syllabe. Actuellement, il n'existe

[31] Avec toutes les lettres de l'alphabet coréen utilisées plus d'une fois, on fabrique 1.656.000 caractères syllabiques coréens.

pas de standard pour les codages alphabétiques qui représentent les textes coréens en alphabet coréen, même si certaines entreprises utilisent leur propre système en privé. Pour l'analyse morphologique du coréen, on a besoin de décomposer les syllabes en lettres alphabétiques coréennes, et donc d'utiliser un système de codage alphabétique.

3.1.1 Les jeux de caractères syllabiques

3.1.1.1 WANSUNG

Le jeu de caractères de la première version WANSUNG couvre 2 350 syllabes codées sur 2 octets. Les lettres de l'alphabet latin sont codées sur un octet. Les autres alphabets et signes sont codés en utilisant deux octets. Les syllabes coréennes sont aussi codées en deux octets. La différence entre un octet de l'alphabet latin et le premier octet d'une syllabe coréenne est la valeur d'un bit de marquage. Par exemple, un caractère 'a' est la valeur en binaire : 0b 01100000 et la syllabe '가' est codée par la valeur : 0b 10100100 10100001. Si la valeur du premier bit d'un octet est un'1', il indique que cet octet est utilisé pour le numéro de page de codage et l'octet suivant est le décalage dans la page que le premier octet a indiquée.

Le codage de ce jeu de caractères est utilisé dans les systèmes d'exploitation ou les logiciels d'éditeur de texte. Mais ce premier standard ne suffisait pas pour décrire tous les mots contemporains ainsi que les onomatopées. La première version ne couvrait pas toutes les syllabes. Les syllabes manquantes ont été ajoutées par la suite, d'une telle manière que leur position ne respectait plus l'ordre alphabétique. Ainsi il est très difficile de retrouver les lettres d'alphabet qui construisent une syllabe avec ce système, on doit utiliser un tableau qui fait la correspondance entre syllabes et séquence de lettres de l'alphabet coréen, on utilise aussi un tableau pour trier les mots selon l'ordre alphabétique.

3.1.1.2 JOHAB

Le système de codage JOHAB a été proposé comme réponse aux problèmes du système WANSUNG. Le JOHAB résout en particulier le problème du désordre des syllabes.

Puisque dans ce système, chaque syllabe coréenne est représentée par les seize bits de deux octets, le premier bit est 1 bit de marquage qui indique que ce caractère est une syllabe coréenne, les quinze autres bits sont découpés en 3 parties de cinq bits, chacun représente le groupe de consonnes initiales, le groupe de voyelles, le groupe de consonnes finales. Ce codage respecte l'ordre alphabétique. Ainsi le tableau 3.2 montre le codage de la syllabe '가 : *ga*' qui est codée en JOHAB par 0x8861.

Tableau 3-2 Signification des bits d'une syllabe dans le JOHAB

<table>
<tr><td>1</td><td>00010</td><td>00011</td><td>00001</td></tr>
<tr><td rowspan="3"></td><td rowspan="3"></td><td rowspan="2"></td><td>manque de consonne finale : <E></td></tr>
<tr><td>première de voyelle : ' ㅏ '</td></tr>
<tr><td colspan="2">première consonne de consonne initiale 'ㄱ'</td></tr>
<tr><td colspan="4">marque de bit pour les caractères de syllabes coréennes</td></tr>
</table>

Le JOHAB n'est pas un standard agréé mondialement, cependant, ce codage est utilisé par un logiciel d'éditeur de textes qui a été adopté par le gouvernement coréen pour les communications officielles.

3.1.1.3 UNICODE

Le standard UNICODE est un système de codage de caractères conçu pour l'échange et l'affichage de textes écrits dans les différentes langues du monde moderne. Il est distinct des autres systèmes que nous avons décrits précédemment par le fait que la taille d'un code du codage est fixée sur deux ou quatre octets. Les caractères syllabiques coréens sont codés selon l'ordre alphabétique comme JOHAB. De plus il permet la représentation de plus d'idéogrammes que WANSUNG pour la description des mots sino-coréen. Dans le WANSUNG, on a 4888 idéogrammes, UNICODE a près de vingt mille idéogrammes. Le tableau ci-dessous montre les nombres des caractères UNICODE et WANSUNG.

Tableau 3-3 Comparaison de contenu de lettres entre UNICODE et WANSUNG

Nom de gamme	Nombre de lettres	
	UNICODE	WANSUNG
Latin-1 Supplément	127	33
Latin Extended-A	129	18
Grec	109	48
Cyrillique	235	66
Hiragana	90	83
Katakana	95	86
Hangul Compatibility Jamo	94	94
CJK Unified Ideographs	21 000	4 620
Hangul Syllables	11 172	11 172
CJK Compatibility Ideographs	301	268
Other	29 267	878
TOTAL	49 617	17366

Il existe deux zones, *Hangul compatibility Jamo*[32], *Hangul Jamo*[33], pour l'alphabet coréen dans le système UNICODE. La zone *Hangul Jamo* existe seulement dans UNICODE. Chaque code *Hangul Jamo* représente soit un groupe de consonnes initiales (ex. ㄱ : *g* : 0x1100, ㅆ :*ss* 0x110A), soit un groupe de voyelles (ex. ㅏ :*a* :0x1161, ㅞ :*uôi* :0x1170), soit un groupe de consonnes finales(ex. ㄱ :*g* : 0x11A8, ㅆ :*ss* 0x11BB). La zone *Hangul compatibility Jamo* existe aussi dans le système de codage WANSUNG. L'ordre des caractères est identique entre les deux systèmes de codage. Chaque code *Hangul compatibility Jamo* est soit un groupe de consonnes (ex. ㄱ: *g* : 0x3131, ㅆ :*ss* 0x3146) soit un groupe de voyelles (ex. ㅏ :*a* :0x314F, ㅞ :*uôi* :0x315E) sans distinguer entre consonnes initiales et consonnes finales. La zone *Hangul Jamo* contient donc plus de lettres que *Hangul compatibility Jamo.*

Quand dans les textes codés par UNICODE, on trouve la lettre d'alphabet coréen isolée, la valeur de cette lettre est une valeur de la zone *Hangul Compatibility Jamo*.

Nous considérons les lettres de l'alphabet coréen qui existent dans le texte comme des symboles.

[32] L'annexe 2 à droite
[33] L'annexe 2 à gauche

Nous utilisons les lettres de *Hangul Jamo* pour représenter les syllabes en lettres alphabétiques coréennes et pour distinguer les caractères en lettres de *Hangul Compatibility Jamo*.

Les idéogrammes sur WANSUNG sont au nombre de 4888 et ils se trouvent dans les deux zones *CJK Unified Ideographs* et *CJK Compatibility Ideographs* dans l'UNICODE.

La correspondance entre idéogrammes et syllabes graphiques coréennes est complexe.

(1)
a) 이 (*_i* : 貳 : deux)

b) 이 (*_i* : 耳 : oreille)

c) 이 (*_i* : 李 : nom de famille)

Du point de vue de l'écriture coréenne, a, b et c sont homophones et homographes.

Tableau 3-4 Un idéogramme homographique

	coréen	sens	UNICODE	WANSUNG
龜	구 *gu*	tortue	U+9F9C	0xCFCF
龜	귀 *gui*	nom de pays	U+F907	0xD0A2
龜	균 *gyun*	lézarde	U+F908	0xD0B8

Dans le tableau (3-4), l'idéogramme polysémique '龜' est hétérophone. En informatique chaque idéogramme homographique a des valeurs différentes. Dans ce tableau, en UNICODE, l'idéogramme associé au mot *gu* « torture » existe dans la zone *CJK Unified Ideographs*. Les autres existent dans la zone *CJK Compatibility Ideographs* qui contient un complément d'idéogrammes utilisés en Chine, au Japon, en Corée.

De plus, en Corée du sud, il y a un phénomène de différence de prononciation, dans certaines conditions, la consonne initiale du mot disparaît. On l'appelle 두음법칙 : *du_eum_bôb_chig* « règle de la voix de début». Dans le tableau (3-5), le caractère '流' correspond à deux syllabes '유 : *_yu*' en début de mot et '류 : *_lyu*' dans les autres positions.

Tableau 3-5 Variantes par *du_eum_bôb_chig*

	Sino-coréen	coréen	Sens
流	物流	물류 : *mul_lyu*	logistique
	流速	유속 : *yu_sok*	vitesse du courant d'une rivière ou de la mer
	流俗	유속 :*yu_sok*	coutume

Mais certains noms de famille ne respectent pas la règle *du_eum_bob_chig*, comme le nom de famille '柳' : '유 : *_yu*' et '류 : *_lyu*'. Nous traitons les noms de famille associés comme entrées différentes et les décrivons dans le dictionnaire en deux entrées différentes.

유,柳,N+NF,Appellation
류,柳,N+NF,Appellation

La grammaire de la Corée du nord n'adopte pas cette règle.

Nous traitons le texte avec UNICODE. C'est un standard mondial. L'ordre des syllabes est l'ordre de l'alphabet coréen. On n'a pas besoin d'un tableau pour convertir les syllabes et les trier selon l'ordre de l'alphabet coréen. Pour les mots sino-coréens, on peut les décrire avec des idéogrammes plus nombreux que sur le WANSUNG.

3.1.2 Codages alphabétiques

Ces codages ne sont utilisés que pour représenter ou pour analyser les mots coréens en phonèmes. Comme nous l'avons vu, les caractères syllabiques coréens ne sont pas des lettres de l'alphabet. Ce ne sont pas des idéogrammes qui peuvent représenter un concept. Un caractère syllabique coréen représente une syllabe. Cependant le début d'une syllabe peut faire partie d'un morphème, et la fin de la syllabe faire partie d'un autre. Dans ce cas, on a besoin d'exprimer une syllabe de l'alphabet coréen par les phonèmes. Par exemple, le mot qui se termine par une voyelle se soude avec le suffixe qui commence par une syllabe sans consonne initiale. Dans plusieurs cas, il existe une contraction ou une élision ou la réduction et la constitution de voyelle composée ou de demi voyelle.

Selon la taille des éléments qui expriment une syllabe en octet, nous détaillerons trois systèmes de codage alphabétique.

Codage sur 2 octets

Le codage JOHAB est un cas représentatif pour présenter une syllabe coréenne de l'alphabet coréen. Un code peut être utilisé directement pour exprimer une syllabe et aussi peut servir pour extraire les lettres de l'alphabet coréen qui sont codées sur 5 bits.

Codage sur 3 octets

On peut vouloir traiter les syllabes qui se composent avec l'alphabet coréen ancien. Deux octets ne sont pas suffisants pour représenter toutes les syllabes anciennes. On utilise alors un octet pour les consonnes initiales, un pour les voyelles et un pour les consonnes finales.

Codage sur n octets

Pour l'analyse morphologique du coréen, on doit noter en détail les consonnes et les voyelles. On représente une syllabe avec la séquence des éléments fondamentaux de l'alphabet coréen.

Tous les jeux de caractères alphabétiques peuvent être différents selon le but de la recherche ou l'option linguistique. Normalement, les utilisateurs ne connaissent pas les jeux de caractères alphabétiques.

3.1.3 Transcodage entre syllabes et lettres d'alphabet coréen

Nous traitons tous les mots coréens au niveau des morphèmes de l'alphabet coréen. Pour la description explicite sans restriction d'utilisation des caractères coréens, nous donnons les moyens de conversion entre syllabe et lettres d'alphabet coréen aux utilisateurs de traitement du texte coréen avec le système UNITEX. Une consonne finale 'ㄳ' est traitée en informatique comme une consonne qui est le troisième élément (consonnes finales) mais les personnes ayant étudié la phonologie coréenne considèrent cette consonne comme une consonne composée d'un phonème 'ㄱ : [g]' et d'un phonème 'ㅅ : [s]'.

Pour la flexibilité d'utilisation et d'application des logiciels de conversion entre syllabes et lettres alphabétiques coréennes, nous proposons à l'utilisateur une interface pour les transcodages. Cette interface est gérée par l'utilisateur grâce à un graphe utilisé pour transcoder les chaînes de lettres coréennes en syllabes, et à un tableau défini par l'utilisateur, utilisé pour transcoder les syllabes vers l'alphabet coréen.

Nous avons créé deux programmes pour la conversion entre syllabes et lettres de l'alphabet coréen. Chaque programme utilise un fichier de données. La figure (3-1) montre l'entrée de programme et les programmes. Le programme « Syl2jamo » change les syllabes vers les lettres de l'alphabet coréen avec un tableau de conversion. Le programme « Jamo2syl » convertit les textes coréens en lettres de l'alphabet coréen vers les syllabes avec un transducteur qui produit les syllabes des chaînes de lettres de l'alphabet coréen.

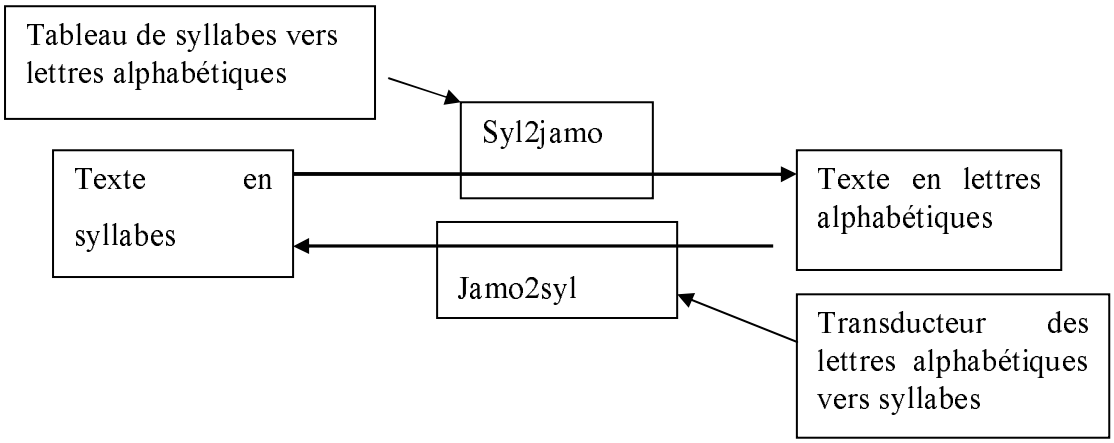


Figure 3-1 Transcodage entre syllabes et alphabet coréen

3.1.3.1 Conversion en syllabes vers les lettres alphabétiques

Pour les syllabes codées par UNICODE, on peut obtenir les éléments d'une syllabe par le calcul en trois parties : la consonne initiale, la voyelle, la consonne finale. D'abord, on peut les obtenir par le calcul :

Consonne initiale : ([syl] – 0xac00) / (21 * 28)
Voyelle : (([syl] – 0xac00) / 28) modulo 21
Consonne finale : ([syl] – 0xac00) modulo 28

Le « [syl] » est la valeur d'une syllabe coréenne codée en UNICODE. La valeur « 0xac00 » est la valeur initiale de la gamme « Hangul Syllabes ».

A partir du code de chaque partie, on obtient les éléments alphabétiques en ajoutant la valeur du code à la valeur initiale de chaque partie de zone « Hangul Jamo » d'UNICODE. Mais si on veut les éléments de la zone « Hangul compatibility Jamo », on a besoin d'un tableau de conversion entre l'ordre de chaque partie et l'élément qui lui est associé. De plus, si on veut les éléments en phonèmes fondamentaux, on doit encore décomposer chaque partie en éléments fondamentaux.

Alors nous donnons un tableau (annexe A31) pour gérer la décomposition des syllabes. Le fichier se décompose en 4 parties : une partie pour définir les caractères spéciaux et trois parties pour les éléments d'une syllabe. Dans la première partie, on peut définir les caractères spéciaux par la chaîne des caractères entre les crochets angles « <> » comme la marque de limite de syllabes qui n'existe pas en réalité. Nous avons utilisé la chaîne de caractères <SS> pour la marque de limite de syllabes.

Chaque partie commence par la ligne de nom de la partie. Les lignes dans chaque partie sont triées selon l'ordre de la consonne initiale, la voyelle, la consonne finale d'une syllabe dans la zone *Hangul Jamo* d'UNICODE.

Chaque autre partie se décompose en 4 parties en colonnes. La première colonne donne la valeur des caractères dans la zone *Hangul Jamo*.

La deuxième colonne est la valeur des caractères dans la zone *Hangul compatibility Jamo*. Pour la description des formes fléchies des morphèmes, on tape les lettres alphabétiques qui sont codées par la zone *Hangul compatibility Jamo* avec l'éditeur de texte,

puis on doit les transcoder en *Hangul Jamo*.

Nous traitons les chaînes de caractères après le symbole '#' comme commentaire.

Le tableau de l'annexe A31 est utilisé pour transcoder les syllabes vers les caractères alphabétiques. Le tableau de l'annexe A32 est le fichier pour la translittération de syllabes en alphabet latin.

3.1.3.2 Conversion des lettres alphabétiques en syllabes

Pour le transcodage des lettres alphabétiques en syllabes, nous donnons un transducteur. La sortie de ce transducteur comporte les actions. Le transducteur émet une valeur qui est calculée par les expressions. Nous représentons par une expression régulière le format des sorties :

EXP := <**DEP_ELE**> | <**DEP_ELE**>,<**EXP**>
DEP_ELE := <**Defi_controle**> | <**VAR**><**ASSIGN**><**EP_CALCUL**>
EP_CALCUL := <**VAR**>|<**VAR**> <**OP**> <**VAR**>
OP := + | - | * | /
VAR :=<**PRE_VAR**> | <**USER_VAR**>
PRE_VAR := %ret%
USER_VAR=%<**STR**>%
<**STR**> := (a | …| z |) <**STR**>/<E>
ASSIGN := '='
Defi_controle := **COUT** | **CINIT** | **FINI**

Par exemple, un transducteur de décodage produit la syllabe '간' à partir d'une chaîne de caractères en alphabet coréen « _ㄱㅏㄴ ».

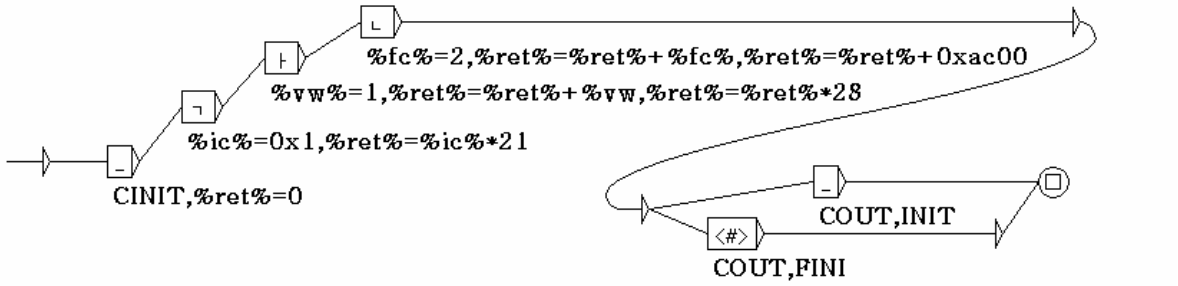


Figure 3-2 Transcodage d'une chaîne de caractères de l'alphabet coréen en une syllabe

Le parcours de ce transducteur commence à reconnaître le symbole de marque de syllabe « _ » et chaque état arrivé fait l'exécution des commandes de la ligne de sortie. L'ordre d'exécution des commandes est de gauche à droite. La sortie d'état de ce transducteur est des actions de calcul et assignement pour obtenir une valeur.

Le mot « COUT » est une commande qui émet en sortie la valeur de la variable « %ret% ». Le mot « CINIT » fait retourner le parcours à l'état initial du transducteur. Le mot « FINI » arrête le parcours du transducteur.

Nous montrons notre transducteur de décodage des caractères alphabétiques vers des syllabes à l'annexe 3.

Le graphe ci-dessous montre une autre application avec ce transducteur. Il est utilisé pour reconnaître le dernier élément des morphèmes nominaux coréens. Il émet en sortie une des valeurs suivantes : V (voyelle), C (consonne sauf 'l'), L (consonne 'l') selon le dernier élément de morphèmes nominaux. Ces valeurs vont être utilisées pour les variations phonétiques des postpositions.

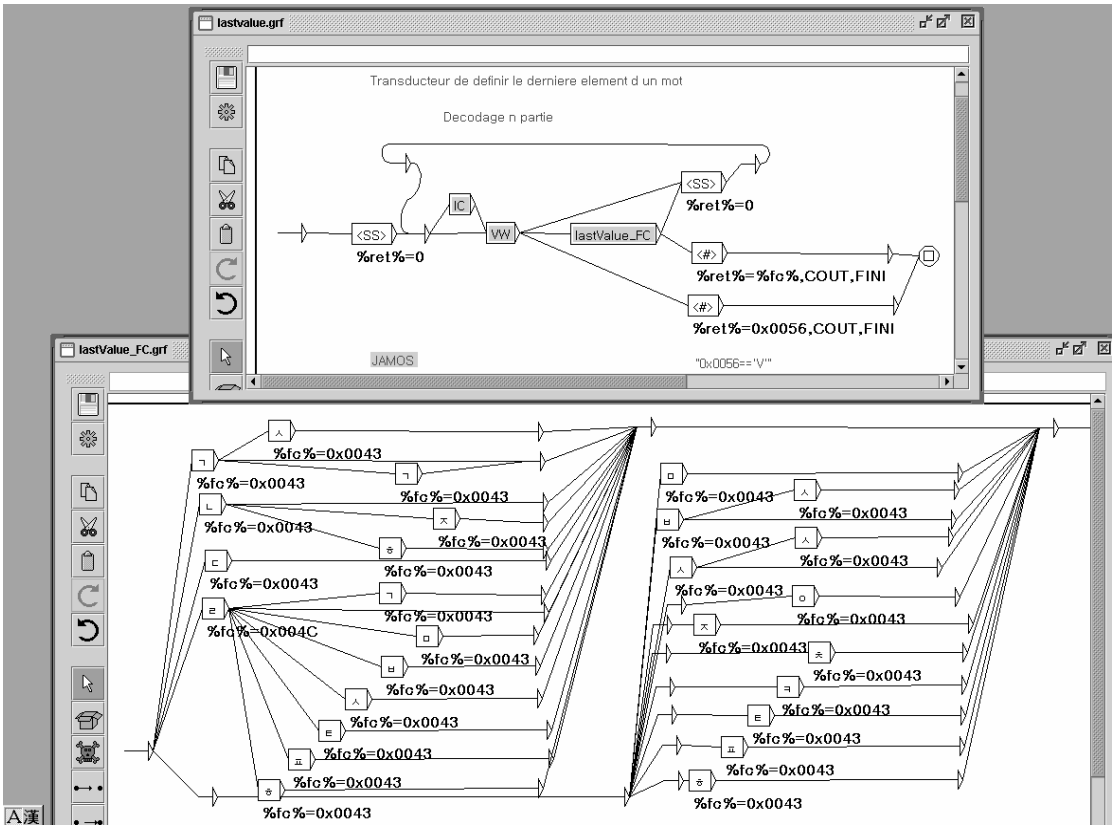


Figure 3-3 Transducteur qui connaît le dernier élément

3.2 Description des séquences des morphèmes d'un mot

En général, les dictionnaires contiennent les mots sous forme canonique avec les informations linguistiques sémantiques, grammaticales, flexionnelles et syntaxiques. Les mots dans le texte sont individuellement sépares par les séparateurs. Ces mots sont des mots fléchis. Pour reconnaître les mots fléchis et pour l'efficacité du traitement automatique, on utilise les dictionnaires sous forme de compression. Pour la langue française, les dictionnaires électroniques sont construits à partir du dictionnaire qui contient les mots sous forme canonique. Dans l'UNITEX, la figure (3-4) montre la construction du dictionnaire électronique à utiliser pour reconnaître des mots dans le texte.

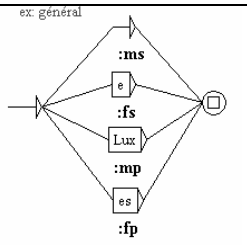
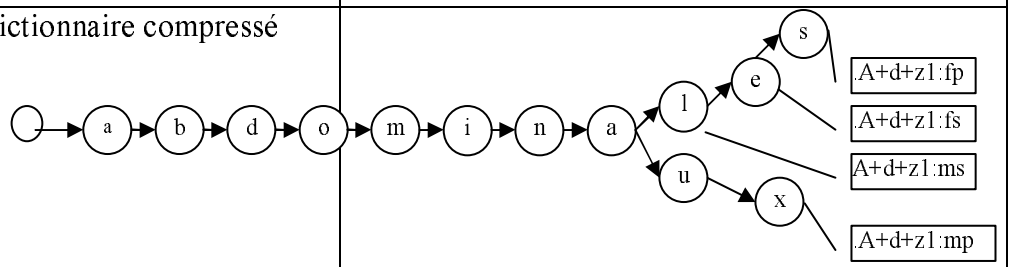
dictionnaire	format d'entrées
DELAS	abdominal,.A76+d+z1
	 transducteur de flexion
DELAF	abdominal,.A+d+z1:ms abdominale,abdominal.A+d+z1:fs abdominales,abdominal.A+d+z1:fp abdominaux,abdominal.A+d+z1:mp
dictionnaire compressé	

Figure 3-4 Construction de dictionnaire électronique français

Les linguistes éditent d'abord le dictionnaire des entrées simples : DELAS. Après, on utilise le transducteur de flexion pour obtenir les formes fléchies de chaque lexique. DELAF est le dictionnaire qui contient les formes fléchies. On peut obtenir le dictionnaire compressé sous forme d'automate. Chaque état terminal contient les informations linguistiques.

Nous allons expliquer la description des séquences des morphèmes d'un mot coréen avec le transducteur.

La figure (3-5) montre les séquences de morphèmes avec une racine 작다 : *jak_da* « être petit(e) » et le suffixe -으시 :-*_eu_si* (honorifique de sujet), le suffixe -_었 : -*_ôss* (temps passé) et le suffixe -다 : -*_da* (suffixe terminal déclaratif).

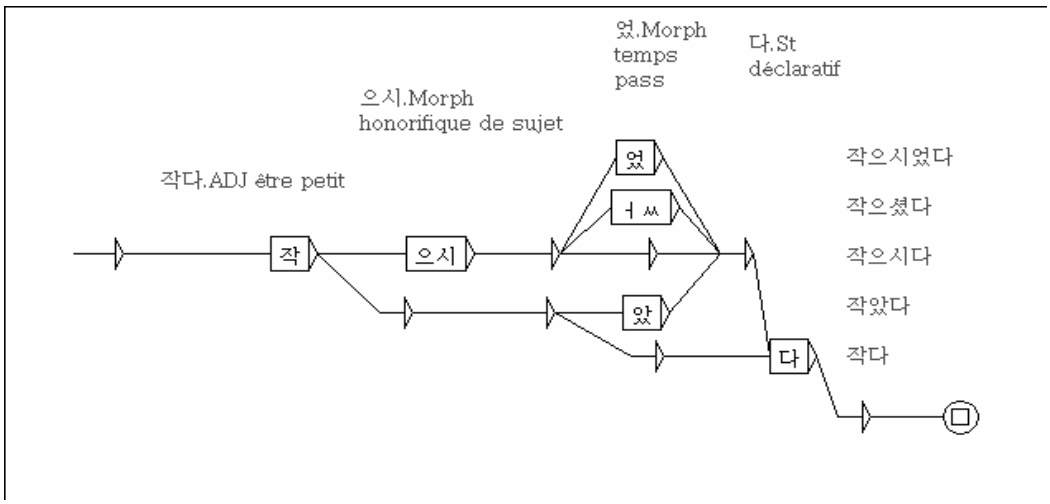


Figure 3-5 Séquences des morphèmes avec la racine "*jak_da*"

Les boites qui expriment les états contiennent les morphèmes.

La forme *jak_da* est l'infinitif (suffixe -*_da*). Nous allons expliquer les séquences des morphèmes selon la variation phonétique. La forme de racine dans le texte est sans -*_da* (le suffixe infinitif). Cette racine a les caractéristiques phonétiques : l'harmonie de voyelle Yang, terminé par une consonne. Le suffixe -*_eu_si* se situe à la suite de la racine verbale et adjectivale. Le suffixe -*_ôss* se situe après la racine ou le suffixe, il a plusieurs variantes selon le type de racine. La figure (3-6) exprime les compatibilités entre les types de racine et les variantes d'une voyelle 어 : *_ô*.

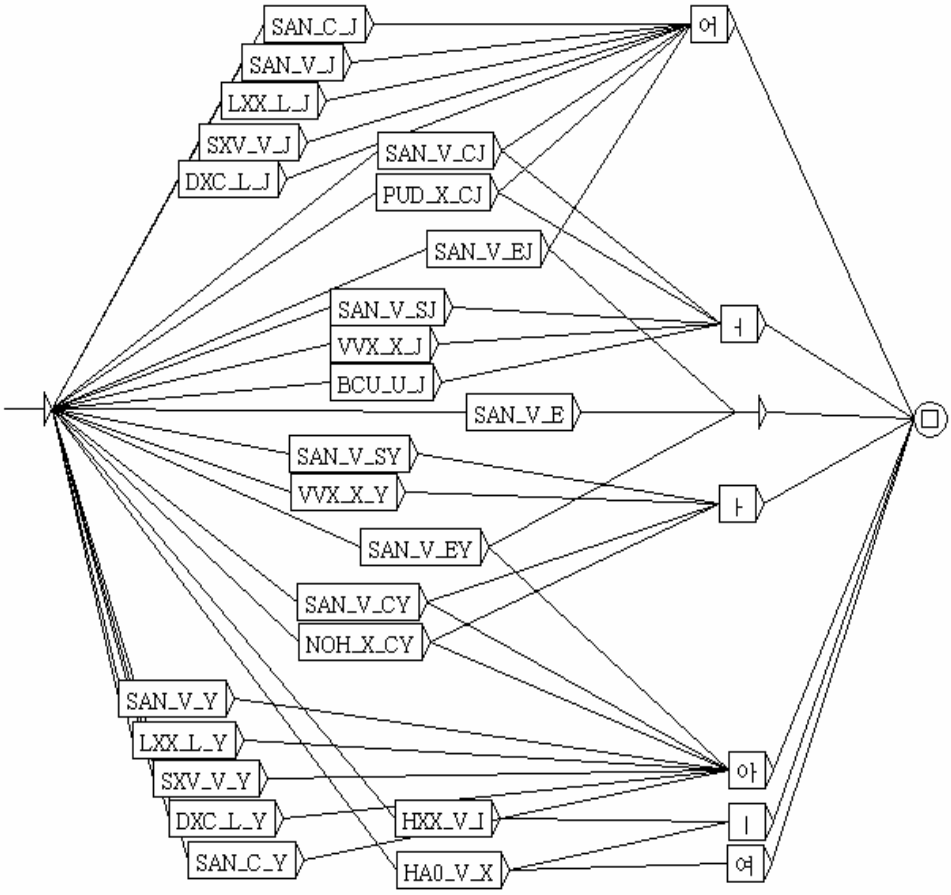


Figure 3-6 Variantes de la voyelle "*_ô*"

Le suffixe -었 : -*_ôss* a deux variante -었 : -*_ôss* et ㅕ씨 : -*ôss* après le suffixe 으시 : -*_eu_si* dans la figure (3-7). « ㅕ씨 » est une chaîne de caractères alphabétiques coréens qui représente une syllabe sans consonne initiale. Le suffixe 다 : -*_da* n'a pas de variante phonétique. Les deux suffixes -*_eu_si* et -*_ôss* doivent exister entre la racine et les suffixes terminaux ou conjonctifs.

La racine 주다 : *ju_da* « donner » est ouverte phonétiquement et elle a la caractéristique de l'harmonie de la voyelle Yin.

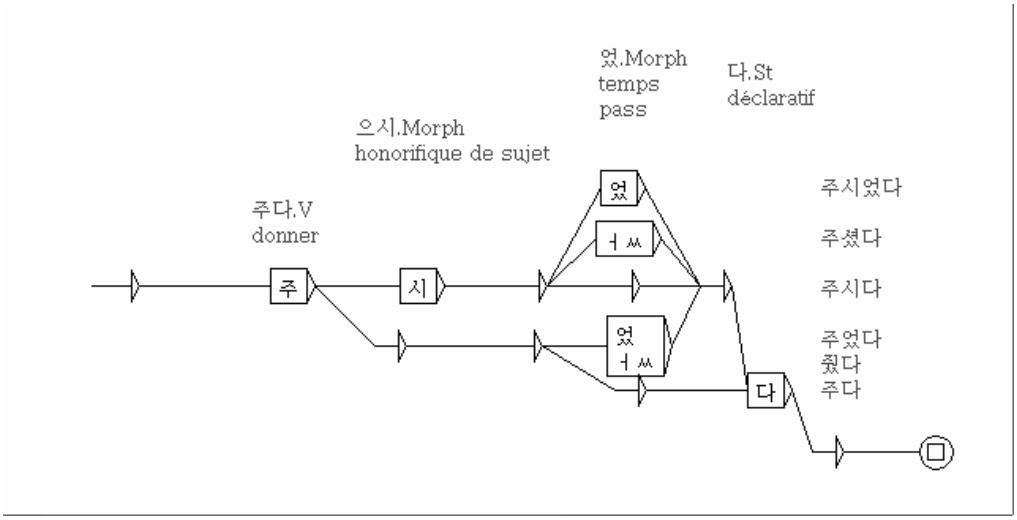


Figure 3-7 Séquence des morphèmes avec la racine *ju_da*

Dans la figure (3-8) la racine 크다 :*keu_da* « être grand(e) » a des variantes.

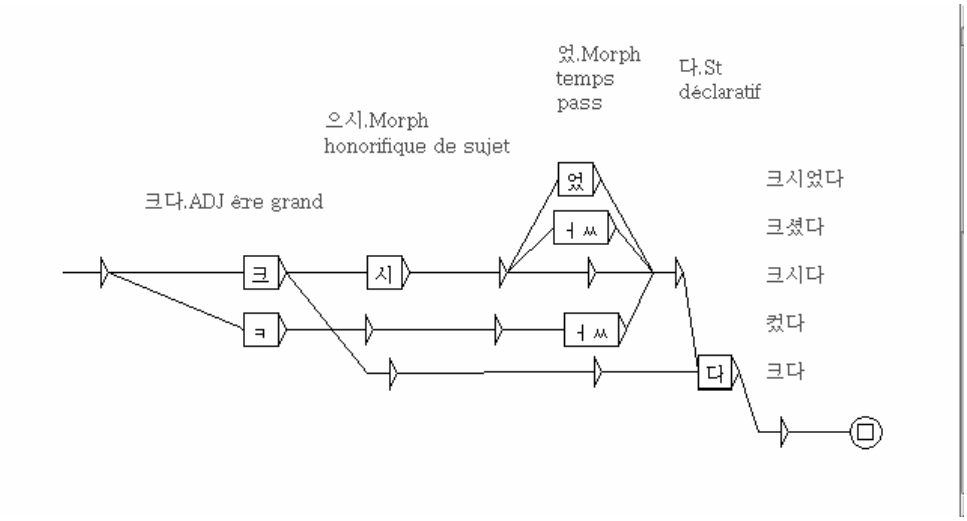


Figure 3-8 Séquences des morphèmes avec la racine *keu_da*

En fait, les racines sont écrites sous forme de liste. Et on exprime les séquences des morphèmes grammaticaux sous forme de transducteur.

Nous exprimons les séquences de morphèmes des exemples avec les racines et les graphes ci-dessous sous forme de transducteur.

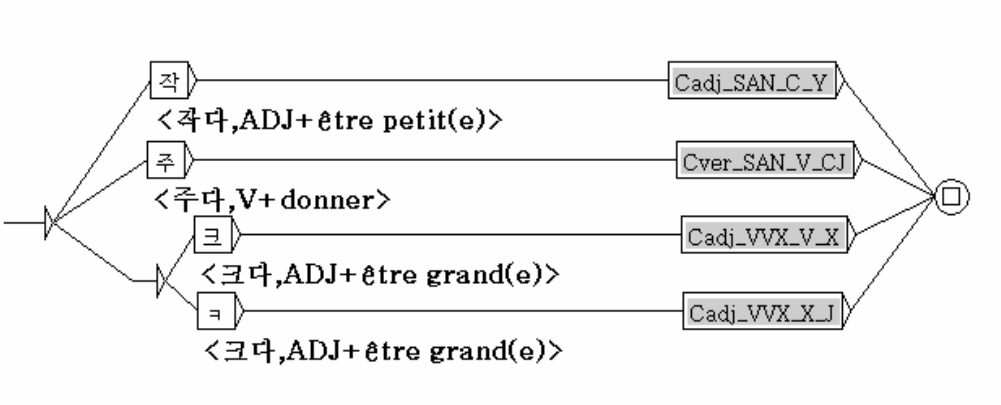


Figure 3-9 Racines avec les sous-graphes sous forme de transducteur

Les informations de racine sont exprimées par la sortie avec les symboles « < », « > ». La partie de sortie est utilisée pour la forme canonique et les informations linguistiques en séparant par le symbole ‘,’.

Les boîtes en gris expriment les sous-graphes qui contiennent toutes les séquences de suffixes avec toutes les possibilités de combinaison de morphèmes qui peuvent se souder à la racine.

En fait, nous avons mis les informations sous la forme d’une liste :

```
;작,,작다,ADJ+être petit,Cadj_SAN_C_Y
;주,,주다,V+donner,Cadj_SAN_V_CJ
;크,,크다,ADJ+être grand(e),Cadj_VVX_V_X
;ㅋ,,크다,ADJ+être grand(e),Cadj_VVX_X_J
```

Cette forme est presque pareille à la forme de DELAF, sauf le champ pour exprimer la compatibilité de racine et de séquence de suffixe qui est identifiée sous le nom de sous-graphe.

Les séquences des suffixes sont exprimées selon la figure suivante.

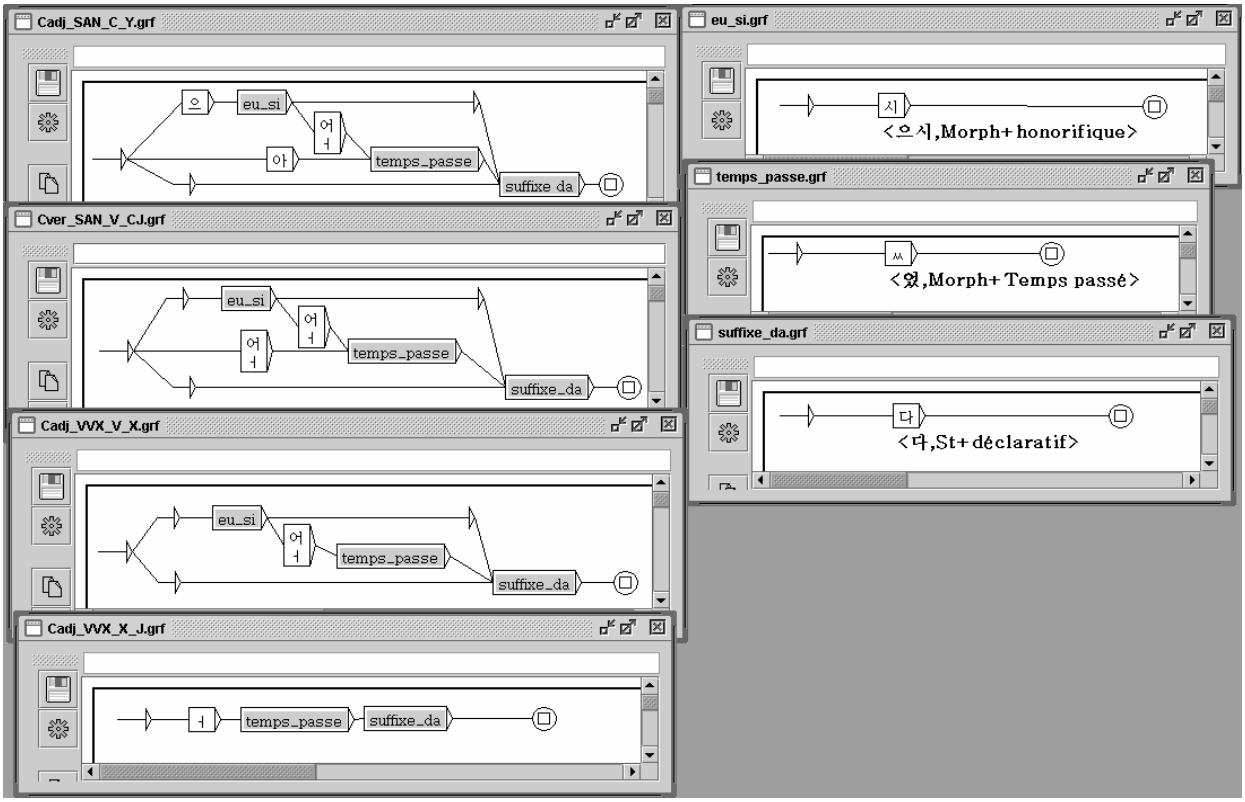


Figure 3-10 Séquences des suffixes avec les sous-graphes

Chaque morphème a ses informations linguistiques comme les racines des exemples.

Ces graphes des séquences de suffixes permettent de produire une liste. La liste suivante montre les séquences de suffixes. « Cadj_SAN_C_Y » pour une racine *jak_da*.

;으시,,으시,Morph+honorifique,;ㅓㅅㅅ,,었,Morph+Temps passé,;다,,다,St+déclaratif,
;으시,,으시,Morph+honorifique,;었,,었,Morph+Temps passé,;다,,다,St+déclaratif,
;으시,,으시,Morph+honorifique,;다,,다,St+déclaratif,
;었,,었,Morph+Temps passé,;다,,다,St+déclaratif,
;다,,다,St+déclaratif,

Nous allons expliquer en détail le traitement de la partie racine et de la partie suffixes aux sections suivantes pour la construction de dictionnaire électronique.

3.3 Construction des dictionnaires

Nous avons montré la description des séquences des morphèmes des mots coréens avec le graphe en transducteur. L'avantage de la description en graphe pour les séquences des morphèmes est qu'on peut décrire et voir facilement les relations autour d'un morphème.

Nous traitons la partie de la racine sous forme de liste et la partie de la séquence des suffixes sous forme de graphe qui représente le transducteur. Après avoir obtenu chaque partie, nous réunissons les deux parties pour former les mots.

Nous avons intégré les fonctions du traitement des dictionnaires dans UNITEX. Les dossiers cités existent sous le système d'UNITEX.

3.3.1 Partie de la racine

Au cours du traitement des dictionnaires de la partie « racine », nous utilisons trois types de dictionnaires avant d'obtenir le dictionnaire sous forme d'automates comprimés qui sont utilisés pour la consultation des mots. : Dictionnaires des racines de base, Dictionnaires des racines avec dérivations, Dictionnaires des racines avec variantes.

3.3.1.1 Dictionnaires des racines de base

Nous commençons par traiter la partie de racine avec les dictionnaires des racines de base. Les dictionnaires sont construits à la main sous forme de listes.

La ligne des dictionnaires des racines de base est divisée en quatre champs par le symbole ','. Le format dans le dictionnaire coréen est : Racine, Racine originelle, Informations linguistiques, Informations sur suffixes flexionnels.

(2)

춥다,,A,BCUVAR
흐르다,,VS,LLEVAR
발전,發展,N+/dPRED2+/dPRED3,NC
사랑스럽다,,A,Cadj_SEU_LEB
가공적이다,,A,Cadj_JEK

Le champ « Racine » est le mot coréen sous forme de séquence de syllabes.

Dans la langue coréenne, d'après le dictionnaire coréen *KEUN_ SA_ JôN*, parmi les 140 464 entrées du coréen standard, 56 115 (40%) sont des mots purement coréens, 81 363 (57%) sont des mots sino-coréens et 2 987 (2%) sont des mots étrangers.

Si le mot racine est emprunté, on renseigne le champ *Racine originelle*. Et on peut aussi considérer la forme originelle comme forme canonique du point de vue de la variation de l'écriture. Si on veut trouver un mot '發展[34]'dans le texte, on doit automatiquement chercher aussi 발전 :*bal_jôn* qui est la forme de ce mot en écriture coréenne. Les noms de famille coréens sont des mots sino-coréens. Il existe 274 noms de famille : 262 en une syllabe, 12 en deux syllabes. Le prénom varie d'une syllabe à trois syllabes. Dans les journaux, on peut voir les noms personnels écrits en idéogrammes sans juxtaposition avec la forme en écriture coréenne.

Nous traitons tous les idéogrammes en convertissant les syllabes coréennes. Nous traitons 4 888 idéogrammes qui existent dans le système de codage WANSUNG, bien qu'il y ait environ 40 000 idéogrammes dans UNICODE.

Le champ des informations linguistiques est utilisé pour représenter les informations sur la racine. Chaque élément dans ce champ est distingué par un symbole '+'. Les éléments qui commencent par la séquence « +/d » indiquent la forme de dérivation de cette racine.

La séquence suivant la séquence « +/d » est un nom de transducteur de dérivation qui nous donne des mots dérivés avec la racine. Les transducteurs des dérivations doivent être situés dans le dossier « Korean /derivation ».

Pour le code de dérivation « +/dPRED2», la figure (3-11) montre le transducteur « PRED2 » qui produit un verbe dérivé avec le suffixe -*_ha_da* .

[34]발전 *bal_jôn :*développement

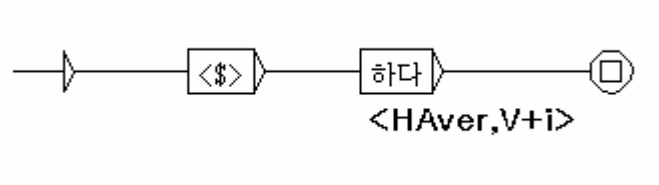


Figure 3-11 Un transducteur de dérivation avec suffixe verbal

La séquence de symboles «< $ >» montre une action du transducteur qui est la copie de la racine. Les suffixes soudés à la racine dépendent de la racine. Nous donnons un champ « Information sur suffixes flexionnels » pour présenter la compatibilité. Nous avons construit 8 sous graphes de transducteurs de dérivation des racines.

3.3.1.2 Dictionnaires des racines avec dérivations

Après le traitement de la dérivation, on obtient un dictionnaire avec dérivations. Les lignes des dictionnaires des racines avec dérivations sont identiques à celles des dictionnaires de base sans symbole de dérivation : « /d ». En plus, les racines dérivées sont présentes.

(3)
춥다,ADJ,Cadj_BCUVAR
흐르다,,VS+INT,LLEVARV
발전,,N+PRED1+PRED2,NC
발전하다,,VS+INT,HADAV1
발전되다,,VS+INT,Cver_SAN_C_Y
사랑스럽다,,A,Cadj_SEU_LEB
가공적이다,,A,Cadj_JEK

A cette étape, on utilise le champ « Informations sur suffixes flexionnels » pour obtenir les variantes. Il indique le nom du transducteur qui produit les formes de surface de la racine. Ce transducteur produit également les informations sur les suffixes compatibles avec chaque variante.

Les transducteurs sont situés dans le dossier « Korean /variation ». S'il n'y a pas de fichier qui ait ce nom dans le dossier « Korean /variation », on ne crée pas de variante. S'il

existe un transducteur qui a le même nom que ce champ, on crée les variantes de la racine.

Par exemple, le graphe « Cadj_BCUVARJ.grf » montre le transducteur qui crée la forme canonique et la variante. Nous avons construit 71 sous graphes de transducteurs de variation des racines.

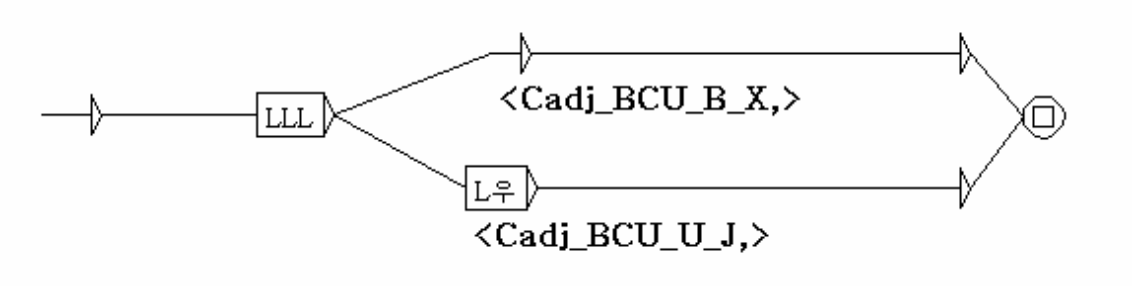


Figure 3-12 Transducteur de forme variante « Cadj_BCUVARJ »

Nous utilisons les fonctions de flexion d'UNITEX. La manipulation des caractères se situe au niveau de l'alphabet coréen. Nous ajoutons une fonction qui change une information sur les séquences des suffixes suivants. Les informations entre les symboles '<', '>' indiquent l'ensemble des suffixes suivants.

3.3.1.3 Dictionnaires des racines avec variantes

Après le traitement des champs « ***Informations sur suffixes flexionnels*** », on obtient les formes fléchies.

Format de la ligne :

[; Variante, Canonique, Origine, Informations linguistiques, Compatibilité]$^{+}$

Exemples dans les dictionnaires des racines avec variantes.

(4)

;춥,춥,,A,Cadj.BCU_B_X
;추우,춥,,A,Cadj.BCU_U_J
;흐르,흐르,,V,Cver_LLE_V_X
;흘ㄹ,흐르,,V,Cver_LLE_X_J
;발전,발전,,N+/PRED1,NC
;발전,발전하,,VS+INT,HADA1
;발전되,발전되다,,VS+INT,HADA1

;사랑,사랑스럽다,,A,Cadj_SEU_LEB
;가공,가공적,,ADJ,Jek_i_da

Le champ « compatibilité » contient le nom du fichier qui représente la compatibilité de la racine avec les séquences des suffixes flexionnels. Ils sont situés dans le dossier 'Korean/suffixe'.

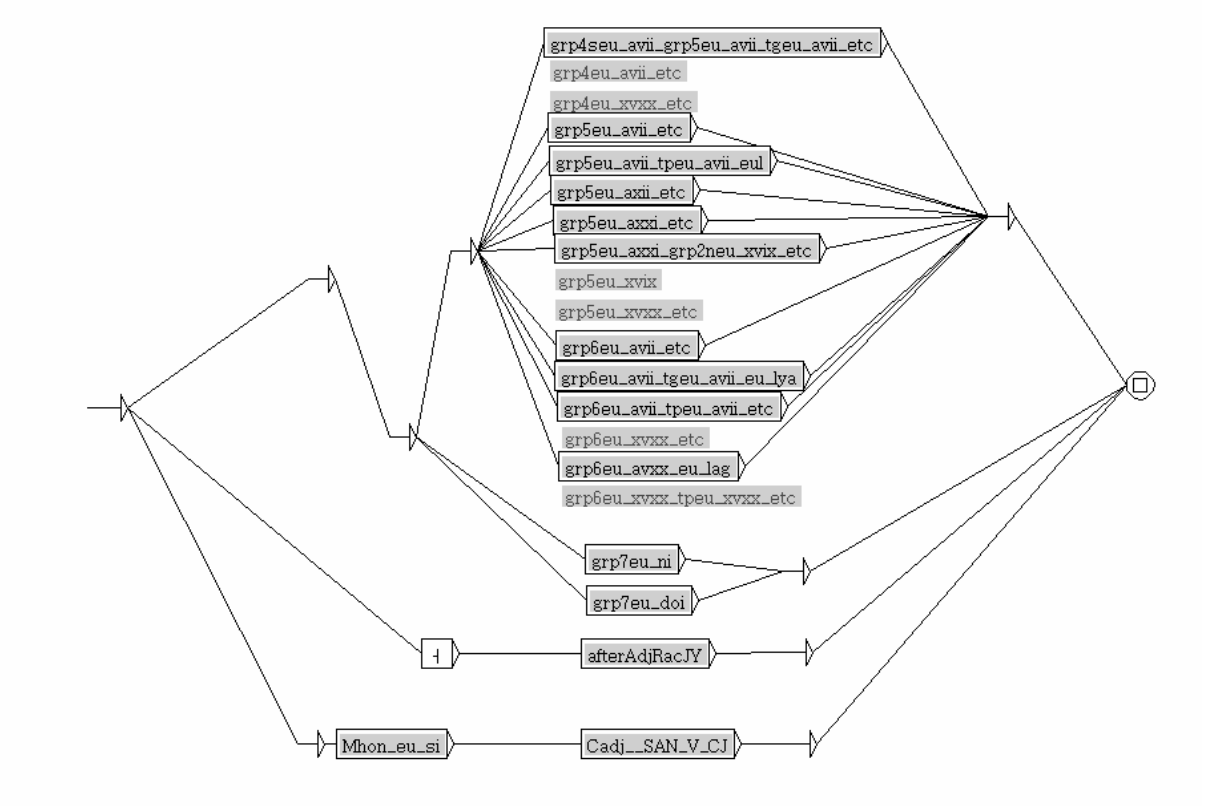


Figure 3-13 Graphe "Cadj_BCU_U_J.grf"

Si le mot n'accepte aucun suffixe, ce champ est vide ou avec le symbole <E>.

Par exemple, le graphe (3-13) « Cadj_BCU_U_J.grf » montre le sous-graphe de la description de la séquence des suffixes des variantes qui ont la compatibilité « Cadj_BCU_U_J ». Ce sous-graphe représente les séquences de suffixes soudés à la variante qui se termine par la voyelle 'u'.

Le suffixe -적 :的 :-_*jôk* est un suffixe de dérivation adjectivale.

(5)

a. 발전적 증상
bal_jôn_jôk jeung_sang
[bal_jôn][jôk] [jeung_sang]

: Développement +St. symptôme
Le symptôme du développement
b. 마르크스적 생각과 행동
ma_leu_keu_seu_jôg_jôk saing_gag_goa haing_dong
[ma_leu_keu_seu][jôk] [saing_gag][goa] [haing_dong]
"Karl Marx"+Suf.det pensé.Post.coor action
La pensée et l'action de type marxiste

Le graphe « moph°jek.grf » montre la séquence des suffixes avec le suffixe -*_jôk* qui n'a que deux séquences des suffixes : -*_i_da* (adjectif), *-eu_lo* (postposition d'adverbe). Si ce suffixe se termine sans un autre suffixe, le mot est un modifieur (5b).

La séquence « <#> » interrompt le parcours de l'automate. Ce sont les marques de la fin de séquence de suffixes.

Le mot en grisé « Cida_C » indique le sous-graphe qui présente les séquences des suffixes qui commencent par le suffixe *_i_da* : «être» après les morphèmes qui se terminent par une consonne.

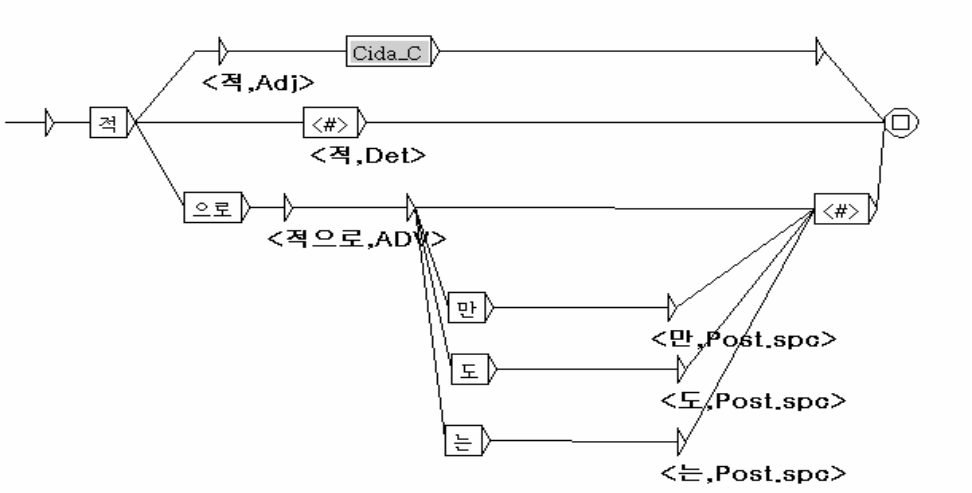


Figure 3-14. Graphe du suffixe -적 : -*_jôk*

3.3.1.4 Données sur les dictionnaires des racines

Nous utilisons les dictionnaires composés par Jee-Sun NAM [NAM 1997] Nous ajoutons les informations sur les compatibilités des suffixes dans les entrées des dictionnaires de base.

Dans le dictionnaire des noms simples, nous changeons les informations associées à la dérivation en ajoutant le symbole de dérivation. Pour les suffixes de nom, nous ajoutons ici les trois éléments de la compatibilité phonétique, NC pour les noms terminés par des consonnes sauf « ㄹ : l », NL pour les noms terminés par la consonne « ㄹ : l ». NV pour les noms terminés par une voyelle.

Le dictionnaire des noms

Type de compatibilité	Nombre
NV	5 794
NL	1 085
NC	7 244
total	14 123

La distribution du code de dérivation sur les entrées de noms simples, 14 123.

Type du nom dérivé	nombre	suffixe
Pred1	3 136	-ha_da transitif
Pred2	2 687	-ha_da intransitif
Pred3	1 778	-doi_da intransitif
ADJH	232	-ha_da adjectif
Total	7 823	

La distribution du code sur le dictionnaire des verbes

Type du code	Code	Nombre
Flexion	SANVARV	521
	SANVARC	2 454
	BCUVARJ	12
	BCUVARY	2

	HADA1	2
	DXCVAR	38
	LEUVAR	2
	LXXVAR	136
	LLEVAR	290
	NOHVAR	39
	PUDVAR	1
	SXVVAR	30
	VXVVAR	57
	Somme	3 603
Dérivation	-ge_li	1 851
	-ji	300
	-teu_li	123
	-ha	3 028
	-honorifique	7
	somme	5 309
Adverbes dérivés		1 338
Total		10 240

Pour les verbes et les adjectifs, nous donnons deux types de groupes pour la compatibilité de la séquence des suffixes : flexion, dérivation.

Le dictionnaire des adjectifs

Type du code	Code	Nombre
Flexion	SANVARV	187
	SANVARC	158
	BCUVARJ	82
	BCUVARY	35
	HXXVAR	87
	LEUVAR	10
	LXXVAR	36
	LLEVAR	40
	SXVVAR	1
	VXVVAR	28
	HADA0	2
	somme	684
Dérivation	-lob	59
	-maj	26

	-ebs	94
	-ida	88
	-iss	6
	-jek	721
	-hada	3 268
	-seu_lob	357
	somme	4 593
Adverbes dérivés		628
Total		5 905

Le dictionnaire des mots invariables

Type	Code	
Déterminante	DET	103
Injection	INJ	360

3.3.2 Partie du suffixe

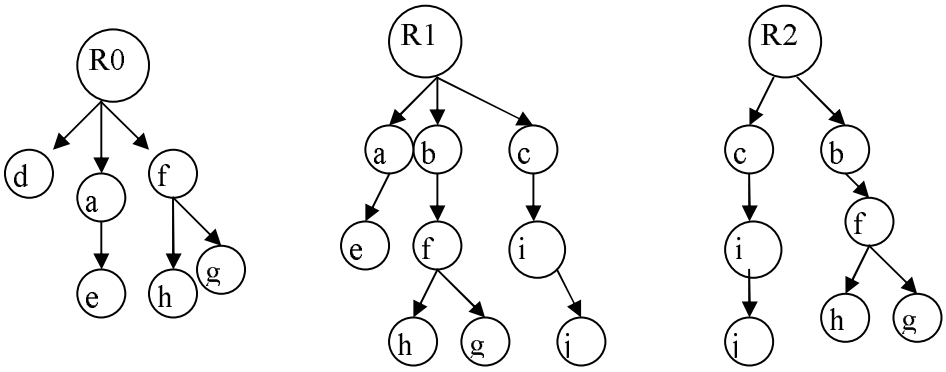
La séquence des suffixes est représentée sous forme de graphe par le transducteur. Les graphes présentent des séquences des suffixes sous forme fléchie. Nous créons les graphes qui ont un nom identique à celui du champ de « Information sur suffixes flexionnels » dans les dictionnaires des variantes de racines.

Le format de description des morphèmes dans une boîte dans le transducteur est la suivante : **Forme fléchie / < Forme canonique, Informations linguistiques >**

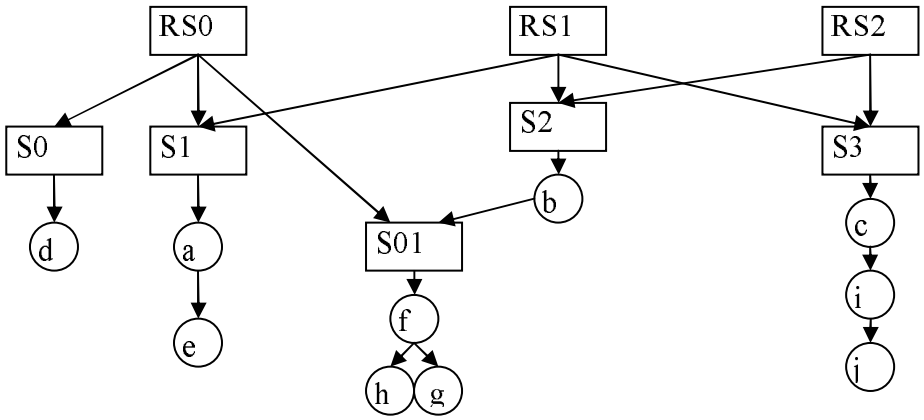
C'est le format d'une ligne dans l'éditeur de graphe du système UNITEX ; Dans le graphe, la partie « forme fléchie » est affichée dans la boîte qui présente un état de transducteur. La partie « informations » à la suite du symbole « / » est affichée en-dessous de la boîte ; c'est la sortie d'un état.

3.3.2.1 Structuration des séquences de morphèmes

Pour commencer, nous exprimons les séquences de suffixes selon les compatibilités de la racine (R1, R2, R3) sous forme d'automate comme la figure (3-15a), les états initiaux indiquent la compatibilité avec chaque racine. Nous pouvons aussi exprimer les chemins des arbres en ajoutant des nœuds (S0, S1, S01, S2, S3) qui indiquent les chemins communs (3-15b). Le nœud ajouté et les chemins à partir de ce nœud construisent un sous-automate.



a) Des arbres des séquences chaque compatibilité



b) Des arbres des séquences avec sous-arbres

Figure 3-15 Arbres de séquences

La figure (3-16) montre l'expression par les sous graphes. La figure (3-17) montre les graphes des séquences de suffixes verbaux compatibles avec les racines verbales non variables et terminées par les voyelles (à gauche).

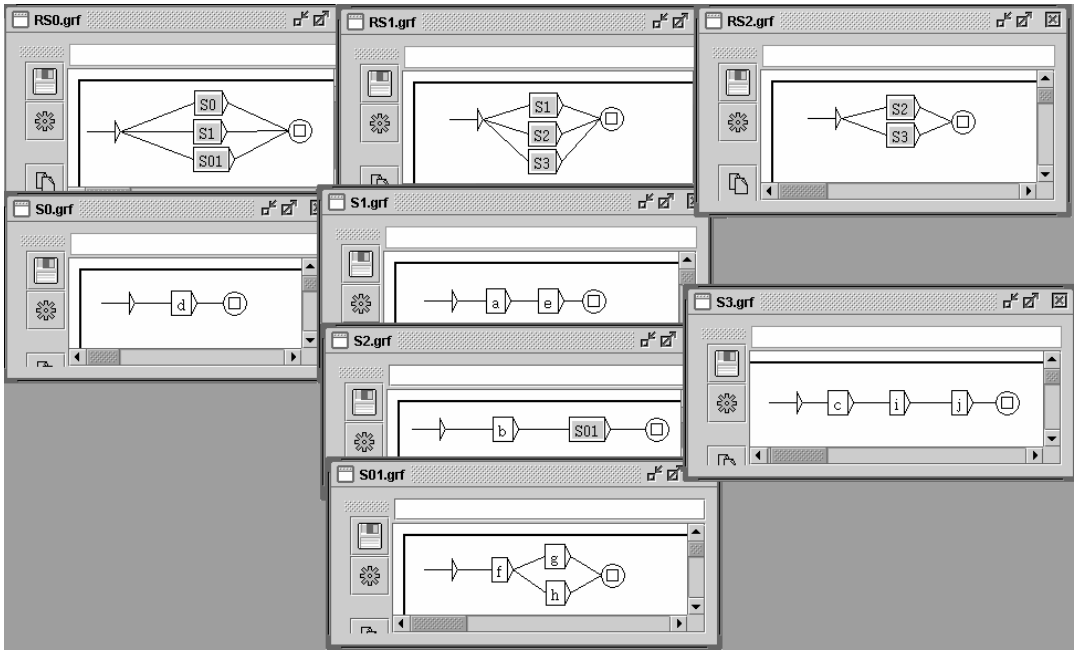


Figure 3-16 Expression sous forme sous graphes

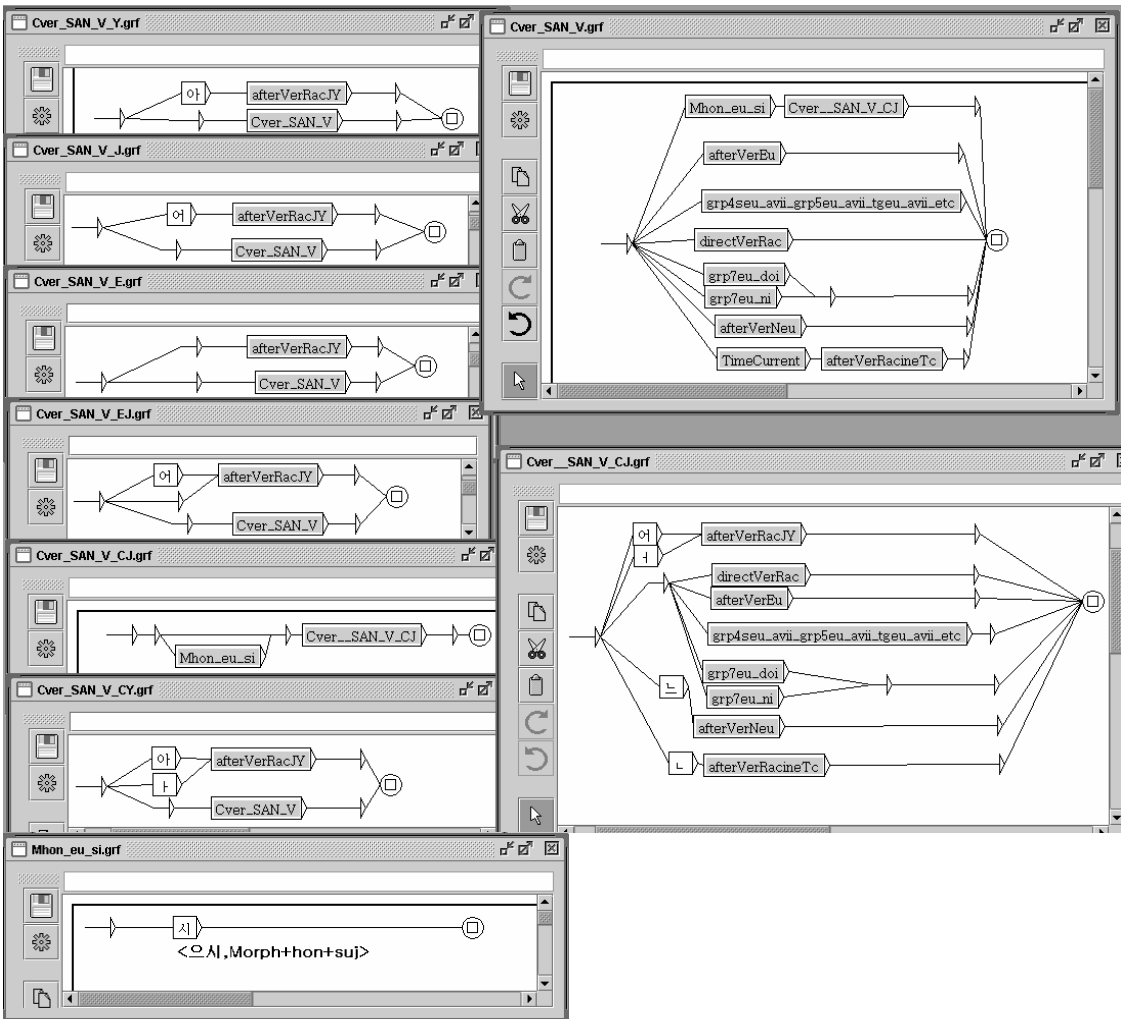


Figure 3-17 Construction des séquences de suffixes avec sous graphes

3.3.2.2 Description des morphèmes du suffixe

Dans la description des séquences des suffixes avec les sous-graphes, certaines formes sont rattachées au morphème précédent ou suivant. La figure (3-18) montre une forme rattachée au morphème précédent.

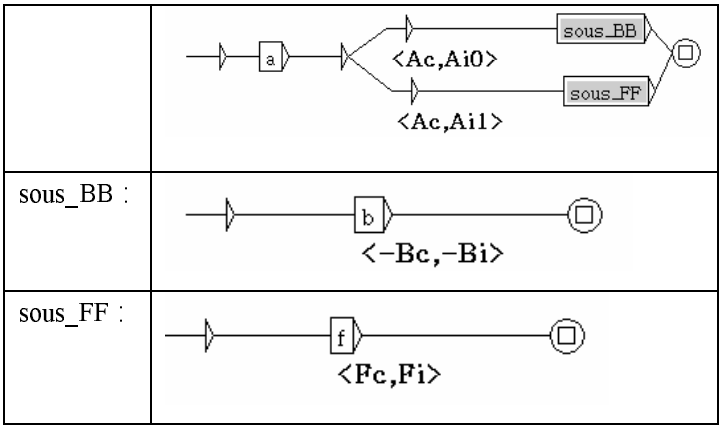
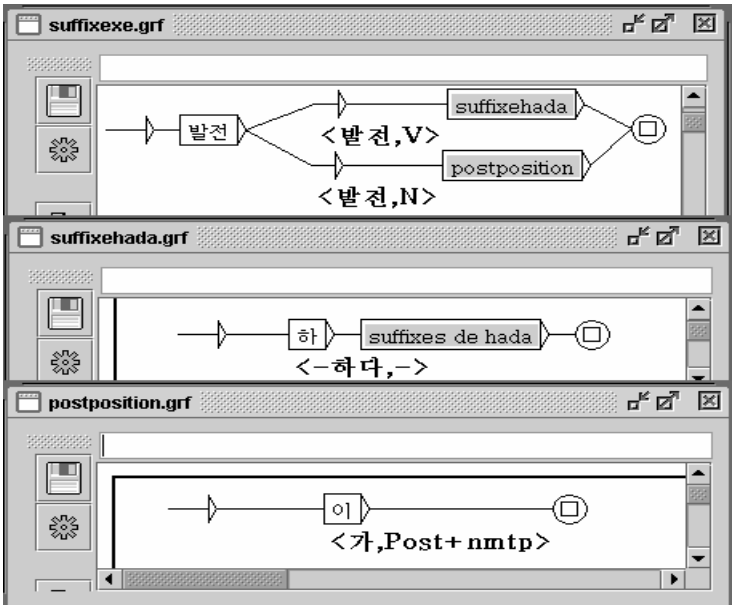


Figure 3-18 Séquences avec une forme rattachée au morphème précédent

Le symbole '-' signifie la concaténation de forme canonique à la forme canonique du morphème précédent et l'ajout d'informations aussi dans les champs de morphème précédent, dans ce cas, la séquence {ab} possède la forme canonique 'AcBc' et les informations 'Ai+Bi' « ab/AcBc ;Ai+Bi» et la séquence {af} correspond à deux éléments « a/Ac,Ai0 + f/Fc,Fi».

Par exemple, Le mot « *bal_jôn* » existe en lui-même et avec les séquences nominales, c'est un nom : *développement*. S'il est soudé avec le suffixe 하다 : *-_ha_da* (faire), il devient une racine du verbe : 발전하다 : *bal_jon_ha_da* (développer).



La figure (3-19) montre une forme rattachée au morphème suivant. La séquence 'Gd'.est analysée « Gd/GcDc, Di+Gi » et

la séquence ‘Ge’ est analysée « Ge/GcEc,Ei+Gi ».

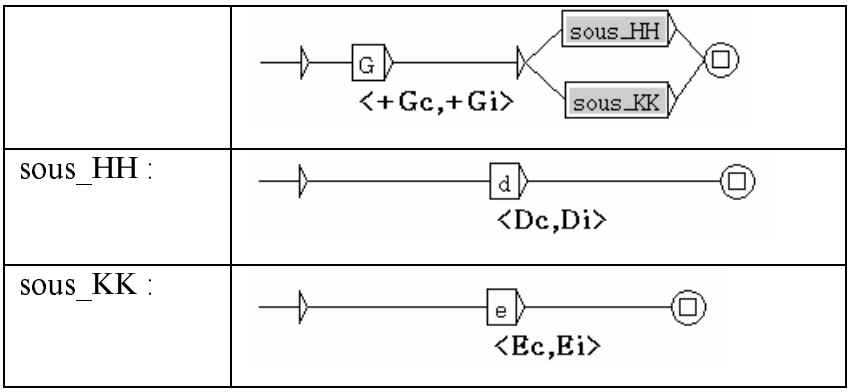


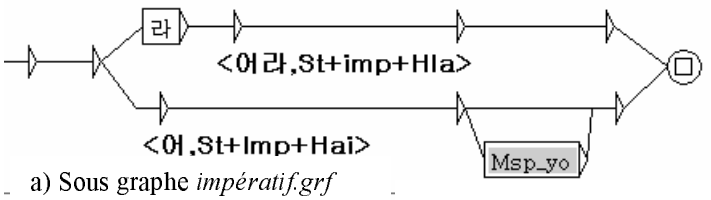
Figure 3-19 Séquences avec le morphème dépendu du morphème suivant

Exemples de morphèmes

Le graphe « TimeFuture.grf » ci-dessous montre le morphème qui représente le suffixe du futur.

Figure 3-20 Le graphe de morphème de temps futur « Timefuture.grf »

Le graphe «impératif.grf » montre le suffixe terminal impératif.



a) Sous graphe *impératif.grf*

b) Sous graphe *Myo_yo.grf*

Figure 3-21 Les suffixes terminaux impératifs 어라 :-*_ô_la* et 어 :-*_ô*

Le suffixe -어라 *:-_ô_la* a une syllabe ‘*-_ô*’ qui a plusieurs variantes. Nous situons ce sous-graphe après chaque variante de la syllabe ‘*_ô*’.

Les suffixes contractés

Dans l'écriture coréenne, le symbole de la citation est officiellement « " ».

Les exemples (6) comportent du discours rapporté.

(6)

a. "그는 집으로 갔다"라고 했다.
"*geu_neun jib_eu_lo gass_da*"***la_go hass_da***
[geu][neun] [jib][[eu_lo] [ga_da][ôss][da][i_la_go] [ha_da][ôss][da].
Lui+Post.nmtp maison+Post.ver aller+Mtpass+St.déc+Post.citation faire+Mtpass+St.déc
b. 그는 집으로 갔다라고 했다.
*geu_neun jib_eu_lo gass_da_**la_go hass_da***
c. 그는 집으로 갔다고 했다.
*geu_neun jib_eu_lo gass_da_**go hass_da***
d. 그는 집으로 갔댔다.
geu_neun jib_eu_lo gass_daiss_da

On a dit « il est allé à la maison ».
On a dit qu'il est allé à la maison.

Toutes les phrases de (6) sont identiques, la différence entre (6a) et (6b) est l'existence de symboles de citation. Entre (6b), (6c) et (6d), c'est la forme contractée. La séquence « Post.citation + « espacement » + *ha_da* » s'est contractée. Dans la phrase (6d), cette séquence a disparu et les deux mots sont contractés.

Le graphe « contract_da_go__ha » montre les suffixes disparus à cause de la contraction. Mais nous mettons les informations avec les formes fléchies vide « <E> » pour soutenir les informations linguistiques disparues.

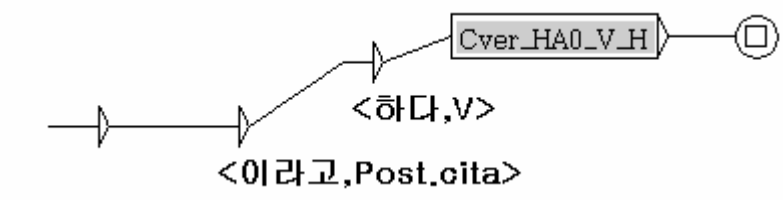


Figure 3-22 Eléments disparus à cause de la contraction

Cette contraction provoque une augmentation considérable du nombre des mots coréens. Avant le suffixe « -*_i_la_go* », tous les mots coréens en séquence des morphèmes peuvent se situer. Cette séquence produit environ (137 516 533 X 4051) mots coréens d'après le tableau (3-9).

Les suffixes dérivatifs

Le suffixe -음 :-*_eum* est un suffixe de nominalisation soudant les racines verbales et adjectivales. Avec un suffixe -직 :-*_jig* qui représente la déduction, la séquence -_eu*m_jig* se soude seulement aux racines verbales et sert à former un adjectif.

(7) 그 사과는 먹음직스럽다
geu sa_goa_neun môg_eum_jig_seu_lôb_da.
[geu] [sa_gos][neun] [mog_da][eum][jig][seu_lôb][da].
Ce.Det pomme+Post.nmtp manger+Suf.nominal+Suf.adj +St.déc
Cette pomme semble délicieuse.

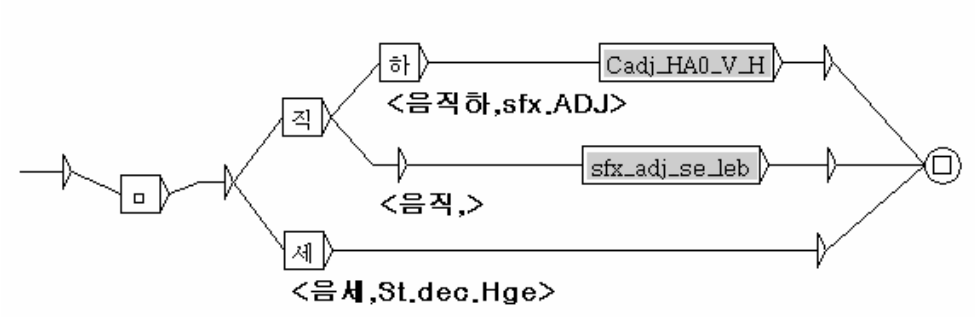


Figure 3-23 Suffixes avec une syllabe "-*_eum*"

Le graphe « sfx_adj_seu_lôb » représente le suffixe qui sert à former un adjectif : «-스럽- :-*seu_lô*-». Ce suffixe se soude aux racines de nom.

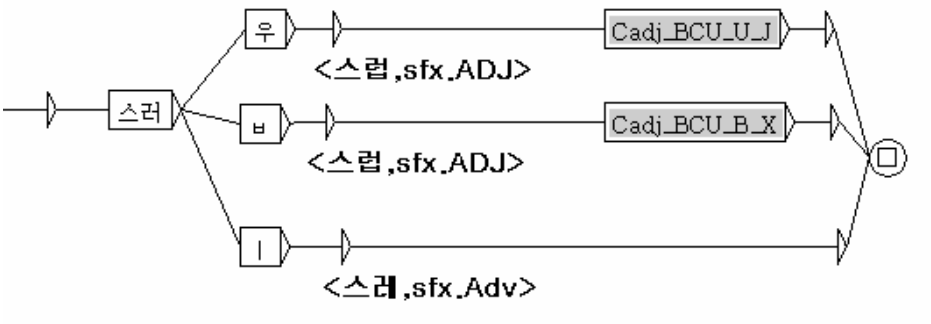


Figure 3-24 Séquences pour plusieurs suffixes « sfx_adj_seu_leb.grf »

Le suffixe terminal -게 :-*_ge* est un suffixe adverbial. Il se soude avec n'importe quelle racine verbale et adjectivale. Cette racine est devenue un adverbe dérivé. Il existe d'autres suffixes adverbiaux : -이 : -*_i*, -히 : -*_hi*. Dans le cas du suffixe adjectif -스럽- :-*_se_lôb*-, le suffixe adverbial -*_i* peut se souder avec une variante du suffixe -스러 :-*se_lôb*. Ce suffixe a aussi les séquences de suffixes adjectivaux comme les racines adjectivales qui ne se terminent pas par la consonne 'b'.

Les suffixes en formes facultatives

Les racines verbales 떨어뜨리다 : *ddôl_ô_ddeu_li_da*, 떨어트리다 : *ddôl_ô_ddô_li_da*, 떨어터리다 : *ddôl_ô_tô_li_da* « faire tomber quelque chose de manière violente » sont acceptées comme les mots standards.

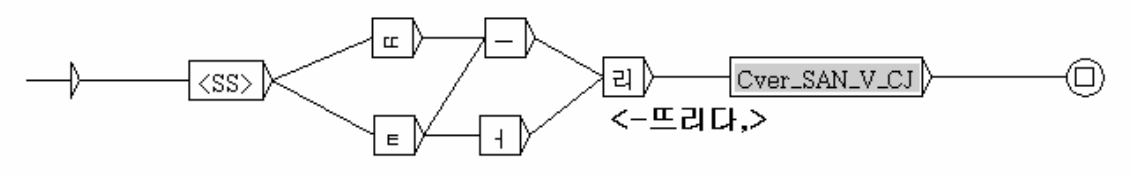


Figure 3-25 Séquences pour multiforme.

Nous traitons les mots qui se terminent par un des suffixes -*_ddeu_li_da*, -*_ddô_li_da*, -*_tô_li_da* comme un mot avec un suffixe -*_ddeu_li_da*. On peut découper ces mots en séquence : une racine+<ô : Suffixe conjonctif>+*ddeu_li*+<Suffixes verbaux>. Le suffixe conjonctif <-_ô> avec la racine verbale fabrique la séquence des verbes composés. Le verbe suivant est devenu un auxiliaire. En l'absence d'un mot verbal 뜨리다 :*ddeu_li_da*, c'est toujours un suffixe verbal. On suppose que c'était un verbe auxiliaire mais qu'il est devenu un suffixe.

3.4 Construction de dictionnaires comprimés

Après avoir construit les dictionnaires sous forme de liste et de graphe, nous convertissons les dictionnaires en une forme comprimée afin de les utiliser en réalité pour la consultation des mots dans le texte.

La figure ci-dessous montre les étapes de la construction des données comprimées.

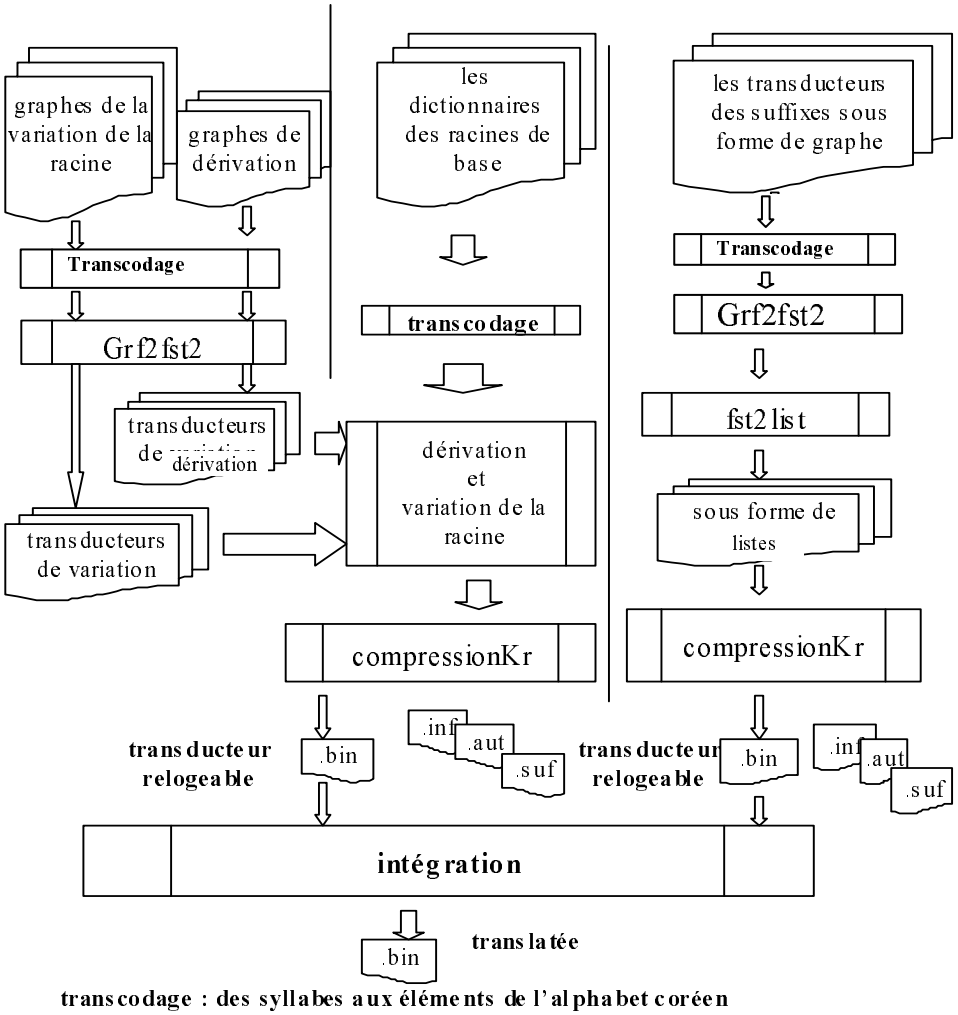


Figure 3-26 Etapes pour obtenir un dictionnaire comprimé

La partie racine au centre du graphe montre le processus de traitement des dictionnaires sous forme de liste. Le processus pour obtenir le dictionnaire électronique coréen commence par la construction des dictionnaires de base, ensuite les transducteurs de

dérivation et variation sont appliqués pour obtenir les dérivées et les allophones.

Les dictionnaires de racine sont comprimés. Une entrée comprimée est un morphème de surface avec la forme canonique et les informations.

Les graphes des transducteurs des séquences des morphèmes grammaticaux sont convertis en listes du même format que les dictionnaires des racines, puis comprimés.

Après la compression, les fichiers comprimés partiels contiennent des transitions relogeables. Une transition non relogeable contient une valeur absolue d'un état but. La transition relogeable contient une valeur qui indique un état but qui existe dans l'autre fichier partiel.

Nous réunissons des dictionnaires partiels en un dictionnaire qui contient toutes les séquences de morphèmes dans les mots.

3.4.1 Traitement des graphes de séquences des morphèmes grammaticaux

Les graphes représentent les séquences des morphèmes. Le transducteur se compose d'états où l'on présente les morphèmes sous forme fléchie. Chaque branche représente l'ordre des morphèmes de gauche à droite.

Le programme « fst2list »[35] réalise l'extraction des séquences des morphèmes qui sont des états de l'automate ou du transducteur. Les éléments sont récupérés en séquence par un parcours de l'automate ou du transducteur. Nous l'utilisons pour transformer des graphes en liste, pour une unification de la forme des entrées du programme comprimé par rapport avec les dictionnaires sous forme de liste. Dans l'annexe 5, nous montrons l'utilisation du programme « fst2list ».

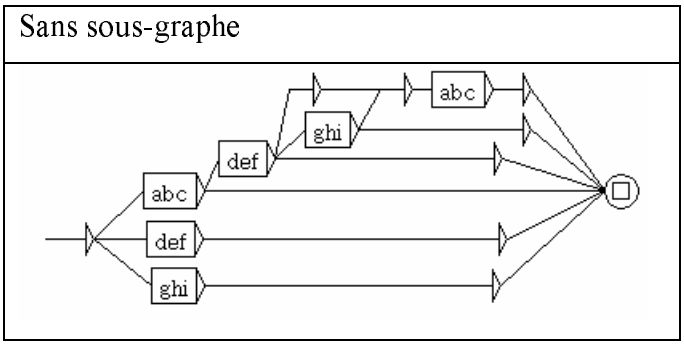
Le format du résultat de l'extraction sous forme de liste est le suivant.

[; Variante, Canonique, Origine, Informations linguistiques, Compatibilité]*

Il en est de même du format du dictionnaire de racines avec variantes, où la différence est la présence de plusieurs morphèmes par rapport aux dictionnaires avec variantes de racines.

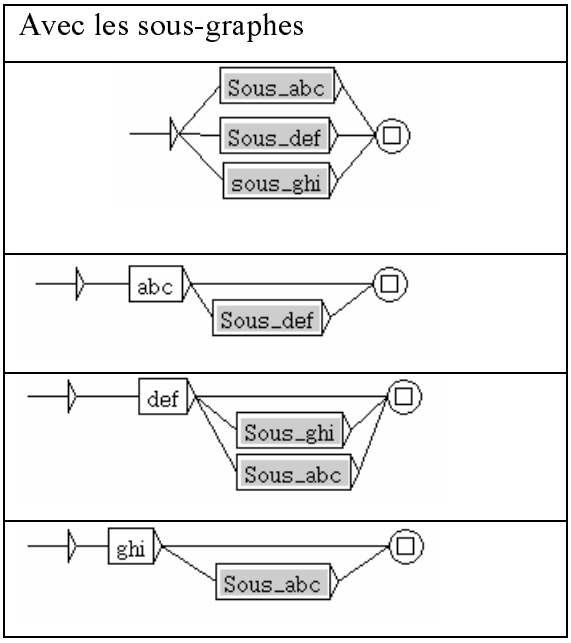
La construction du sous-graphe pour les parties communes aide à diminuer la taille de la description, et à éviter la répétition des descriptions. Elle donne de plus une solution au problème des chemins cycliques. Nous n'avons pas le moyen de résoudre parfaitement des chemins cycliques.

Sans sous-graphe

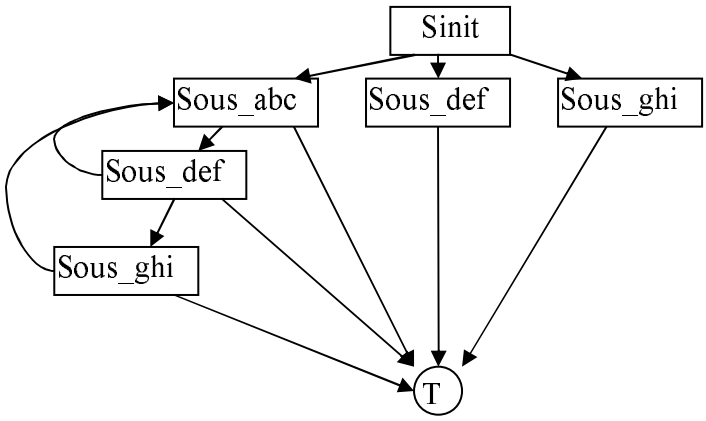


Quand nous transformons ce graphe en réseau de transitions récursif (RTN) avec des sous-graphes, nous pouvons créer des cycles.

[35] Le programme énumère les chemins du transducteur ou de l'automate.



Avec cette méthode, il existe des appellations cycliques. Le graphe ci-dessous montre les appels aux sous-graphes. Le 'T' est terminal.



Le premier transducteur ne contient que des sous-graphes. Les noms de sous-graphes sont identifiés aux noms de compatibilité qui se trouvent dans les dictionnaires de racines dans le champ de la compatibilité. Le premier transducteur est un automate multi-initiaux [REV 1991]. Le graphe (3-27) montre un de tel graphe.

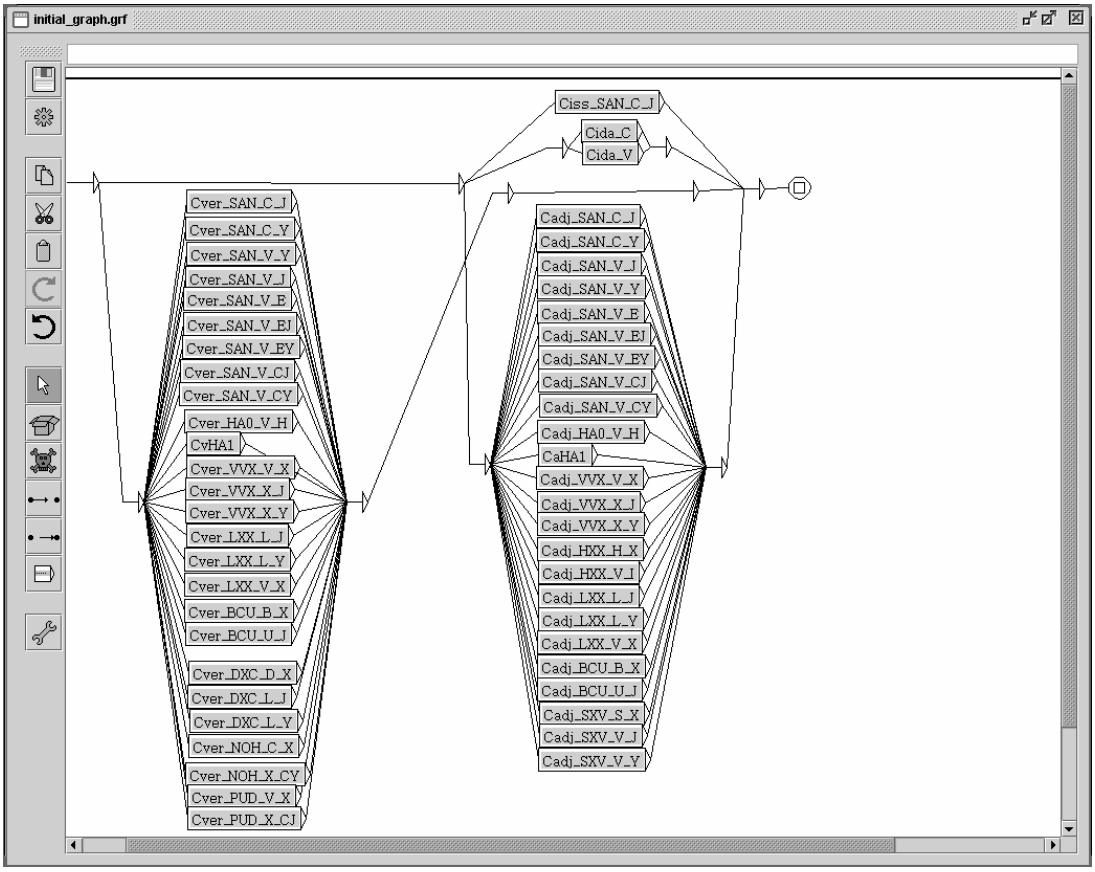


Figure 3-27 Premier graphe des séquences de suffixes verbaux et adjectivaux

Nous avons construit 228 sous graphes pour les séquences des suffixes adjectivaux et verbaux, 175 sous graphes pour les postpositions, 5 sous graphes pour les suffixes dérivés.

3.4.2 Méthode de compression du dictionnaire

Nous traitons les données des dictionnaires sous forme comprimée, nous avons essayé deux méthodes : l'automate et le transducteur.

Nous avons construit un dictionnaire comprimé de plusieurs parties de racines ou de séquences de suffixes. La structure de l'automate de base est empruntée à UNITEX, nous avons modifié la structure pour la liaison des parties entre les racines et les suffixes. Dans le dictionnaire comprimé sous UNITEX avec la structure d'automate d'arbre, toutes les informations se situent à la fin des nœuds.

Nous faisons la compression de toutes les parties de racines et suffixes. Après la compression de chaque partie, il existe des transitions relogeables résolues lors de l'unification des parties comprimées individuellement.

Structure de données en automate

- **Structure d'un état**
  - Index des informations
  - Nombre de transitions
  - Tableau de transitions
- **Structure d'une transition**
  - Caractère
  - Adresse d'état but

Pour le premier essai de représentation des données des dictionnaires, nous construirons des automates. Cette méthode est fondée sur UNITEX et nous y ajoutons le traitement des nœuds terminaux. Les nœuds terminaux dans la partie de la racine portent l'information de compatibilité qui indique le nom de l'automate des séquences de suffixes. La figure (3-28) montre les chemins de l'automate. La séquence de caractères « sb » est la marque de début de syllabe.

Pour les séquences des suffixes, le chemin présente la séquence des caractères. A la fin de la séquence, il existe un nœud qui préserve les informations du mot. Les informations forment deux parties.

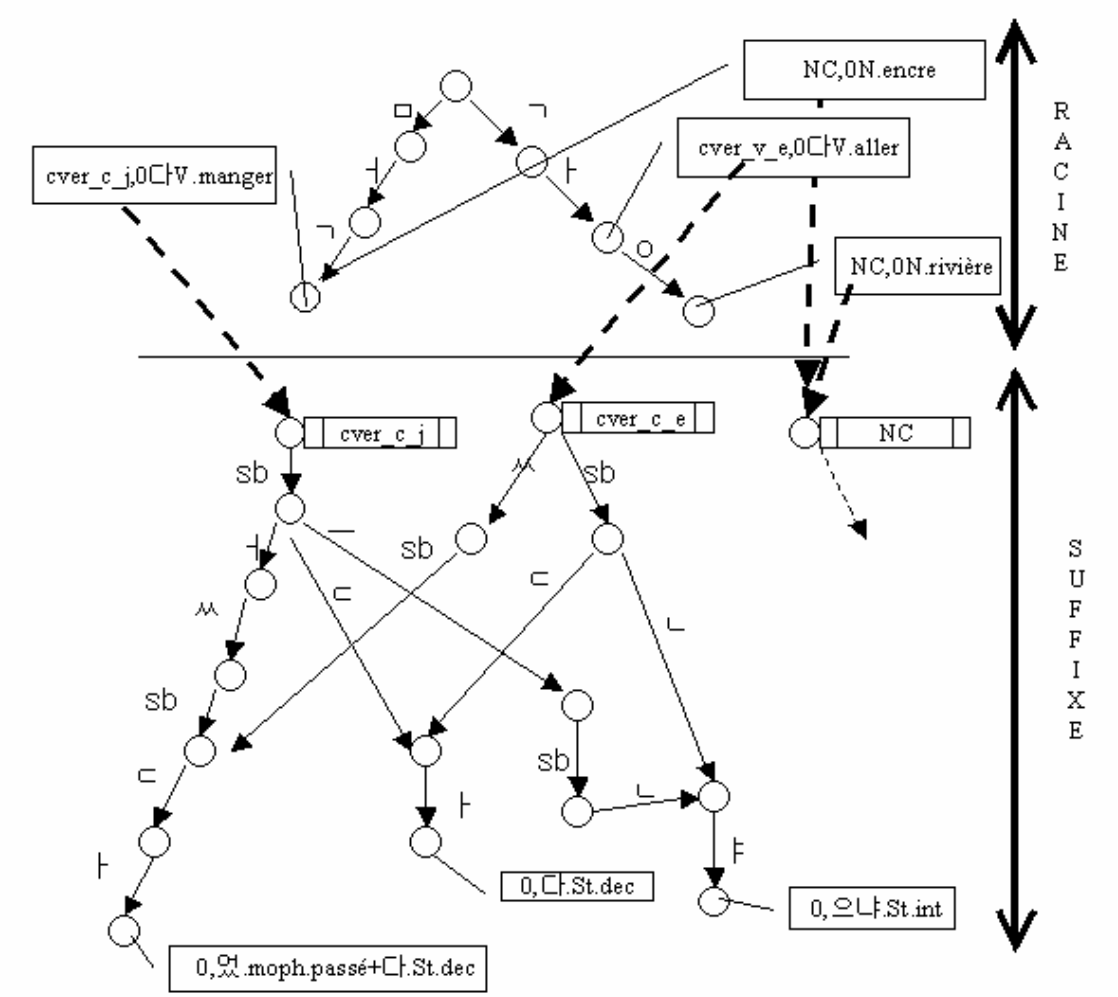


Figure 3-28 Graphe de liaison des racines et suffixe du dictionnaire (méthode par automate)

La première partie contient les informations sur la compatibilité à droite. Si elle est vide (0), le parcours est terminé. La deuxième partie contient les informations linguistiques. Quand on consulte les dictionnaires, on lie les noms des suffixes obtenus par parcours de l'arbre racine vers la racine correspondante de l'état initial de l'arbre des suffixes.

Avec cette méthode, après le parcours pour identifier les séquences d'un mot, toutes les informations de chaque séquence sont situées à l'état terminal.

Structure de données en transducteur

- **Structure d'un état**
  - Nombre de transitions
  - Tableau des transitions
- **Structure d'un élément de transitions**
  - Caractère
  - Sorti/Contrôle si Caractère == 0
  - Etat but/index de nom de sous-graphe

Dans la structure d'état de la méthode par transducteurs, chaque nœud ne contient aucune information. Mais chaque transition porte ses informations.

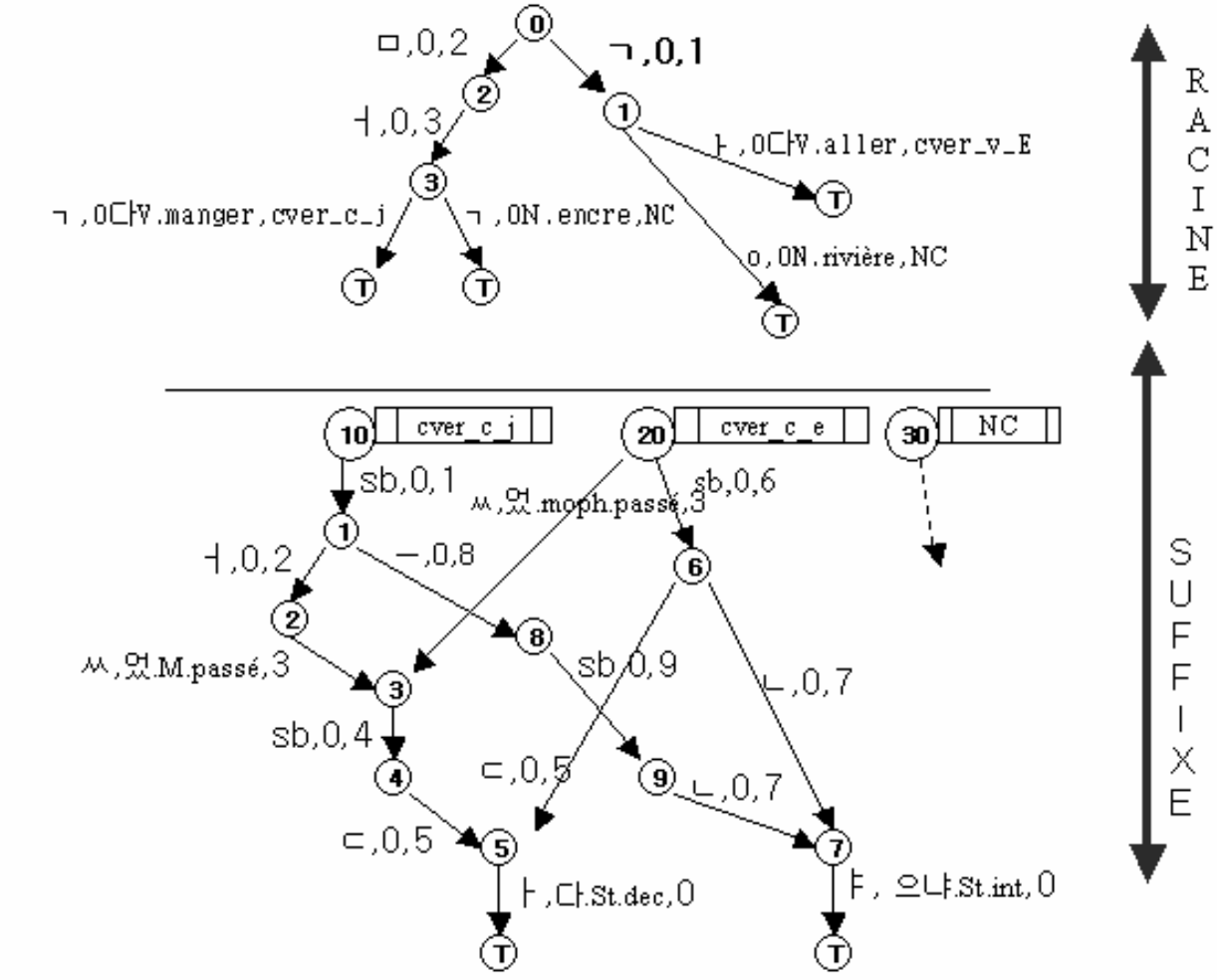


Figure 3-29 Graphe de transducteur de la racine et de la séquence de suffixes

Si la valeur de l'état but est 0, le parcours est terminé. Pendant la consultation, on arrive à la fin de la séquence consultée, on trouve un mot. Si la valeur de l'état but n'est pas 0,

on va au nœud suivant.

Le tableau suivant montre la comparaison entre deux structures. Nous avons utilisé six dictionnaires. L'unité des nombres est le kilo octets

Ce résultat montre que la structure de transducteur est préférable à celle d'automate sur les séquences des suffixes. Nous avons utilisé la deuxième méthode pour l'étape suivante : réunion des dictionnaires. La deuxième méthode est meilleure pour le traitement des séquences de morphèmes.

Tableau 3-6 Comparaison de la taille en kilo octet

Structure	Noms des dictionnaires						Total	Réunion
	adjectifs	noms	verbes	Suf.av	Suf.N	etc.		
Arbre	76	135	145	334	4	234	928	
Transducteur	102	204	194	39	2	19	560	555

Réunion des transducteurs de racines et de suffixes

Après la compression de chaque transducteur de racine et de séquence de suffixes, les fichiers comprimés sont les dictionnaires comprimés « relogeables » : certaines transitions n'indiquent pas une adresse absolue d'état but mais un nom qui indique un des sous-graphes qui existe dans l'autre dictionnaire. Nous effectuons la résolution des branches non encore résolues en même temps que la réunion des transducteurs relogeables. Après la réunion des dictionnaires comprimés relogeables, nous obtenons un seul dictionnaire comprimé (3-30).

Pendant la réunification des dictionnaires comprimés, nous remplaçons les noms des graphes par les états initiaux correspondants.

Nous pouvons encore utiliser les fichiers comprimés relogeables de la partie des séquences de suffixes avec les autres dictionnaires de racine.

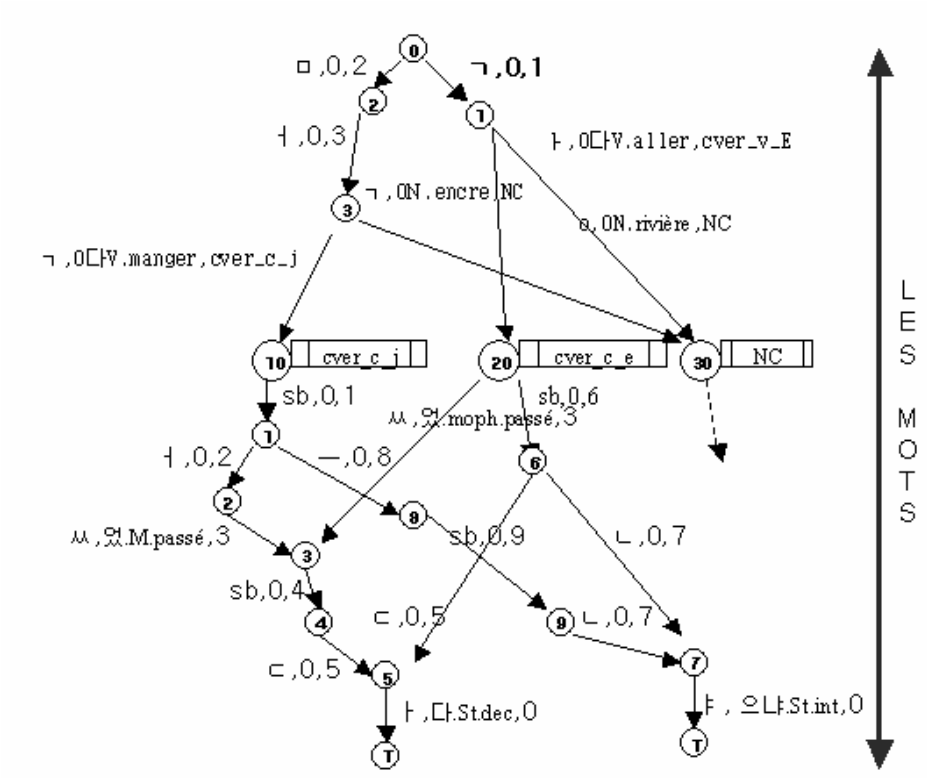


Figure 3-30 Un dictionnaire réuni et comprimé

3.5 Résultat

La taille du dictionnaire réuni est presque autant que le total des tailles de modules d'entrée au dictionnaire réuni (tableau 3-8).

Tableau 3-7 Données sur les dictionnaires comprimés avant la réunion.

	Nombre d'entrées	Nombre de nœuds	Nombre de nœuds après la minimisation	Nombre de transitions	
				sans sortie	Avec sortie
Verbes	10 852	30 417	15 534	5 006	19 088
Noms	21 958	25 986	17 146	10 513	16 394
Adjectifs	6 283	17 496	10 102	1 608	10 939
Suffixes verbaux et adjectivaux	127 253	264 986	2 077	2 403	3 237
Suffixes nominaux	208	387	67	46	78
Total	166 554	339 272	44 926	19 576	49 736

Le tableau suivant montre le nombre approximatif des mots contenus dans le dictionnaire. Le suffixe a des sous-graphes, nous comptons le nombre de chemins des sous-graphes.

La deuxième colonne donne les nombres d'entrées par types de suffixe dans le dictionnaire. La troisième est le nombre de chemins de transducteur.

Nous montrons les étiquettes des morphèmes à l'annexe 5.

Tableau 3-8 Nombre des mots coréens

Type de suffixe	Nombre de racine	Nombre de séquences des suffixes	Nombre de mots
ADV	1964	5	9820
Cadj_BCU_B_X	176	383	67408
Cadj_BCU_O_Y	35	687	24045
Cadj_BCU_U_J	141	2682	378162
Cadj_BCU_U_X	35	1995	69825
Cadj_HA0_V_H	1	3736	3736
Cadj_HXX_H_X	87	380	33060
Cadj_HXX_V_I	87	2693	234291
Cadj_LXX_L_J	32	746	23872
Cadj_LXX_L_Y	8	746	5968
Cadj_LXX_V_X	40	2062	82480
Cadj_SAN_C_J	172	3071	528212
Cadj_SAN_C_Y	26	3071	79846
Cadj_SAN_V_CJ	126	3736	470736
Cadj_SAN_V_CY	1	3736	3736
Cadj_SAN_V_E	43	3049	131107
Cadj_SAN_V_EJ	14	3736	52304
Cadj_SXV_S_X	1	351	351
Cadj_SXV_V_Y	1	2427	2427
Cadj_VVX_V_X	74	2427	179598
Cadj_VVX_X_J	45	687	30915
Cadj_VVX_X_Y	29	687	19923
CaHA1	3491	3824	13349584
Cida_C	78	1691	131898
Cida_V	12	4744	56928
Ciss_SAN_C_J	6	2409	14454
Cv_Geli	1851	3980	7366980
Cver_SAN_V_CJ	7	1988	13916
Cver_BCU_B_X	14	457	6398
Cver_BCU_O_Y	2	686	1372
Cver_BCU_U_J	12	2851	34212
Cver_BCU_U_X	2	2165	4330
Cver_DXC_D_X	68	407	27676
Cver_DXC_L_J	63	2841	178983
Cver_DXC_L_Y	5	2841	14205
Cver_HA0_V_H	2	3980	7960
Cver_LXX_L_J	290	1034	299860
Cver_LXX_V_X	290	2215	642350
Cver_NOH_C_X	39	2628	102492
Cver_NOH_X_CY	39	1732	67548
Cver_PUD_V_X	1	2608	2608
Cver_PUD_X_CJ	1	1372	1372
Cver_SAN_C_J	266	3314	881524
Cver_SAN_C_Y	357	3294	1175958
Cver_SAN_V_CJ	3807	3980	15151860
Cver_SAN_V_CY	146	3980	581080
Cver_SAN_V_E	214	3294	704916
Cver_SAN_V_EJ	293	3980	1166140
Cver_SAN_V_J	71	3294	233874
Cver_VVX_V_X	195	2608	508560
Cver_VVX_X_J	172	686	117992
Cver_VVX_X_Y	23	686	15778
CvHA1	8851	4051	35855401
ddeuliVer	123	3980	489540
Morph_JEK	721	10	7210
NC	7244	2356	17066864
NL	1085	5398	5856830
NV	5794	5398	31276012
sfx_adj_se_leb	357	4678	1670046
	39130		**137 516 533**

QUATRIEME PARTIE

4. Application des dictionnaires sur le texte coréen

Dans cette partie nous montrons les étapes de traitement automatique de texte coréen avec ce dictionnaire.

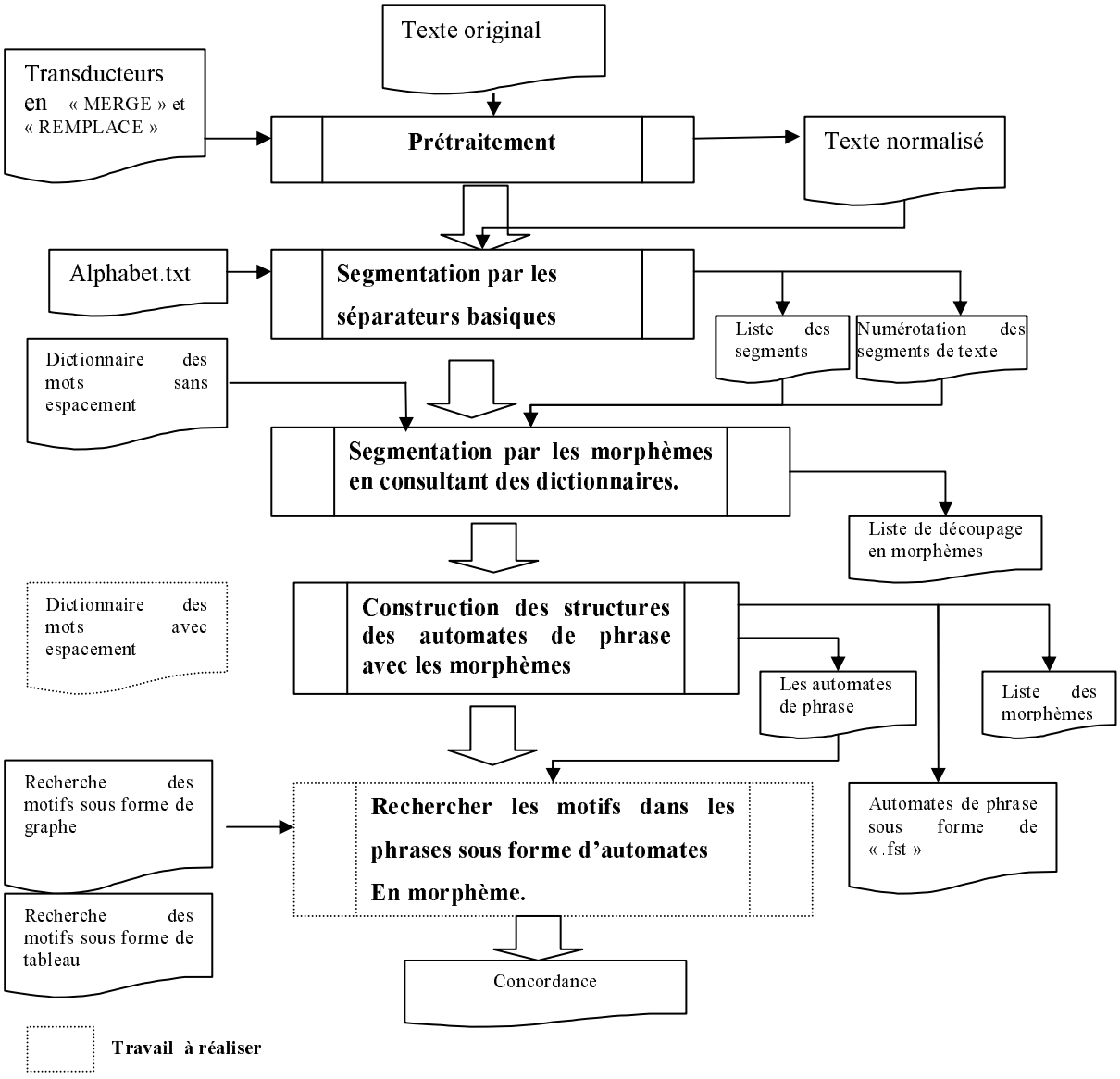


Figure 4-1 Processus de traitement du texte coréen

Nous allons utiliser UNITEX qui est déjà couramment utilisé dans plusieurs langues.

Nous traitons les caractères coréens au niveau de phonèmes qui sont des lettres alphabétiques coréennes. Nous utilisons le dictionnaire comprimé en lettres alphabétiques coréennes. Nous avons ajouté des fonctions de conversion des caractères coréens entre syllabes et lettres alphabétiques coréennes. Nous utilisons la conversion pour l'affichage des résultats en syllabes. La figure (4-1) montre l'architecture du traitement du texte coréen avec UNITEX, la première étape est le prétraitement qui normalise le texte entré. La deuxième est la segmentation, on obtient les segments qui sont les éléments de texte : mots et symboles. Jusque là, nous traitons les textes en syllabes. La troisième est la consultation de dictionnaire de séquences des morphèmes, nous traitons les segments de mots en lettres alphabétiques coréennes. Le résultat de la consultation est les séquences des morphèmes en lettres alphabétiques. Après cette étape, nous affichons automatiquement les résultats en syllabes. La quatrième est la construction de la phrase en automates de phrase pour montrer les résultats de consultation.

4.1 Prétraitement

L'UNITEX traite normalement le texte en UNICODE, actuellement, la plupart des textes coréens sont codés par le système de codage WANSUNG. UNITEX fournit un transcoage.

Nous traitons les textes phrase par phrase. Pour identifier les phrases, nous ajoutons la marque de phrase «{S} » à la fin des phrases.

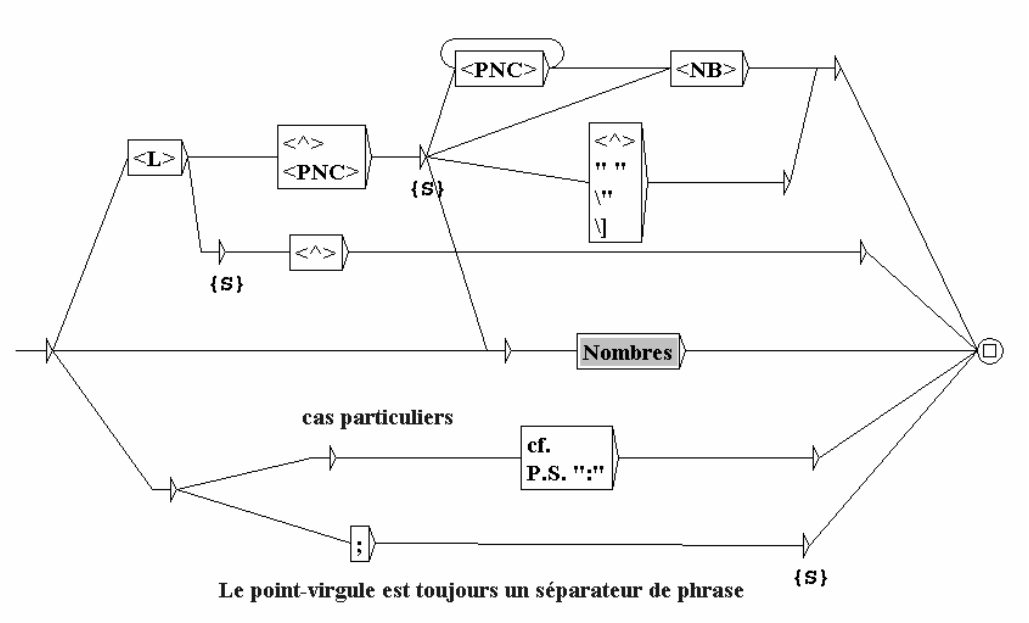


Figure 4-2 Transducteur ajoutant la marque de phrase.

Il existe une restriction pour le texte original. Nous ne pouvons pas traiter le texte qui contient les caractères de changement de ligne. En français, s'il existe des mots coupés par le changement de ligne, on utilise un tiret pour indiquer sa continuation. Mais dans l'écriture coréenne, on change de ligne sans tiret. Pour l'édition de texte, on utilise des papiers qui ont 20 lignes par page et 40 cases par ligne sans tiret, les syllabes et idéogrammes ont une même longueur : chaque carré est réservé à un caractère, mais les caractères latins et les chiffres ont 2 caractères dans un carré. Après l'utilisation des éditeurs informatiques, cette habitude ne change pas.

L'éditeur de texte coréen a des fonctionnalités pour l'écriture verticale pour la présentation de poèmes, les faire-part traditionnels, menus, etc. Nous ne traitons que des

textes en écriture horizontale.

Les symboles pour les citations et focalisations sont les deux paires de symboles « ‘ », « ’ » et « " », « " » sur l’orthographe standard coréenne. Mais on utilise aussi n’importe quelle paire de symboles : ‘「’, ‘」’, ‘『’, ‘』’, ‘【’, ‘】’, ‘◁’, ‘▷’, ‘◀’, ‘▶’. Les symboles 꺽쇠표 *ggôg_ssoi_pyo* (marque de crochet de fer) ‘「’, ‘」’, ‘『’, ‘』’ sont utilisés pour l’écriture horizontale comme les symboles de paire : « ‘ », « ’ » et « " », « " ».

Tableau 4-1 Signes de ponctuation

Classification	signe	application
Terminaison de la phrase	Point : ‘.’., ‘。’	
	Point d’interrogation : ?	
	Point d’exclamation : !	
Suspension	Virgule : ,	Conjonction, apposition
	Point au centre : •	Conjonction
	Barre de fraction : /	
Citation	Guillemets à l'anglaise : ","	
	Accent : ‘ ’	
	(), {}, [], 「」, 『』, <>	
etc	Tilde ~, -	Période et section
	Signe caché X, O, □ …..	Remplaçant les caractères cachés ou inconnus

Le tableau (4-1) montre les signes de ponctuation coréenne, le point au centre ‘•’ est appliqué pour les dates historiques (1a).

(1)
a) les jours historiques : 3•1 운동 *sam_il un_song*: mouvement de l’indépendance (01.03.1919).
b) factorisation de l’affixe
석• 박사[36] : *sôg_bag_sa* :
[석**사** : *sôg_sa* « master »]+[박**사** : *bag_sa* « doctorat »]
Les personnes qui ont les diplômes de master et de doctorat, les diplômes de master et de doctorat)

[36] 석 :碩 : *_sog* (être grand), 박 :.博 : *_bag* (être vaste), 사 . 士 : *_sa* (savant, érudit)

수• 출입[37] *su_chul_ib*
[**수**입 : *su_ib* « importation »]+[**수**출 : : *su_chul* « exportation »]
Importation et exportation

L'exemple (1b) montre l'élision de morphèmes composés avec le point au centre. Ces morphèmes coordonnés contiennent une partie commune (사 : *_sa*, 수 *_su*), comme la factorisation 'ab+ac = a(b+c)'. Ce mot composé s'écrit aussi sans le point au centre ou avec le point : 석박사, 3.1 운동, 삼일운동 : *sam_il_un_dong*.

Le signe '。' est utilisé dans le texte d'écriture verticale pour indiquer la terminaison de phrase.

[37] 수 : 輸 : *_su* (transporter), 출 : 出 : *_chul* (sortir, la sortie), 입 : 入 : *_ib* (entrer, l'entrée).

4.2 Segmentation

Après le prétraitement, la segmentation du texte a pour effet de découper l'ensemble du texte en tokens : mots, chiffres et symboles. Nous obtenons les mots dans le texte délimité par les séparateurs, blanc, symboles, chiffres.

Dans UNITEX, on peut définir les lettres d'alphabet de chaque langue dans le fichier « alphabet.txt » qui a la fonction de définir l'ensemble des caractères pour reconnaître les mots dans le texte selon la langue traitée. Pour le texte coréen, ce fichier contient la liste des syllabes, des idéogrammes et l'alphabet latin. A cause de la restriction de saisir les idéogrammes d'UNICODE avec l'éditeur de texte banal, nous mettons les 4888 idéogrammes qui sont disponibles dans le système de codage WANSUNG. La liste ci-dessous montre quelques lignes du fichier « alphabet.txt » pour la langue coréenne[38].

```
Xx
Yy
Zz
#가힣
伽가
佳가
:
稀희
戱희
詰힐
```

La ligne « #가힣[39] » signifie toutes les syllabes coréennes entre la syllabe 가:*_ga* et 힣 :*_hih*.

Dans les syllabes coréennes, il n'y a pas de distinction entre les lettres majuscule et minuscule. Mais il y a une écriture différente pour les mots sino-coréens qu'on peut écrire en idéogrammes ou en syllabes coréennes. De la même façon que pour l'alphabet latin on écrit les majuscules suivies de la minuscule, nous écrivons dans « alphabet.txt » le caractère sino-coréen suivi de l'écriture en syllabe coréenne, par exemple : « 加가 ».

---

[38] Ce fichier est disponible sous le répertoire de UNITEX/korean

[39] 가 :*_ga* est la première syllabe avec une consonne initiale 'ㄱ', une voyelle 'ㅏ' et une consonne finale vide, '힣' est la dernière syllabe avec une consonne initiale et finale 'ㅎ' et une voyelle 'ㅣ'.

La figure (4-3) montre le résultat de la recherche de la syllabe 미 :*_mi*. Grâce à cette double écriture en sino-coréen et caractère coréen, on obtient toutes les lignes qui contiennent des mots mono-syllabiques « 미 » et aussi « 美 :미 :*_mi* ».

En ce moment, la fonction de recherche des motifs est disponible avec le mot entier mais non avec les morphèmes. Nous ajouterons la fonction qui traite la recherche de motifs en morphèmes avec le texte exprimé par les phrases d'automate.

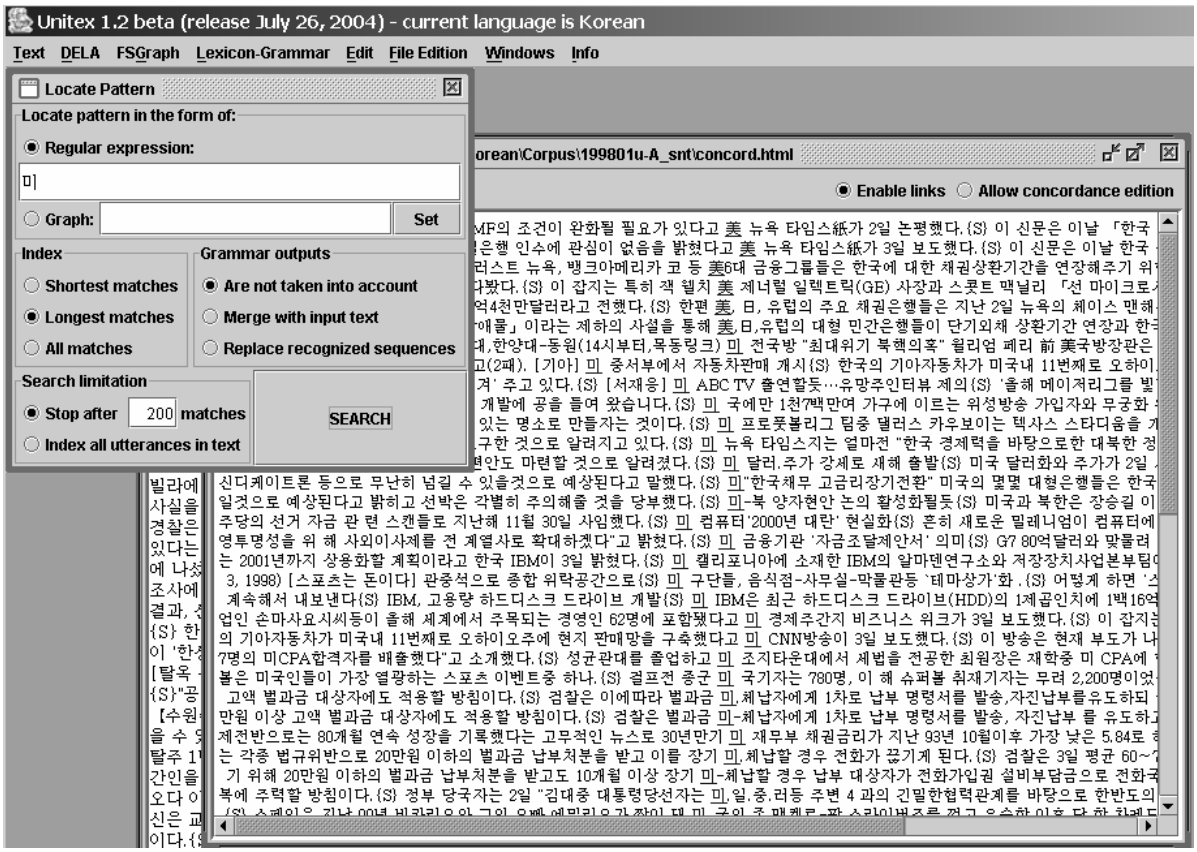


Figure 4-3 Exemple de concordance

Nous ne mettons pas les lettres de l'alphabet coréen dans ce fichier, nous considérons que les lettres de l'alphabet coréen sont dans le texte correspondent à la citation de symboles. Si l'on pense qu'ils sont des caractères des mots coréens, c'est possible de les ajouter. Dans le texte en UNICODE, les lettres de l'alphabet coréen sont codées par les lettres de la zone *Hangul Compatibility JAMO*.

Les résultats de la segmentation sont une liste de segments (4-4), un fichier des données statistiques et un fichier de la numérotation de texte par index de segments.

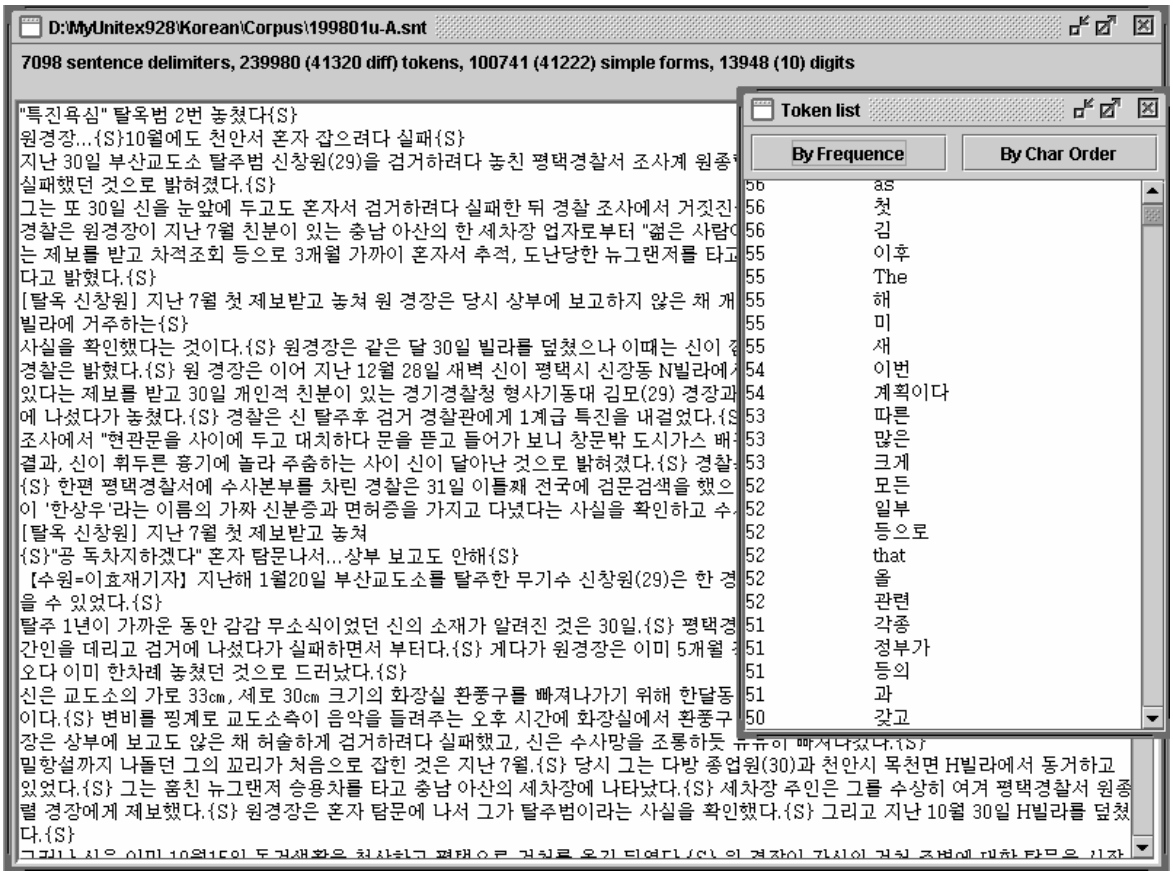


Figure 4-4 Liste de segments

4.3 Consultation des dictionnaires de mots simples

La consultation des dictionnaires sert à la reconnaissance des segments qui ne sont ni des chiffres ni des symboles. Pour les segments français, la consultation donne les informations linguistiques sur les segments mais pour les segments coréens, elle donne les informations linguistiques sur les morphèmes. Un segment est un mot coréen, donc une séquence de morphèmes.

Le dictionnaire comprimé est codé dans l'alphabet coréen qui est défini par la table de conversion de syllabes en lettres de l'alphabet coréen. Alors nous convertissons les mots coréens en *Hangul Jamo*.

Chaque segment doit être traité en lettres alphabétiques coréennes :

(2)

a. Chaîne de caractères syllabiques coréens
사실은 : *sa_sil_eun* ➔ _ㅅㅏ_ㅅㅣㄹ_ㅡㄴ 사실.N.vérité+는.Post.spc
b. Chaîne mêlant des mots sino-coréens et des syllabes coréennes.
超고속➔초고속 : *cho_go_sog* ➔ _ㅊㅗ_ㄱㅗ_ㅅㅗㄱ 초.Det.Ultra+고속.N.grande vitesse
c. Chaîne mêlant des caractères de l'alphabet latin et des syllabes coréennes
UN 은 : *yu_en_eun* ➔UN_ㅡㄴ ONU+는.Post.spc

Dans le cas de (2b), nous convertissons d'abord les caractères sino-coréens en syllabes coréennes, puis nous les convertissons dans l'alphabet coréen. Si l'on écrit un mot étranger avec les caractères de l'alphabet latin, on doit écrire le mot étranger en le séparant des suffixes par exemple (1c) « UN 은 : UN _ㅡㄴ».

Nous pouvons détecter les mots étrangers par leur alphabet. Il est nécessaire de disposer de dictionnaires de mots étrangers avec la même description que les morphèmes coréens. Les mots étrangers ont aussi la compatibilité phonétique (3). Le mot « PC : 피시 : *pi_si* » se termine par une voyelle et le mot « FM : 에프엠 : *e_feu_em* » se termine par une consonne, la postposition de nominatif -_가 : *_ga* varie selon la caractéristique de prononciation selon le coréen.

(3)

a. PC 가 집에 있다.
pi_si_ga jib_e iss_da.
[pi_si][ga] [jib][e] [iss_da][da].
ordinateur+Post.nmtp maison+Post.lieu exister+St.déc.
L'ordinateur existe à la maison.

Il y a un ordinateur à la maison.

b. 여기서는 라디오에 FM 이 잡히지 않는다.
yô_gi_sô_neun la_di_o_e e_feu_em_i jab_hi_ji anh_neun_da.
[yô_gi][e_sô][neun] [la_di_o][e] [e_peu_m][ga] [jab_da][hi][ji] [anh_da][neun][da].
ici+Post.lieu+Post.spc radio+Post.datif FM+Post.nmtp attraper+Suf.passif+Sc « ne pas faire»+Mtc+St.déc.
C'est ici que FM ne passe pas à la radio.

Nous traiterons les mots étrangers comme des racines auxquelles les morphèmes suivants peuvent se souder avec ou sans espacement.

Les symboles sont encore groupés en symboles graphiques, symboles de ponctuation, caractères de contrôle qui ont des valeurs inférieures à 0x20 (ils ne présentent pas d'affichage visible mais gèrent les séquences des caractères entre l'expéditeur et le destinataire dans la transmission de données).

La consultation du dictionnaire de séquences de morphèmes sur un segment donne les séquences des morphèmes connus. Nous sauvegardons ces résultats du découpage des segments par dictionnaires sous la forme d'une expression de chaînes de morphèmes.

« Expression de chaînes de morphèmes »

ESMorph= : (<SeqM>) | (<SeqM>)<SymSep0><ESMorph> | <ø>

SeqM= : {<EMorph>} | {<EMorph>}<SymSep1><SeqM>

EMorph = : Forme fléchie, Forme canonique (Forme originale). Informations linguistiques

SymSep0= : '+'

SymSep1= :' '

Ø : vide

Nous utilisons six symboles '+', ' ', '{', '}', '(', ')' pour exprimer les séquences de morphèmes d'un segment. Le symbole '+' est utilisé pour séparer deux séquences, le symbole ' ' signifie la concaténation des morphèmes. Chaque morphème est distingué par les symboles '{', '}', les symboles '(', ')' entourent chaque séquence. Les accolades entourent chaque morphème. Le symbole « + » signifie la disjonction de séquences de morphèmes,

l'espacement signifie la concaténation des morphèmes dans une séquence. Par exemple, la figure (4-5) montre une expression

(({올,올.N}) + ({오,오다.V+i} {ㄹ,을.Sd+fut}))

qui est : les séquences de 올 : *_ol* : «un brin de laine » ; « qui va venir» pour un segment 올 : *_ol*.

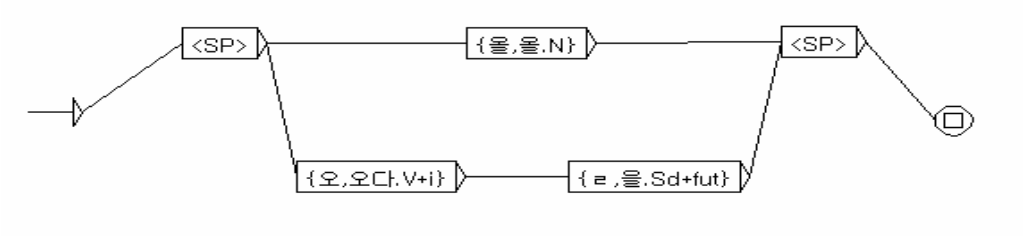


Figure 4-5 Graphe d'un mot 올 : *_ol*

Après la consultation des dictionnaires, nous distinguons les segments analysés et des segments non identifiés. Dans le fichier qui contient les expressions de chaînes des morphèmes, les lignes des segments inconnus sont vides.

La figure (4-6) montre les mots connus par la consultation du dictionnaire en séquence de morphèmes.

Nous n'avons pas encore la construction des dictionnaires des mots composés.

Figure 4-6 Résultat de la consultation de dictionnaire en morphèmes

4.4 Construction des automates des phrases

Après la consultation du dictionnaire, on peut obtenir les séquences de morphèmes des segments, un segment de mot peut être représenté par plusieurs séquences de morphèmes. Nous construirons le texte en automates de la phrase, la figure (4-7) montre l'automate d'une phrase.

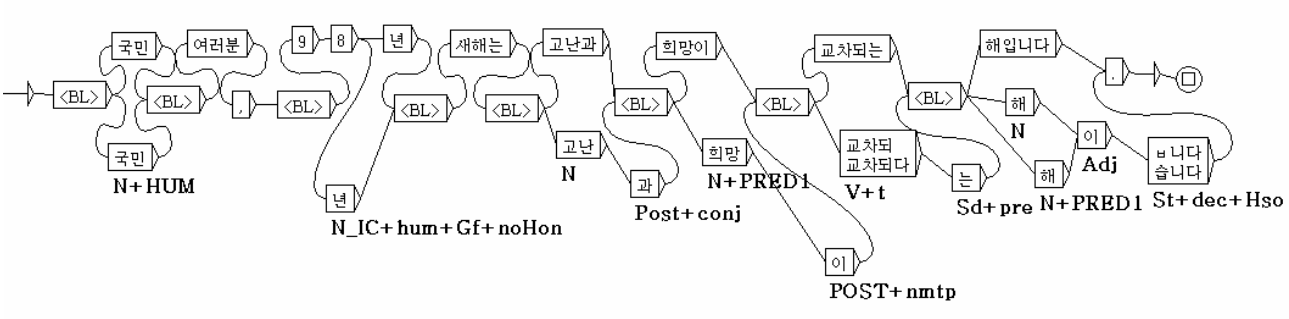


Figure 4-7 Une phrase représentée par un automate

Dans UNITEX, dans le cas des langues qui sont traitées avec un mot comme unité, on n'a pas besoin d'indiquer l'espacement entre les mots, il suffit de présenter une boîte qui contient le mot. Mais nous traitons les mots coréens au niveau de morphème et nous utilisons le symbole «<BL> » pour représenter l'espacement entre les mots.

La figure (4-7) montre une phrase. Les boîtes de segments contiennent les formes de surface des mots sans informations linguistiques. Les boîtes de morphèmes contiennent les morphèmes avec les informations linguistiques. Dans les boîtes de morphème qui ont deux lignes, la première ligne indique la forme fléchie et la deuxième est la forme canonique.

Un mot inconnu est représenté par une seule boîte de segment.

Sur l'ensemble d'un texte on obtient des fichiers (avec extension « .fst2 ») de grande taille. C'est pour cela que nous sauvegardons ces données sous une autre forme :

Structure de l'état de l'automate de texte.

- **type de segment**
- **index de segment**

- **pointeur de l'état suivant**

L'élément de la structure « index de segment » est l'index du segment dans le texte de base. Nous l'utilisons pour identifier le segment dans le texte de base. Nous classifions les éléments dans l'automate de la phrase selon la liste suivante.

- segment original
- segment variante d'écriture : écriture mélangée : syllabes et idéogrammes
- morphème simple
- morphème composé
- chiffre
- symbole contrôle
- symbole graphique

Quand on traite une phrase spécifiée, on génère un fichier de type « .fst2 » d'UNITEX.

A ce stade du traitement on n'a pas détecté toutes les séquences de morphèmes dans les mots, on doit compléter les lexiques. Puisque la levée d'ambiguïtés n'a pas été faite il y a du bruit. Cependant, les résultats sont plus précis qu'avec les techniques classiques, et les données sont plus lisibles qu'avec la morphologie à deux niveaux.

4.5 Recherche de motifs sur les automates des phrases

Cette partie constitue un travail prospectif pour obtenir la structure de la phrase coréenne en utilisant les automates des phrases. Nous allons utiliser ces données pour l'analyse de la structure des phrases et de la morphologie coréennes.

Le système UNITEX permet d'appliquer les grammaires locales aux textes étiquetés lexicalement. La figure (4-8) montre un automate qui reconnaît une séquence de mots

(2) 꽃을 찾는다.
ggoch_eul chaj_neun_da.
[ggoch][eul] [chaj_da][neun][da].
fleur+Post.accu chercher+Mtc+St.déc.

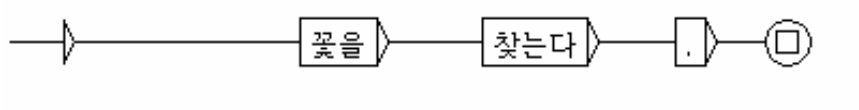


Figure 4-8 Reconnaissance d'une séquence de deux mots.

Avec ce graphe, on peut seulement trouver des phrases qui contiennent les deux mots. Avec le graphe (4-9), on peut trouver toutes les phrases qui contiennent la structure. Ce graphe décrit une séquence : tous les morphèmes étiquetés par « N : nom », les postpositions étiquetées avec « accu : accusatif » et ensuite, les racines verbales : « V » et les suffixes étiquetés avec « Morph », les suffixes étiquetés « St : Suffixe terminal » et le point de la phrase.

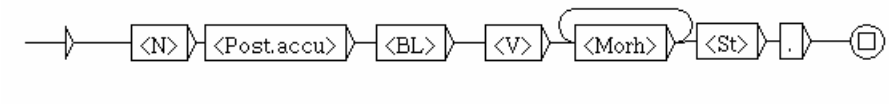


Figure 4-9 Reconnaissance d'une séquence grammaticale.

Nous traiterons les textes sous forme de transducteur étiqueté lexicalement.

Nous allons présenter quelques graphes utilisables pour reconnaître les phrases simples comparatives qui sont construites avec le verbe 못하다 : *mos_ha_da* « ne pas bien faire , ne pas pouvoir faire, ne pas être bien » et la postposition de ressemblance 만큼 : -

_man_keum « autant que » ou la postposition de différence 보다 :-*_bo_da* « plus que »(3).

(3)

a. 그는 너보다 일을 못 한다.
geu_neun nô_bo_da il_eul mos han_da.
[geu][neun] [nô][bo_da] [il][leul] [mos] [ha_da][neun][da].
Lui+Post.nmtp toi+Post.compare travail+Post.accu mal faire+Mtc+St.déc.
Il fait le travail plus mal que toi.
Il ne travaille pas mieux que toi.
= 그는 너보다 일을 못한다.

Lui+Post.nmtp toi+Post.compare travail+Post.accu « mal faire »+Mtc+St.déc.
b. 그는 너보다 일을 잘 한다.
geu_neun nô_bo_da il_eul jal han_da.
[geu][neun] [nô][bo_da] [il][leul] [jal] [ha_da][neun][da].
Lui+Post.nmtp toi+Post.compare travail+Post.accu «bien faire »+ Mtc+St.déc.
Il travaille mieux que toi.

L'expression des morphèmes avec les symboles ' <' et '>' indique la forme canonique ou d'autres informations contenues dans le lexique.

<못> : trouver les morphèmes qui ont la forme canonique 못 :*_mos* « bassin, clou, cor au pied ou à la main, ne pas ».
<ADV> : trouver tous les morphèmes qui ont l'étiquette 'ADV : adverbe'.
<못.ADV> : trouver les morphèmes 못 :*_mos* qui ont l'étiquette 'ADV'.
<거,것> : trouver les morphèmes qui ont la forme fléchie 거 :*_gô* et la forme canonique 것 :*_gôs*'.

Sans symboles, l'expression signifie trouver tous les morphèmes qui sont sous la forme fléchie.

Le sous-graphe « N0.grf » traite la reconnaissance du sujet et montre les pronoms personnels avec les propositions nominatives.

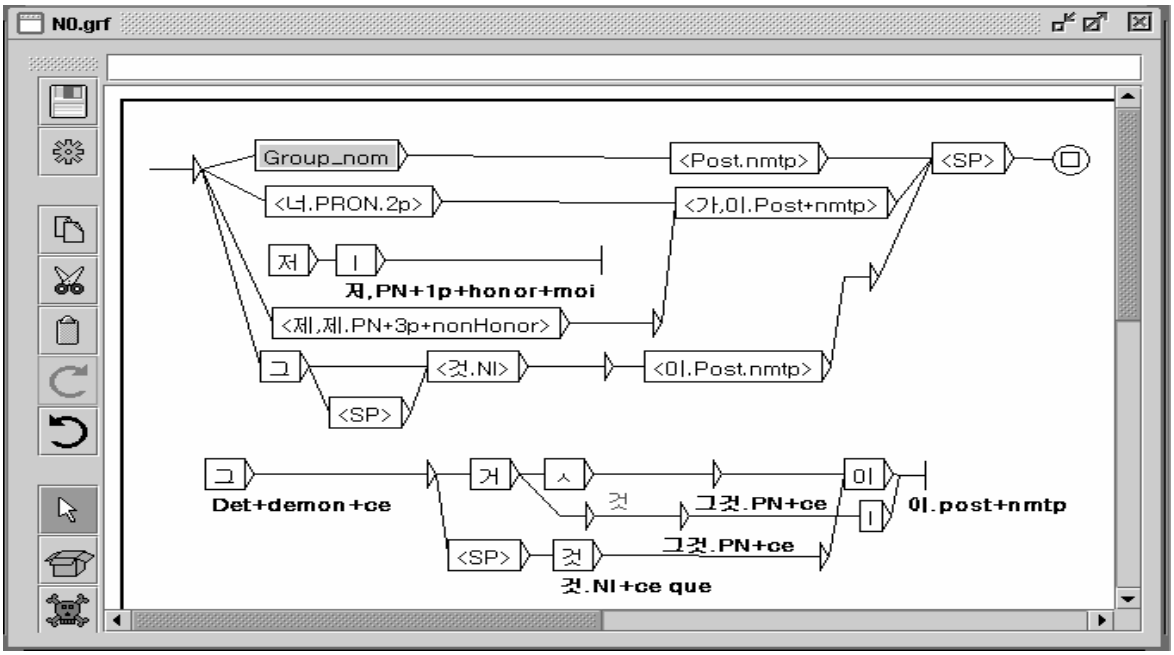


Figure 4-10 Graphe de sujet de la phrase

Le sous-graphe « N1.grf » traite la reconnaissance du groupe nominal avec la postposition accusative.

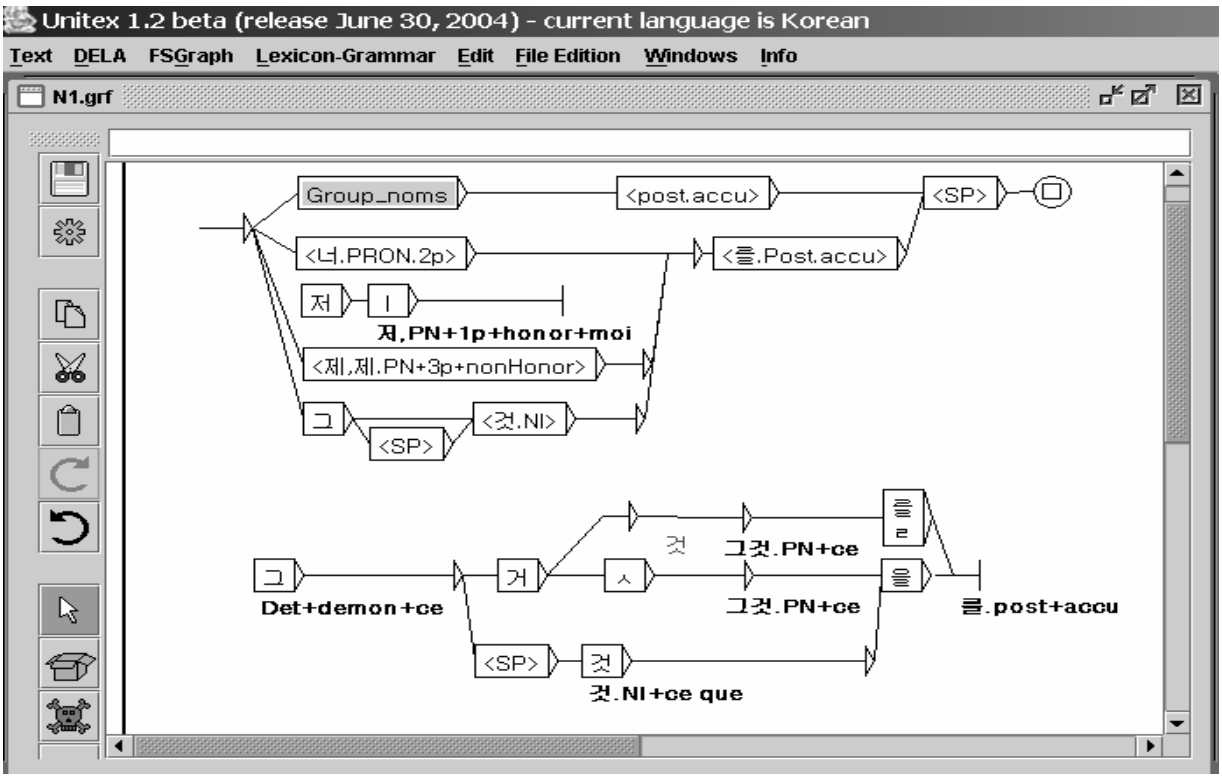


Figure 4-11 Graphe du complément à l'accusatif

Dans la figure (4-12), le graphe « NCOMP.grf » montre le nom ou le groupe de noms avec la postposition de la comparaison pour l'objet de la comparaison. Les chemins au-dessous de la ligne reconnaissent les mêmes postpositions mais pas leurs formes fléchies.

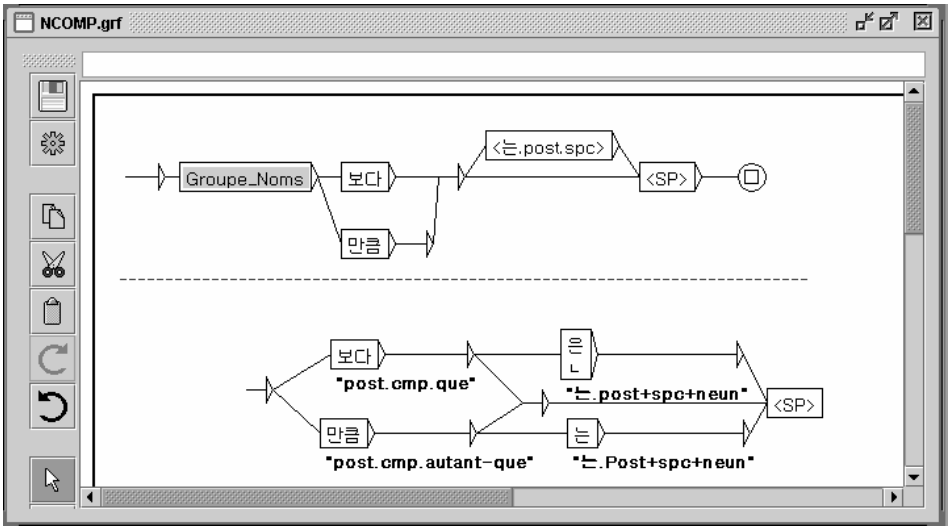


Figure 4-12 Graphe de l'objet de la comparaison

Dans la phrase, les adverbes sont soit en position libre, soit en position non libre. Nous utilisons deux sous graphes pour les adverbes. Le graphe (4-13) « Gp_ADV.grf » montre les adverbes libres comme le complément de circonstance de lieu, de temps, de raison, de condition de cause.

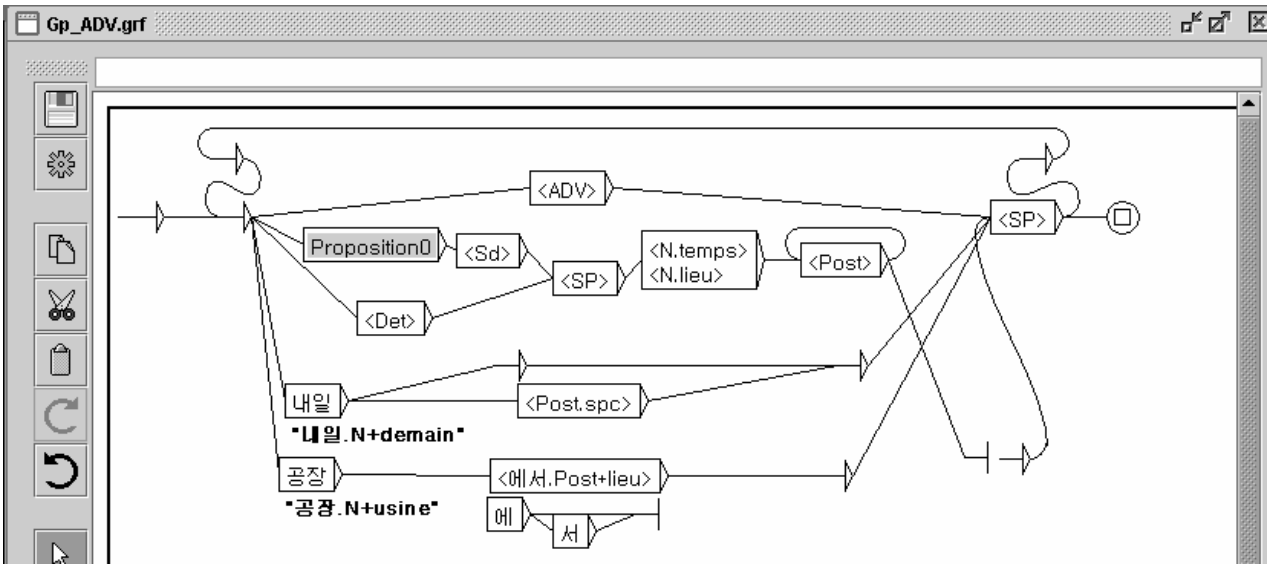


Figure 4-13 Graphe de l'adverbe

Et le graphe (4-14) « ADVS_mos_ha_da.grf » montre les adverbes qui se situent avant des verbes et adjectifs.

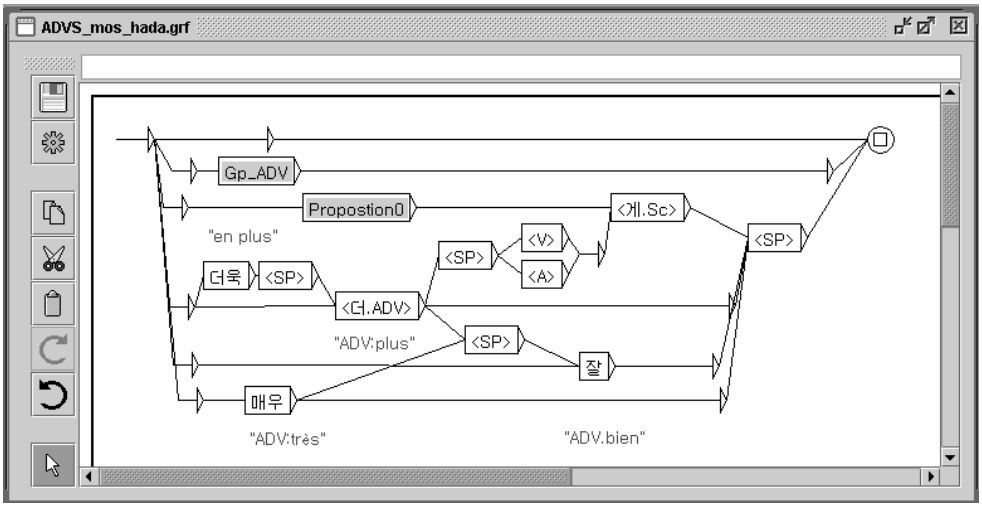


Figure 4-14 Graphe des adverbes avant le mot « *mos_ha_da* »

Les graphes de la description de la structure de la phrase vont être utilisés pour reconnaître et analyser les structures de la phrase coréenne. Cette direction de recherches nécessite la réalisation de nombreuses tâches :

En linguistique,

- Extension des lexiques
- Classification morphologique en détail.
- Description au fur et à mesure des structures de la phrase.

En informatique,

- Traitement d'automates de morphèmes pour recherche de motifs.
- Traitement de séquences de morphèmes figées avec espacement.
- Traitement de séquences de morphèmes lexicaux libres sans espacement.
- Amélioration de la description des variations morphologiques.

Conclusion

Nous avons introduit une méthode de description morphologique, syntaxique et sémantique des mots coréens par automates finis.

Nous donnons un outil de transcodage entre syllabes et alphabet coréen sur UNICODE. Cet outil élimine la contrainte de décrire uniquement au niveau des phonèmes ou uniquement au niveau des syllabes pour l'utilisateur.

Le traitement du texte coréen nécessite une description des morphèmes qui composent les mots. Nous avons adapté les transducteurs comme outil de description des séquences de morphèmes coréens. Chaque morphème de ces séquences reçoit des informations linguistiques et sa compatibilité avec les autres morphèmes est exprimée par les transitions.

Nous décrivons ainsi les relations entre les morphèmes d'un mot coréen. Cette méthode ne nécessite pas de mettre au point des règles séparées ni de recourir à des approximations statistiques.

Nous avons ajouté dans les dictionnaires de racines les informations de compatibilités qui sont des liens vers les transducteurs de séquences de suffixes. Notre classification des racines et des suffixes est fondée sur la phonétique et la morphologie. Les séquences de suffixes sont exprimées par les graphes de transducteur grâce à l'éditeur graphique de UNITEX. Les sorties présentent les informations. Les transitions montrent naturellement l'ordre et la compatibilité entre les morphèmes.

Nous construisons le dictionnaire comprimé des séquences de morphèmes pour les mots coréens. Les textes sont mis sous la forme d'automates de morphèmes, visualisés comme graphes.

Ces automates de morphèmes sont utilisables pour l'analyse de la structure de la phrase coréenne par recherche de motifs.

Nous avons montré que :

- Les automates finis utilisés pour les grammaires locales de mots peuvent être aussi utilisés pour des grammaires de morphèmes : la description par transducteur exprime bien les informations linguistiques sur la syntaxe et la sémantique des morphèmes.

- La description des informations linguistiques sous forme d'automate est compatible avec la construction d'un dictionnaire de mots dans une langue agglutinante.

- Les fonctionnalités d'UNITEX facilitent la gestion de ces données linguistiques et permettent l'analyse du texte.

Ces données linguistiques sont précises, ne font pas appel à des approximations statistiques et leur contenu peut être modifié directement, ce qui permet de contrôler directement les performances du système d'analyse morphologique.

Notre méthode est donc de nature à faire progresser le traitement automatique du texte coréen et d'autres langues agglutinantes.

Bibliographie

1. BAE, Joo-Chai, 2003, 한국어의 발음 : *han_gug_ô_ôi bal_eum* : la prononciation du Coréen, Séoul, Sam_Kyeong_Mun_Hoa_Sa.
2. BAE, Sun-Mee, 2002, *Le Dictionnaire électronique des Séquences Nominales figées en coréen et de leurs formes fléchies, méthodes et application*, thèse de doctorat, Université de Marne-la-Vallée.
3. BERLOCHER Ivan, 2004, *manual de DECOTEX*.
4. BIRD, S., LOPER, E., 2002, *NLTK: The Natural Language Toolkit*, Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia.
5. BRILL, E. , 1995, *Transformation-based Error-driven Learning and natural language processing : A case study in part-of-speech tagging*, Computational Linguistics, 21-4, pp543-565.
6. CHA J.W., LEE G.B, LEE J.H., 1998, *Generalized Unknown Morpheme Guessing for Hybrid POS Tagging in Korean*, Proc. of Workshop on Very Large Corpora, Montréal, pp85-93.
7. CHOI, Sung-Woo, 1999, *Implantation de dictionnaires électroniques du coréen par automates finis* ; thèse de doctorat, Université de Marne-la-Vallée, IGM.
8. CUNNINGHAM, H, 2002, *GATE, a general architecture for text engineering*. Computers and the Humanities 36 ; pp223-254
9. GROSS, Maurice, 1984, *Lexicon-Gramar and the Syntactic Analysis of French*, Association for Computational Linguistics, Morristownn, NJ, USA, pp 275-281.
10. GROSS, Maurice, 1987, *The Use of Finite Automate in The Lexical Representation of natural language*, Electronic Dictionaries and Automata in Computational linguistics, Springer-Verlag, 1987, pp 34-50.
11. GROSS, Maurice, 1997, *The Construction of Local Grammars*, Finite-State Language Processing, MIT Press, 1997, pp 329-354.
12. HAN, Sun-Hae, 2000, *Les Prédicats Nominaux en coréen –Construction à Verbe Support Hada* -, thèse de doctorat, Paris : Université de Paris 7, pp30-34.
13. HAN,C.H, HAN N.R. et KO E.S., *Development and Evaluation of a Korean Treebank and its Application to NLP*, Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002).
14. HONG, Y., KOO, M.W., YANG, G.J., 1996, *A Korean morphological analyzer for*

speech translation system, In ICSLP-1996, pp673-676.

15. KANG, Seung-Shik, 1993, *Syllable-based Morpheme Isolation and Morphological Alternation of Korean Words*, NLPRS'93, Fukuoka, Japan, pp.312-318.
16. KARTTUNEN, Lauri and BEESLEY Kenneth R. 1992. *Two-level Rule Compiler*. Technical Report. ISTL-92-2. Xerox PARC. June 1992. Palo Alto, California.
17. KARTTUNEN, Lauri, 1996. *Directed Replacement*. In *The Proceedings of the 34rd Annual Meeting of the Association for Computational Linguistics. ACL-96*, Santa Cruz, California.
18. KIM D.B., LEE S.J., CHOI K.S., KIM G.Ch., 1994, *A Two-Level Morphological Analysis of Korean*, *COLING*, vol. 1.
19. LAPORTE, Eric, 1988, *Méthode algorithmique et lexicales de phonétisation de textes*, thèse de doctorant, Université Paris 7.
20. LAPORTE, Eric., 2001, *Reduction of lexical ambiguity*. Lingvisticae Investigationes 24:1, pp67-103.
21. LEE, Chang-Yeol, 1997, *La construction de lexiques de formes fléchies et l'analyse morphologique du coréen*, thèse de doctorat, Université PARIS 7.
22. LEE, Eeun-Kyeong, 2000, *Gug_ô_wi yôn_gyô_ô_mi yôn_gu : recherche des suffixes conjonctives du coréen*. Bae_Hag_Sa, Séoul.
23. LEE, Eun-Chul, 1993, *An Improved Method on Korean Morphological Analysis Based on CYK Algorithm*, A Master thesis in Dept. Of Computer Science, POSTECH, Korea.
24. LEE G.B., LEE J.H. et Yoo J.H., 1997, *Multi-level Post-Processing for Korean Character Recognition Using Morphological Analysis and Linguistic Evaluation*", Pattern Recognition 30(8): pp1347-1360.
25. LEE G.B., CHA J.W. et LEE J.H. 1997, *Hybrid POS tagging with generalized unknown-word handling*, International workshop on information retrieval with Asian languages (IRAL), Tsukuba-City, Japan, pp43-50.
26. LEE G.B., CHA J.W., LEE J.H., 2002, *Syllable pattern-based unknown morpheme estimation for hybrid part-of-speech tagging of Korean*, Computational Linguistics 28:1 53-70
27. LEE G.B., LEE J.H., KIM B.Ch. et LEE Y.J., 1997, *A Viterbi-based morphological anlaysis for speech and natural language integration*, Proceedings of the 17th international conference on computer processing of Oriental languages (ICCPOL),

Hong-Kong,pp133-138

28. LEE S.Z., TSHJII J.I, RIM H.C., 2000, *Lexicalized Hidden Markov Models for Part-of-Speech Tagging*, COLING2000, pp481-487.
29. LEE, Jin-Mieung, *Grammaire du Coréen*, P.A.F, 1985, Tome 1.
30. LUCCHESI, C., KOWALTOWSKI, T., 1993, *Applications of Finite Automata Representing Large Vocabularies*. Software - Practice and Experience 23:1, Wiley & Sons, pp15-30.
31. MERIALDO, B., 1994, *Tagging English text with a probabilistic model*. Computational linguistics 20:2, pp155-171
32. NAM, Jee-Sun, 1994, *Classification Syntaxique des Constructions Adjectivales en coréen*, thèse de doctorat, Université Paris 7.
33. NAM, Jee-Sun, 1994, *Dictionnaire des Noms Simples du Coréen, rapport technique*, No 46, LADL, Université Paris 7.
34. NAM, Jee-Sun, 1997, *Système électronique de lexiques coréens DECO*, Mémoire du diplôme d'habitation, IGM, Université Marne-la-Vallée.
35. NAM, Jee-Sun, 1996, *Building a Korean on-line dictionary adequate to parsing systems*, Rapport Technique N-50, LADL, Université Paris 7.
36. NAM, Jee-Sun, 1996, *Dictionary of N-Postpositons, A-Postpositions and V-Postpositions in Korean*, Rapport Technique N-51, LADL, Université Paris 7.
37. MOHRI, Mehryar, 1997, *Finite-State Transducers in Language and Speech Processing*, Computational Linguistics, V 23 I 2, pp269 – 311.
38. PARK J.S., KANG J.G., HUR W., CHOI K.S., 1998, *Machine Aided Error-Correction Environment for Korean Morphological Analysis and Part-of-Speech Tagging*, COLING-ACL pp1015-1019.
39. PAUMIER Sébastien, 2003, *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, Thèse de doctorat, Université de Marne-la-Vallée.
40. REVUZ, Dominique, 1991, *Dictionnaires et Lexiques, Méthodes et algorithmes*, thèse de doctorat, Université Paris 7.
41. SGARBAS K., FAKOTAKIS N., KOKKINAKIS G., 1995, *Two Algorithms for Incremental Construction of Directed Acyclic Word Graphs*, International Journal on Artificial Intelligence Tools, World Scientific, Vol.4, No.3, pp.369-381.
42. SILBERZTEIN, Max, 1993. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*, Masson, Paris.

43. SILBERZTEIN, Max, 1997, *The Lexical Analysis of Natural Language*, Finite-State Language Processing, MIT Press, pp 175-203.
44. SILBERZTEIN, Max, 1999, INTEX, LADL, Université de Paris 7.
45. SHIN J.H., HAN Y.S., PARK Y.C., CHOI K.S., 1995, *A HMM Part-of-Speech Tagger for Korean with Wordphrasal Relations*, Recent Advances in Natural Language Processing, Sofia.
46. SIN, Gi-Cheul, SIN Yong-Cheul, 1988, "큰사전 :*Keun-Sa-Jeon* : *Le grand dictionnaire* ", Press Sam-Sung.
47. SOH, Cheong-Soo, 1996, "국어문법 :*Gug-ô-Mun-Bob* : *Korean Garmmar*", Press de Université HanYang, Korea.
48. VOUTILAINEN Atro, 1995, A syntax-based part-of-speech analyser,EACL,pp157-164.
49. WOODS W. A., 1970, *Transition network grammars for natual language analysis*, Communications of the ACM, Volume 13 Issue 10.
50. The National Academy of the Korean Langage, 1999, "표준국어 대사전 :*Pyo-Jun-Gug-E Dai-Sa-Jen* : *Le grand dictionnaire de langue coréenne standard*", Press Doo-San.
51. Dong-A, 2002"백년옥편 *Bag-Nyôn-og-pyôn* : *le dictionnaire d'idéogramme de bag_nyôn*", Press Dong-A.

Sites web consultés

IGM-Equipe d'Informatique linguistique. 01 Jan 2002, http://infolingu.univ-mlv.fr/

Unicode Home Page. 1998, http ::://www.unicode.org

Finite State Techonology -Xerox XRCE. 03 Feb 2003,
http://www.xrce.xerox.com/competencies/content-analysis/fst/

Penn Treebank project, 07 Sept 2003, http://www.cis.upenn.edu/~treebank/home.html

KLEX, 02, Jan 2005, http://www.cis.upenn.edu/~nrh/klex.html

한국어 언어 학회 *han_gug_ô ôn_ô hag_hoi*(association de la langue du coréen),
http://www.linguistics.or.kr/

한글 학회 *han_geul hag_hoi* (association de Hangeul), http://www.hangeul.or.kr/

21centrury project, 4 Mai 2000, http:://www.sejong.or.kr

Annexes

Annexe 1 Alphabet coréen contemporain

Tableau A Voyelles

Prononciation des voyelles de l'alphabet coréen

l'ordre dans le dictinnaire		Nom du coréen	Prononciation	romanisation		phonétique		nombre des éléments		Transcription	Caractèriq ue de YinYang*
				1984	2000	simple	double	simple	composée		
1	ㅏ	아	a	a	a	+	-	+	-	**a**	Y
2	ㅐ	애	æ	ae	ae	+	-	-	+	**ai**	Y
3	ㅑ	야	ya	ya	ya	+	-	-	+	**ya**	Y
4	ㅒ	얘	yai	yae	yae	+	-	-	+	**yai**	Y
5	ㅓ	어	ʌ	ŏ	eo	+	-	+	-	**ô**	J
6	ㅔ	에	e	e	e	+	-	-	+	**e**	J
7	ㅕ	여	yʌ	yŏ	yeo	+	-	-	+	**yô**	J
8	ㅖ	예	ye	ye	ye	-	+	-	+	**yôi**	J
9	ㅗ	오	o	o	o	+	-	+	-	**o**	Y
10	ㅘ	와	wa	wa	wa	-	+	-	+	**oa**	Y
11	ㅙ	왜	wai	wae	wae	-	+	-	+	**oiô**	Y
12	ㅚ	외	oe	oe	oe	-	+	-	+	**oi**	Y
13	ㅛ	요	yo	yo	yo	+	-	-	+	**yo**	Y
14	ㅜ	우	u	u	u	+	-	+	-	**u**	J
15	ㅝ	워	wɔ	wo	wo	-	+	-	+	**uô**	J
16	ㅞ	웨	we	we	we	-	+	-	+	**uôi**	J
17	ㅟ	위	wi	wi	wi	-	+	-	+	**ui**	J
18	ㅠ	유	yu	yu	yu	+	-	-	+	**yu**	J
19	ㅡ	으	ə	ŭ	eu	+	-	+	-	**eu**	J**
20	ㅢ	의	əi	ui	ui	-	+	-	+	**eui**	J**
21	ㅣ	이	i	i	i	+	-	+	-	**i**	J**

*. Y : Yang ： 양성모음(陽性 母音)*yang_sông_mo_eum* : le son léger et éclairé
J : Yin ： 음성모음(陰性 母音) *eum_sông mo_eum* : le son lourd et sombre
**. Voyelles neutres

Tableau B Consonnes

prononciation des consonnes de l'alphabet coréen

simple	double	composée	nom coréen	transcription latine du nom	romanisation 1984	romanisation 2000	valeur de son : la position dans la syllabe : initial	valeur de son : la position dans la syllabe : final	**notre transcription latine**
ㄱ			기역	gi_yeuk	k/g	g/k	g	g	**g**
ㄴ			니은	ni_eun	n	n	n	n	**n**
ㄷ			디귿	di_geut	d	d/t	d	d	**d**
ㄹ			리을	ri_eul	r/l	r/l	l	l	**l**
ㅁ			미음	mi_eum	m	m	m	m	**m**
ㅂ			비읍	bi_eup	p/b	b/p	b	b	**b**
ㅅ			시옷	si_os	s	s	s	d	**s**
ㅇ			이응	i_eung	ng	ng	NULL	~	**ng**
ㅈ			지읒	ji_euj	ch/j	j	j	d	**j**
ㅊ			치읓	chi_euch	ch'	ch	ch	d	**ch**
ㅋ			키읔	ki_euk	k'	k	kh	g	**k**
ㅌ			티읕	ti_eut	t'	tt	th	d	**t**
ㅍ			피읖	pi_eup	p'	pp	ph	b	**p**
ㅎ			히읗	hi_euh	h	h	h	NULL	**h**
	ㄲ		쌍기역	ssang_gi_yeuk	kk	kk	gg	g	**gg**
	ㄸ		쌍디귿	ssang_di_geut	tt	tt	tt	-	**dd**
	ㅃ		쌍비읍	ssang_bi_eup	pp	pp	bb	-	**bb**
	ㅆ		쌍시옷	ssang_si_ous	ss	ss	ss	d	**ss**
	ㅉ		쌍지읒	ssang_ji_euj	tch	jj	jj	-	**jj**
		ㄳ	기역시옷	gi_yeuk_si_os	gs	gs	-	g	**gs**
		ㄵ	니은지읒	ni_eun_ji_euch	nj	nj	-	n	**nj**
		ㄶ	니은히읗	ni_eun_hi_euh	nh	nh	-	n	**nh**
		ㄺ	리을기역	ri_eul_gi_yeuk	lg	lg	-	g	**lg**
		ㄻ	리을미음	ri_eul_mi_eum	lm	lm	-	m	**lm**
		ㄼ	리을비읍	ri_eul_bi_eub	lb	lb	-	l	**lb**
		ㄽ	리을시읏	ri_eul_si_os	ls	ls	-	l	**ls**
		ㄾ	리을티읕	ri_eul_ti_eut	lt'	lt	-	b	**lt**
		ㄿ	리을피읖	ri_eul_pi_eup	lp'	lp	-	l	**lp**
		ㅀ	리을히읗	ri_eul_hi_euh	lh	lh	-	l	**lh**
		ㅄ	비읍시옷	bi_eub_si_os	bs	bs	-	b	**bs**
*			vide				-	NULL	**<E>**
							19	28	**31**

*. Le cas où la syllabe n'a pas de consonne finale.

-. Sans élément correspondant

Tableau C Enchaînement entre Consonnes initiales et finales

simple	double	compos	enchâinement					
			cas 1		cas 2		cas 3	
			CFAV	CIAF vide(=ㅇ)	CFAV	CIAF =ㅎ	CFAV	CIAF !=ㅎ
ㄱ			NULL	g	NULL	RH	g	R0
ㄴ			NULL	n	NULL	h	n	R1
ㄷ			NULL	d,R3	NULL	RH,R3	d	R0
ㄹ			NULL	l	NULL	h	l	RL
ㅁ			NULL	m	NULL	h	m	R1
ㅂ			NULL	b	NULL	RH	b	R1
ㅅ			NULL	d	NULL	RH	d	R0
ㅇ			~	-	~	h	ng	-
ㅈ			NULL	j	NULL	RH	d	R0
ㅊ			NULL	ch	NULL	RH	d	R0
ㅋ			NULL	kh	NULL	h	g	R0
ㅌ			NULL	th,R3	NULL	RH	d	R0
ㅍ			NULL	ph	NULL	h	b	R0
ㅎ			NULL	-	NULL	h	NULL	R2
	ㄲ		NULL	gg	g	RH	g	R0
	ㅆ		NULL	ss	d	RH	d	R0
		ㄱㅅ	g	s	g	RH	g	R0
		ㄴㅈ	n	j	n	h	n	R1
		ㄴㅎ	n	NULL	n	h	n	R2
		ㄹㅅ	l	s	l	RH	l	R1
		ㄹㅌ	l	th	b	h	b	R2
		ㄹㅍ	l	ph	l	h	l	R2
		ㄹㅎ	l	NULL	l	h	l	R2
		ㅂㅅ	b	s	b	RH	b	R0
vide			NULL	-	NULL	-	NULL	-

CFAV : Consonne Finale de la première de deux syllabe soudées
CIAF : Consonne Initiale de la deuxième de deux syllabe soudées
Les actions entre consonne finales et consonne initiales [40]

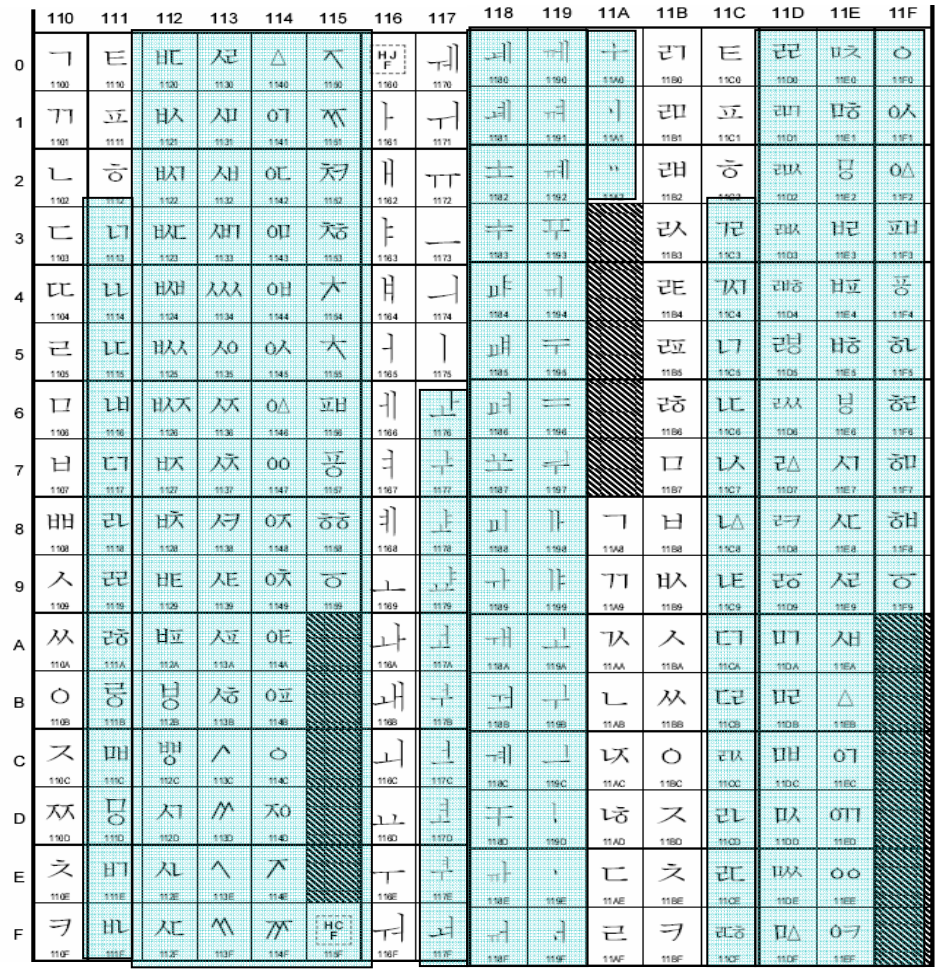
	Action	F(CFAV,CIAF)		
RH	occlusive	f('h', Cx)=f(Cx, 'h')➔f(0 ;Y)	Cx= g,d,b,j Y = k, t,p,ch	낳다[**나타**] : *nah_da* [**na_ta**] (donner naissance)
R1	Glottale	f(Cx, Cy)➔ f(Cx ; Cyy)	Cx=g,d,b Cy=g,d,b,s,j Cyy=gg,dd,bb,ss,jj	늑대[**늑때**] : *neug_dai*[**neug_ddai**] (loup)
R3	palatale	f(Cx, Ci)➔f(0, Cii)	Cx=d,t ; Ci=_i; Ci =j, ch	닫이다[**다치다**] :*dad_i_da*[**da_ji_da**](être fermé)
R4	nasale	f('g', Cn)➔f('ng', Cn) f(Cl, Cn)➔f('m', Cn) f(Cp, Cn)➔f('n', Cn),	Cn = m, n Cl=b, ph Cp=d,s,j,ch ; t ;h, ss	먹는다[**멍는다**] : *môg_neun_da*[**mông_neun_da**] (manger) 십리[**심리**] : *sib_li* [**sim_li**] (4 kilo métre) 꽃말[**꼰말**] : ggoch_mal[**ggon_mal**] (parole de fleur)
R5	latérale	f('n' ; 'l')=f('l', 'n')➔f('l' ; 'l')		난리[**날리**] : nan_li[**nal_li**] (trouble)

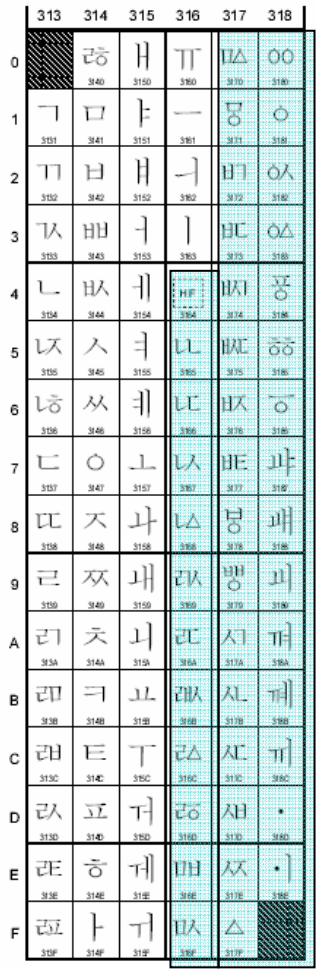
[40] http://www.korean.go.kr/

Annexe 2 Deux zones de l'alphabet coréen dans UNICODE

HANGUL JAMO

HANGUL Compatibility JAMO





Zone de lettres alphabétiques coréennes anciennes

0x1100 : le début de la partie des consonnes initiales
0x1160 : le début de la partie des voyelles
0x11A8 : le début de la partie des consonnes finales

0x3130 : des consonnes
0x314F : des voyelles

Annexe 3 Tableaux de transcodage des syllabes vers alphabets

Tableau A31 Conversion de syllabes en lettres alphabétiques coréennes fondamentales

map of Korean alphabet in the UNICODE
<SS> 0x318D # boundary of syllable
Initial_Consonants
0x1100 ㄱ 0x1100 #0x3131
0x1101 ㄲ 0x1100 0x1100 #0x3132
0x1102 ㄴ 0x1102 #0x3134
0x1103 ㄷ 0x1103 #0x3137
0x1104 ㄸ 0x1103 0x1103 #0x3138
0x1105 ㄹ 0x1105 #0x3139
0x1106 ㅁ 0x1106 #0x3141
0x1107 ㅂ 0x1107 #0x3142
0x1108 ㅃ 0x1107 0x1107 #0x3143
0x1109 ㅅ 0x1109 #0x3145
0x110A ㅆ 0x1109 0x1109 #0x3146
0x110B ㅇ #0x3147
0x110C ㅈ 0x110C #0x3148
0x110D ㅉ 0x110C 0x110C #0x3149
0x110E ㅊ 0x1112 0x110C #0x314A
0x110F ㅋ 0x1112 0x1100 #0x314B
0x1110 ㅌ 0x1112 0x1103 #0x314C
0x1111 ㅍ 0x1112 0x1107 #0x314D
0x1112 ㅎ 0x1112 #0x314E
Vowels
0x1161 ㅏ 0x1161 #0x314F
0x1162 ㅐ 0x1161 0x1175 #0x3150
0x1163 ㅑ 0x1175 0x1161 #0x3151
0x1164 ㅒ 0x1175 0x1161 0x1175 #0x3152
0x1165 ㅓ 0x1165 #0x3153
0x1166 ㅔ 0x1165 0x1175 #0x3154
0x1167 ㅕ 0x1175 0x1165 #0x3155
0x1168 ㅖ 0x1175 0x1165 0x1175 #0x3156
0x1169 ㅗ 0x1169 #0x3157
0x116A ㅘ 0x1169 0x1161 #0x3158
0x116B ㅙ 0x1169 0x1175 0x1165 #0x3159
0x116C ㅚ 0x1169 0x1175 #0x315A
0x116D ㅛ 0x1175 0x1169 #0x315B
0x116E ㅜ 0x116E #0x315C
0x116F ㅝ 0x116E 0x1165 #0x315D
0x1170 ㅞ 0x116E 0x1165 0x1175 #0x315E
0x1171 ㅟ 0x116E 0x1175 #0x315F
0x1172 ㅠ 0x1175 0x116E #0x3160
0x1173 ㅡ 0x1173 #0x3161
0x1174 ㅢ 0x1173 0x1175 #0x3162
0x1175 ㅣ 0x1175 #0x3163
Final_Consonants

FCNULL	-		#the final consonant is empty
0x11A8 ᆨ	0x1100		#0x3131
0x11A9 ᆩ	0x1100	0x1100	#0x3132
0x11AA ᆪ	0x1100	0x1109	#0x3133
0x11AB ᆫ	0x1102		#0x3134
0x11AC ᆬ	0x1102	0x110C	#0x3135
0x11AD ᆭ	0x1102	0x1112	#0x3136
0x11AE ᆮ	0x1103		#0x3137
0x11AF ᆯ	0x1105		#0x3139
0x11B0 ᆰ	0x1105	0x1100	#0x313A
0x11B1 ᆱ	0x1105	0x1106	#0x313B
0x11B2 ᆲ	0x1105	0x1107	#0x313C
0x11B3 ᆳ	0x1105	0x1109	#0x313D
0x11B4 ᆴ	0x1105	0x1110	#0x313E
0x11B5 ᆵ	0x1105	0x1111	#0x313F
0x11B6 ᆶ	0x1105	0x1112	#0x3140
0x11B7 ᆷ	0x1106		#0x3141
0x11B8 ᆸ	0x1107		#0x3142
0x11B9 ᆹ	0x1107	0x1109	#0x3144
0x11BA ᆺ	0x1109		#0x3145
0x11BB ᆻ	0x1109	0x1109	#0x3146
0x11BC ᆼ	0x110B		#0x3147
0x11BD ᆽ	0x110C		#0x3148
0x11BE ᆾ	0x110E		#0x314A
0x11BF ᆿ	0x110F		#0x314B
0x11C0 ᇀ	0x1110		#0x314C
0x11C1 ᇁ	0x1111		#0x314D
0x11C2 ᇂ	0x1112		#0x314E

Tableau A32 Tableau Conversion de syllabes en lettres latines

```
#
# convert table to alphabet
<SS>    _           # boundary of syllable
Initial_Consonants
0x1100 0x3131 G         #0x3131
0x1101 0x3132 G G       #0x3132
0x1102 0x3134 N         #0x3134
0x1103 0x3137 D         #0x3137
0x1104 0x3138 D D       #0x3138
0x1105 0x3139 L         #0x3139
0x1106 0x3141 M         #0x3141
0x1107 0x3142 B         #0x3142
0x1108 0x3143 B B       #0x3143
0x1109 0x3145 S         #0x3145
0x110A 0x3146 S S       #0x3146
0x110B 0x3147           #0x3147
0x110C 0x3148 J         #0x3148
0x110D 0x3149 J J       #0x3149
0x110E 0x314A C H       #0x314A
0x110F 0x314B K         #0x314B
0x1110 0x314c T         #0x314C
0x1111 0x314D P H       #0x314D
0x1112 0x314E H         #0x314E
Vowels
0x1161 0x314F A         #0x314F
0x1162 0x3150 A I       #0x3150
0x1163 0x3151 I A       #0x3151
0x1164 0x3152 I A I     #0x3152
0x1165 0x3153 E         #0x3153
0x1166 0x3154 E I       #0x3154
0x1167 0x3155 I E       #0x3155
0x1168 0x3156 I E I     #0x3156
0x1169 0x3157 O         #0x3157
0x116A 0x3158 O A       #0x3158
0x116B 0x3159 O A I     #0x3159
0x116C 0x315A O I       #0x315A
0x116D 0x315B I O       #0x315B
0x116E 0x315c U         #0x315C
0x116F 0x315d U E       #0x315D
0x1170 0x315E U E I     #0x315E
0x1171 0x315F U I       #0x315F
0x1172 0x3160 I U       #0x3160
0x1173 0x3161 E U       #0x3161
0x1174 0x3162 W I       #0x3162
0x1175 0x3163 I         #0x3163
Final_Consonants
FCNULL - - - # the final consonant is empty
0x11A8 0x3131 G         #0x3131
0x11A9 0x3132 G G       #0x3132
0x11AA 0x3133 G S       #0x3133
0x11AB 0x3134 N         #0x3134
0x11AC 0x3135 N J       #0x3135
0x11AD 0x3136 N H       #0x3136
0x11AE 0x3137 D         #0x3137
0x11AF 0x3139 L         #0x3139
0x11B0 0x313a L G       #0x313A
0x11B1 0x313b L M       #0x313B
0x11B2 0x313c L B       #0x313C
```

0x11B3 0x313d L S #0x313D
0x11B4 0x313e L T #0x313E
0x11B5 0x313f L P #0x313F
0x11B6 0x3140 L H #0x3140
0x11B7 0x3141 M #0x3141
0x11B8 0x3142 B #0x3142
0x11B9 0x3144 B S #0x3144
0x11BA 0x3145 S #0x3145
0x11BB 0x3146 S S #0x3146
0x11BC 0x3147 N G #0x3147
0x11BD 0x3148 J #0x3148
0x11BE 0x314a C H #0x314A
0x11BF 0x314B K #0x314B
0x11C0 0x314c T #0x314C
0x11C1 0x314c P #0x314D
0x11C2 0x314e H #0x314E

Annexe 4 Transducteurs de transcodage des alphabets vers les syllabes

Le graphe ci-dessous reconnaît les chaînes de lettres alphabétiques coréennes et produit les chaînes en syllabes coréennes codées par UNICODE.

Le premier graphe décompose en sous graphes, le sous graphe « JAMOS » reconnaît les caractères alphabétiques coréen qui ne sont pas des éléments dans la syllabe, selon l'ordre des éléments de syllabe.

Le symbole '<SS>' indique le début de séquence d'une syllabe. Les séquences qui commencent par ce symbole sont reconnues comme syllabes. Après ce symbole, le sous graphe IC est le sous graphe qui reconnaît la partie de consonne initiale. VW est pour la partie de voyelles et FC est pour la partie de consonnes.

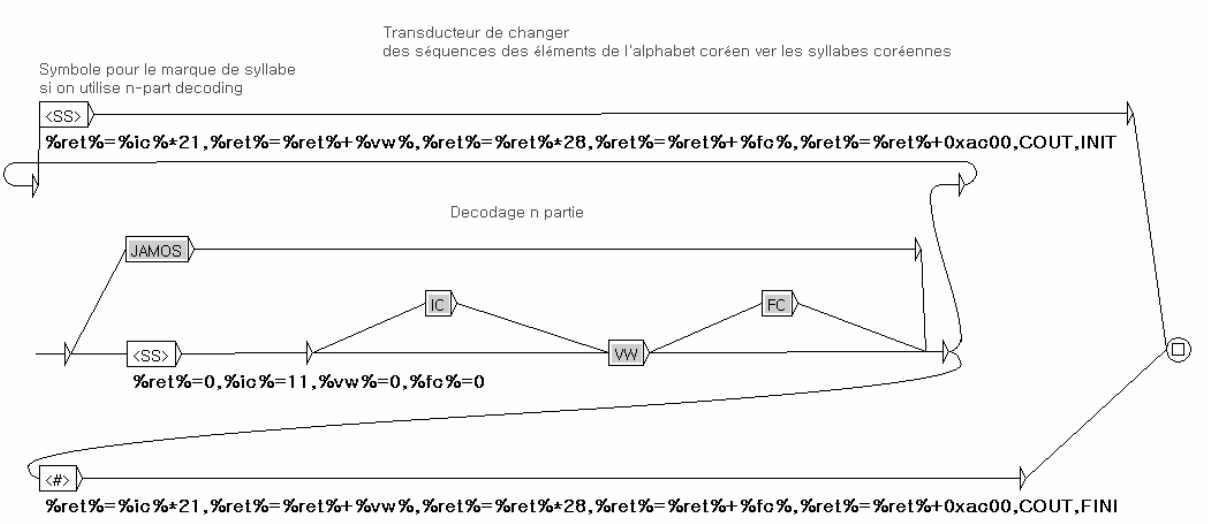


Figure A Le premier graphe de transcodage de séquence de caractères alphabétiques vers des syllabes

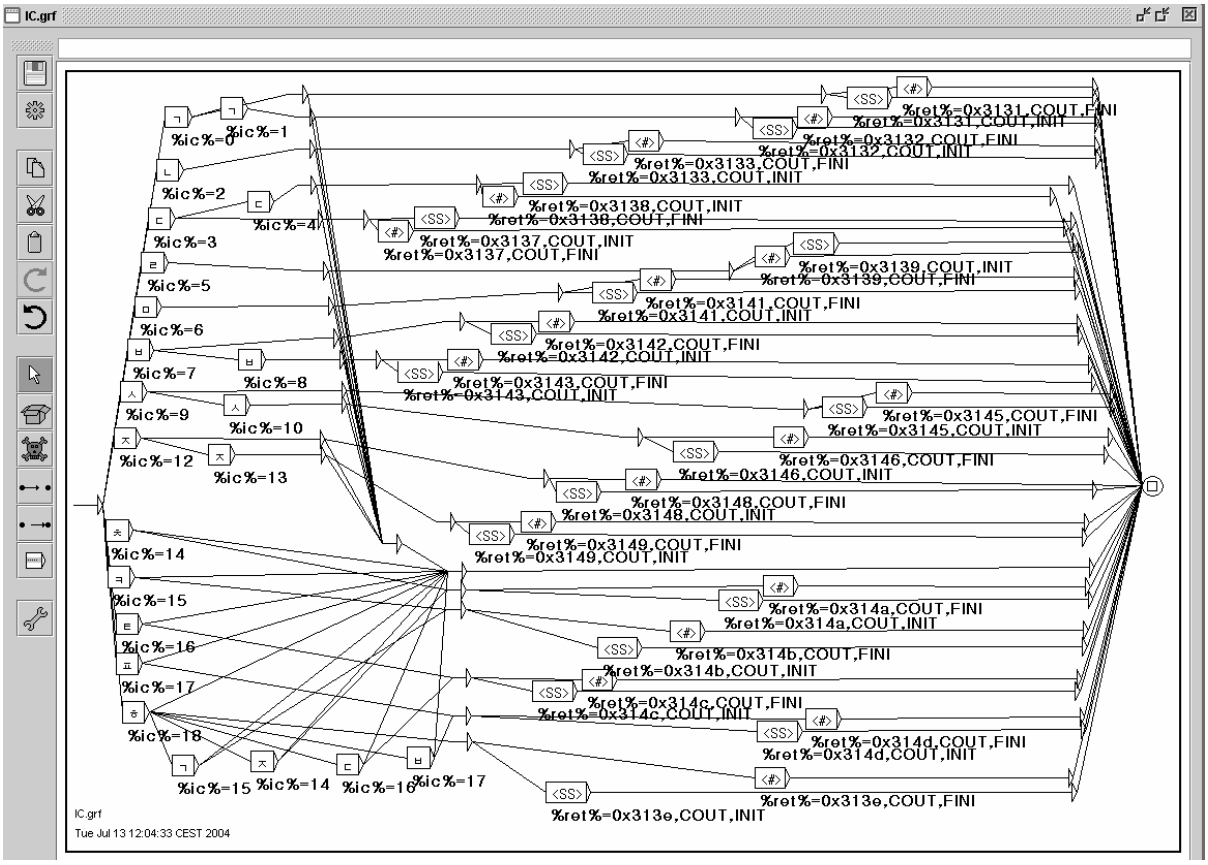


Figure B Sous-graphe pour obtenir la valeur de la consonne initiale

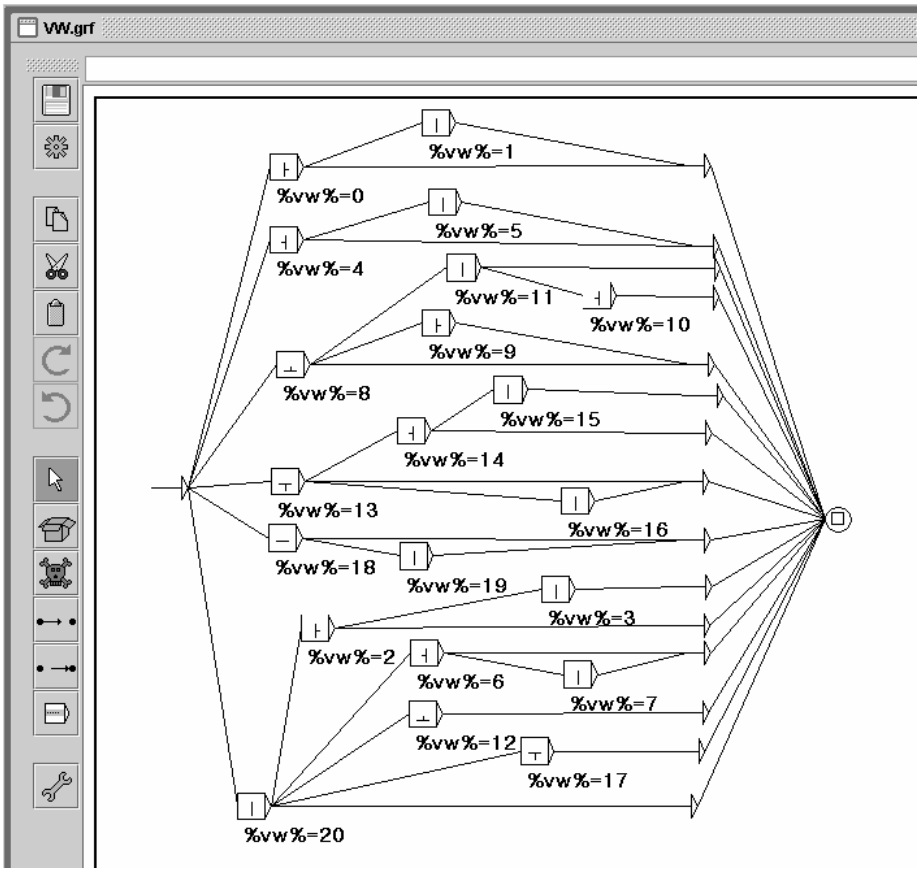


Figure C Sous-graphe pour obtenir la valeur de la voyelle

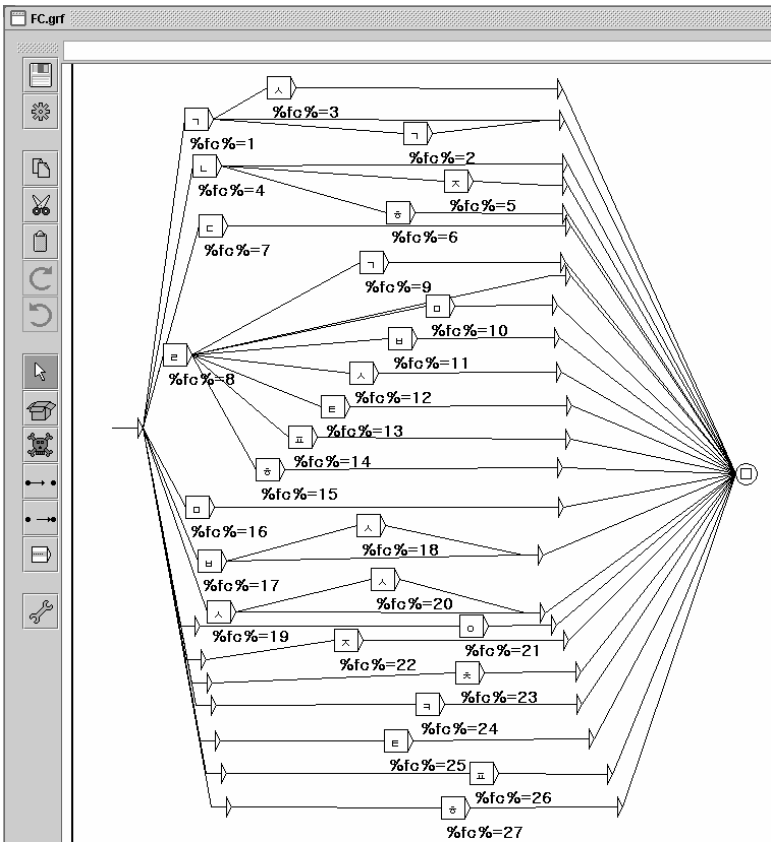


Figure D Sous-graphe de FC, pour obtenir la valeur de la consonne finale

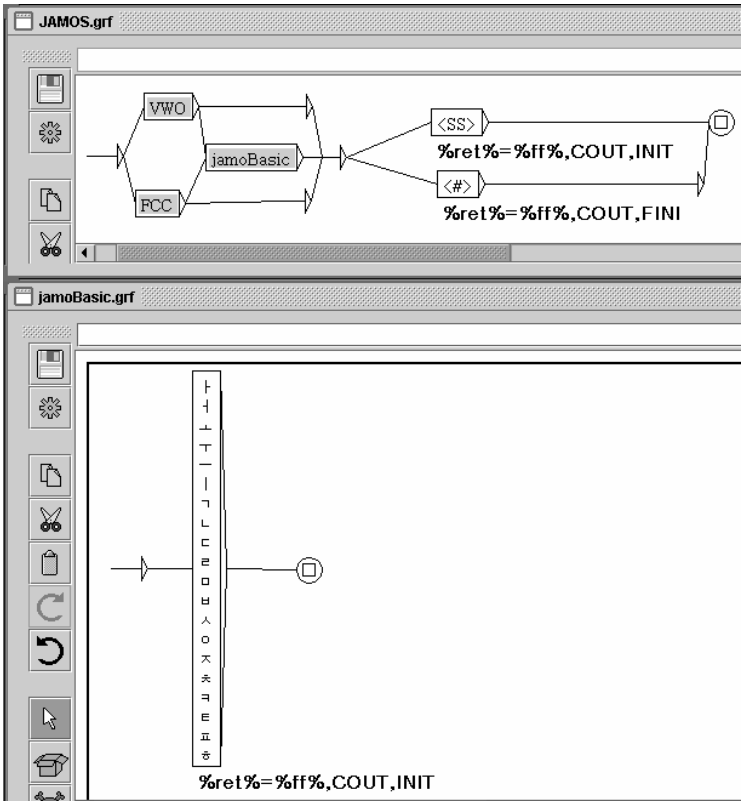


Figure E Sous-graphe pour reconnaître les séquences de non syllabes

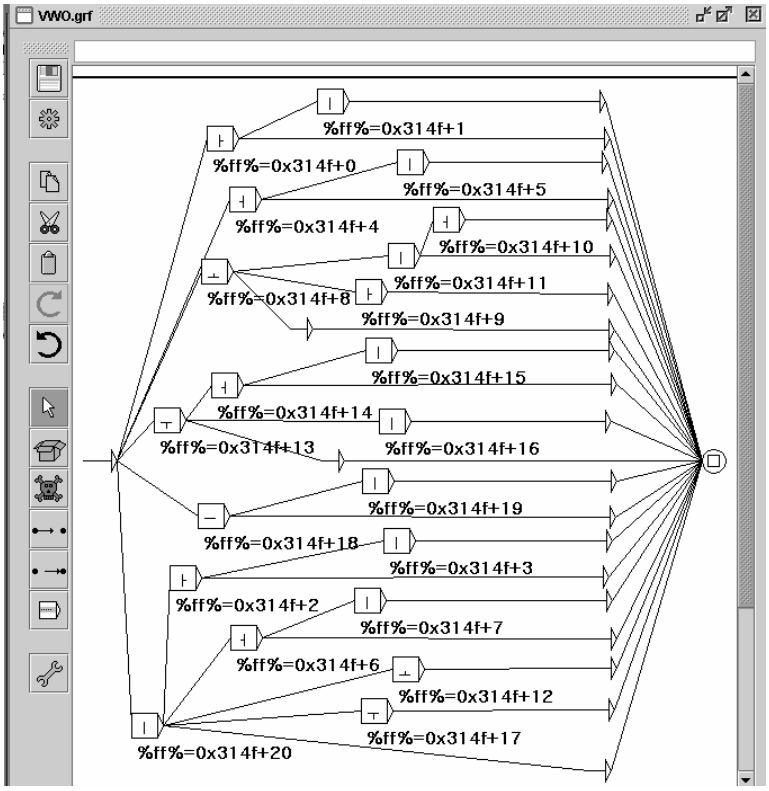


Figure F Sous-graphe des voyelles dans la séquence de non syllabe

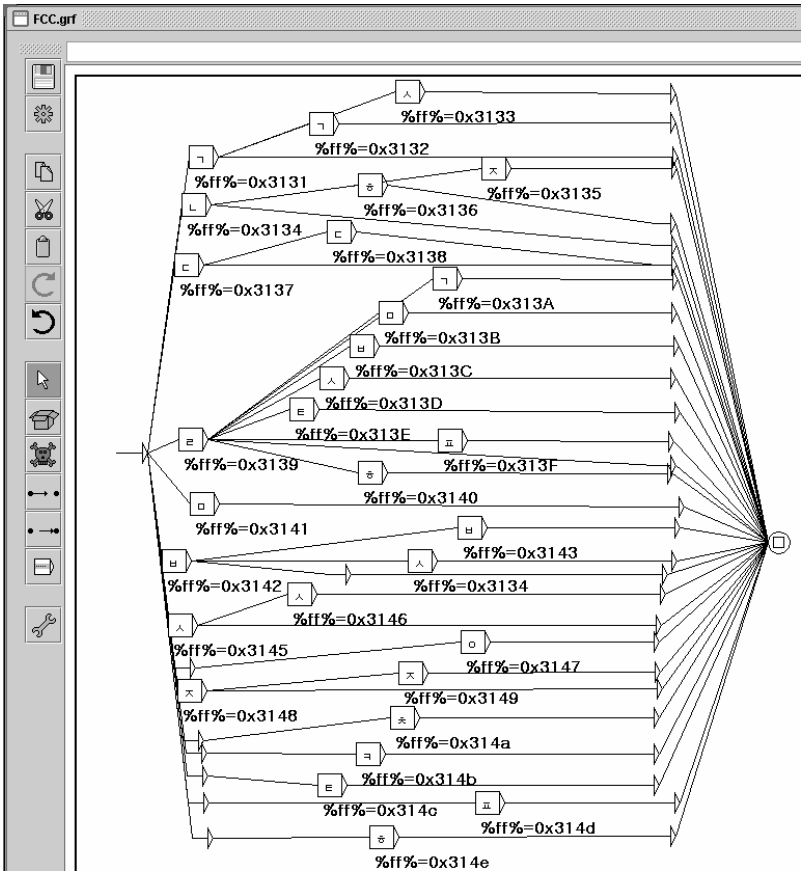


Figure G Sous graphe des consonnes dans la séquence de non syllabes

Annexe 5 Liste des étiquettes

Tableau D Liste des étiquettes des suffixes

A	N+HUM	N+PRED2+PRED2	V+aux
A+Cop	N+HUM+ADJH	N+PRED2+PRED3	V+i
A+Ha_der_	N+HUM+ANM	N+PRED3	V+i+ADVana
A+neg	N+HUM+ETR	N+PREDH	V+i+DE
ADV	N+HUM+ETR+PREDH	N+PREDH+PRED3	V+i+JMA
ADV+DIR	N+HUM+ETR+PREDHA	N+PREDHA	V+i+JMA+VC
ADV+LOC	N+HUM+Gf	N+Per+Hum	V+i+PS
ADV+ModiV	N+HUM+Gm	N+TIM	V+i+PS+VC
ADV+PosN	N+HUM+NVK	N+unknown+count	V+i+RHM
ADV+PosV	N+HUM+NVS	NI	V+i+RHM+DE
ADV+neg	N+HUM+PLT	NI+HUM	V+i+VC+neg
ADV+tellement	N+HUM+PRED1	NI+HUM+Gf+noHon	V+t
ADV+unknown+raison	N+HUM+PRED1+PRED2	NI+HUM+Gm	V+t+CS
DET+adj	N+HUM+PRED1+PRED3	NI+HUM+Gm+noHon	V+t+CS+VC
DET+anonymat	N+HUM+PRED2	NI+HUM+hon+Gm	V+t+DE
DET+unknown	N+HUM+PREDH	NI+LOC	V+t+PS
DET+unknown+count	N+HUM+PREDHA	NI+Nunit+N_NC	V+t+PS+VC
DET+unknown+quel	N+NF	NI+Nunit+nonN_NC	V+t+RHM
DET	N+NF+Udef	PRO+LOC	V+t+RHM+DE
DET+demon1	N+NF+appelation	PRO+anonymat	V+t+RHM+VC
DET+demon1+p2	N+NI	PRO+hum+unknown	V+t+RHMA
DET+demon1+p3	N+NVK	PRO+hum+unknown+qui	V+t+RKM
DET+demon2	N+NVK+PRED1	PRO+lieu+unknown	V+t+VC
DET+demon3	N+NVK+PRED1+PRED2	PRO+noHon+p2	V+t+i
N	N+NVK+PRED2	PRO+p1	V+t+i+DE
N+ADJH	N+NVK+PREDHA	PRO+p1+hon	V+t+i+RHM
N+ADJH+PRED1+PRED2	N+NVM	PRO+p1+plur	V+t+i+RHM+DE
N+ADJH+PRED1+PRED3	N+NVM+PRED1	PRO+p2	V+t+i+RHMA
N+ADJH+PRED2	N+NVM+PRED2	PRO+p2+hon+vous	V+t+i+RKM
N+ANM	N+NVM+PREDH	PRO+p3+Gf	V+t+i+RKM+VC
N+ANM+ETR	N+NVM+PREDHA	PRO+p3+Gm	V+t+i+VC
N+ANM+ETR+PREDHA	N+NVS	PRO+self	INJ
N+ANM+PRED1	N+NVS+PRED1	PRO+self+hon	
N+ANM+PREDH	N+NVS+PRED2	PRO+unknown	
N+ANM+PREDHA	N+NVS+PREDHA	PRO+unknown+quoi	
N+ETR	N+PLT	PRO+unknown+raison	
N+ETR+ADJH	N+PLT+ANM	PRO+unknown+time	
N+ETR+PRED1	N+PLT+ETR	V	
N+ETR+PRED1+PRED2	N+PLT+ETR+PREDHA	V+DE	
N+ETR+PRED2	N+PLT+PREDH	V+PRO	
N+ETR+PREDH	N+PLT+PREDHA	V+PRO+unknown	
N+ETR+PREDHA	N+PRED1	V+RKM	
V+i+RHM+DE+VC	N+PRED1+PRED2	V+VC	
V+i+RHM+VC	N+PRED1+PRED2+PRED3	N+PRED2	
V+i+RHMA	N+PRED1+PRED3	N+PRED2+ADJH+PRED3	
V+i+RHMA+DE	V+i+RKM+VC	V+i+VC+PS	
V+i+RKM	V+i+VC	V+i+VC+RKM	

Tableau E Liste des étiquettes des suffixes

sfx+A	Post+datif+src	St+excl+Hao
sfx+ADJ	Post+dest+point	St+excl+Hge
sfx+ADV	Post+dir	St+excl+Hla
Morph	Post+genitif	St+imp+Hai
Morph+acom	Post+obj+noHon	St+imp+Hao
Morph+asp+de	Post+nmtf	St+imp+Hge
Morph+fut	Post+plus+spc	St+imp+Hla
Morph+hon	Post+point+src	St+imp+Hso
Morph+hon+o	Post+position	St+inf
Morph+hon+suj	Post+raison	St+int
Morph+pass	Post+title	St+int+Hao
Morph+pf	Post+tool	St+int+Hge
Morph+plur	Post+spc	St+int+Hla
Morph+pp	Sc	St+int+Hso
Morph+pre	Sc+Subor	St+prop
Post+accu	Sc+cond	St+prop+Hao
Post+allatif	Sc+coor	St+prop+Hge
Post+aux+aussi	Sc+subor	St+prop+Hso
Post+aux+only	Sd	St+prop+Hla
Post+cita	Sd+fut	Suf+each
Post+comp+egal	Sd+pass	Suf+egal
Post+comp+nonEgal	Sd+pre	Suf+group
Post+comp+min	Sncomp	Suf+plur
Post+comp+unique	St+dec	
Post+conj	St+dec+Hai	
Post+datif	St+dec+Hao	
Post+datif+dest	St+dec+Hge	
Post+datif+dest+hum	St+dec+Hla	
Post+datif+dest+hum+dir	St+dec+Hso	
	St+dec+Pi	
	St+excl	
	St+excl+Hai	

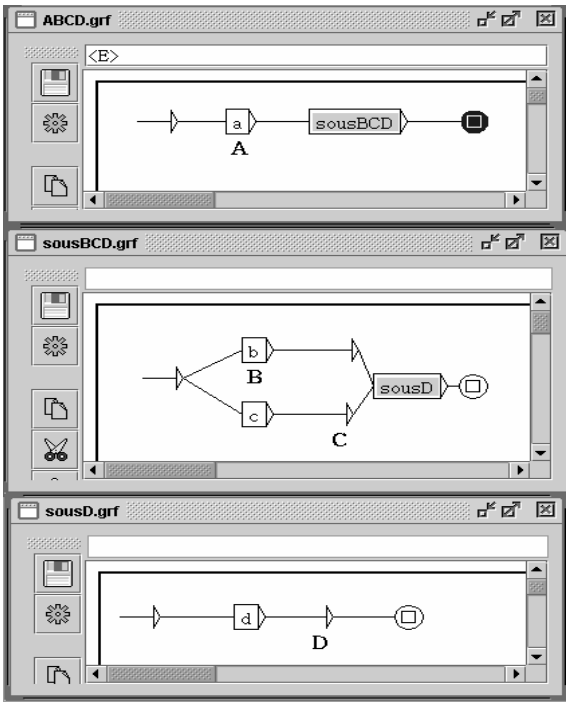
Annexe 6 Extraction des éléments sous l'automate par le programme « fst2list »

Nous donnons le programme « fst2list » pour montrer les éléments dans les états de l'automate ou du transducteur. C'est un outil pour vérifier les travaux de sous graphes.

Il fabrique les chemins de l'automate et du transducteur à partir du fichier avec l'extension « .fst2 » qui est obtenue du fichier « .grf » qui contient les expressions d'automate ou transducteur.

Chaque chemin dans le graphe est présenté en forme de liste des données qui sont obtenues dans l'ordre pendant le parcours de l'automate ou du transducteur par une ligne. Les données sont les entrées ou les sorties dans les nœuds.

Par défaut, il traite le transducteur comme l'automate sans compter les sorties. Si on choisit le mode de transducteur, il montre les entrées de l'état et les sorties de l'état. Selon l'option, on peut voir les séquences des entrées et sorties en paires ou non. L'exemple suivant montre les résultats d'un graphe « ABCD.grf » avec les sous-graphes appelés selon les options de l'affichage.



a)
1) fst2list ABCD.fst2

acd
abd

2) fst2list -t s -o abcdsep.txt ABCD.fst2

acdACD
abdABD

3) fst2list -t s -f a -o abcdass.txt ABCD.fst2

aAcCdD
aAbBdD

4) fst2list -p s -o sousgraphes.txt ABCD.fst2

[1 th automata ABCD]
a{sousBCD}
the automate ABCD, 1 path, 0 path stopped by cycle, 0 error path
[2 th automata sousBCD]
c{sousD}
b{sousD}
the automate sousBCD, 2 path, 0 path stopped by cycle, 0 error path
[3 th automata sousD]
d
the automate sousD, 1 path, 0 path stopped by cycle, 0 error path

Avec l'option « -p s », on peut obtenir tous les chemins des sous-graphes qui sont appelés à partir du premier graphe. On peut utiliser la vérification des chemins et la localisation de l'élément auquel on inspire de l'intérêt.

En utilisation de ce programme, nous pouvons obtenir les formes fléchies avec une racine et les séquences de suffixe pour le coréen avec un processus « figure a3-1 ».

Nous montrons un exemple de l'utilisation de ce programme pour extraire toutes les formes fléchies qui sont la combinaison d'une racine «먹다 : *mog_da* :manger » et ses séquences de suffixes verbaux.

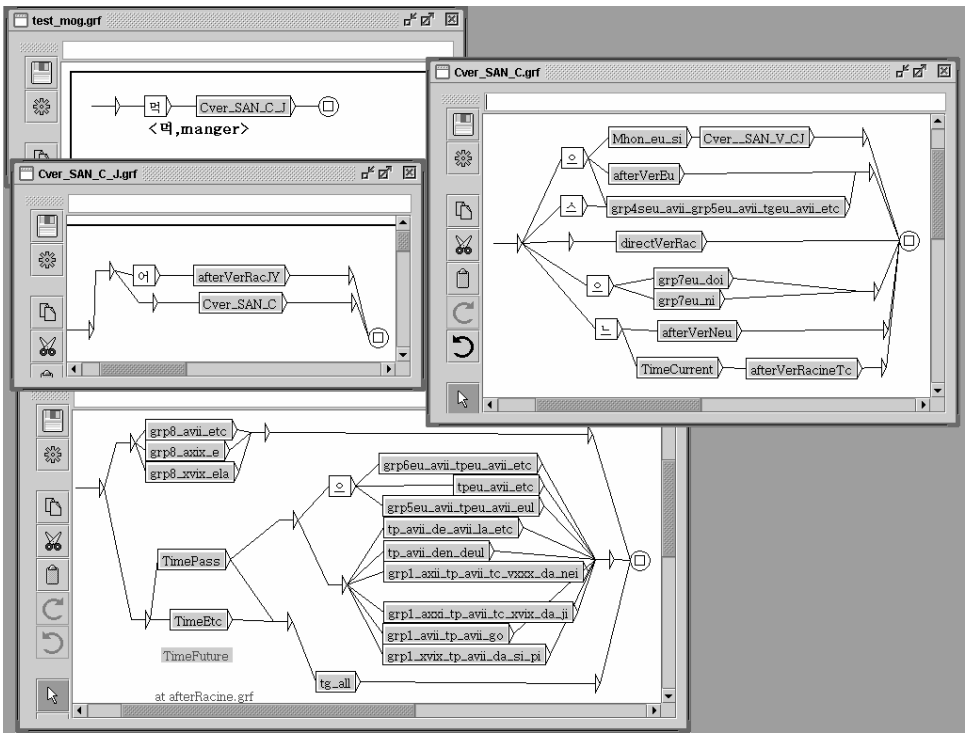
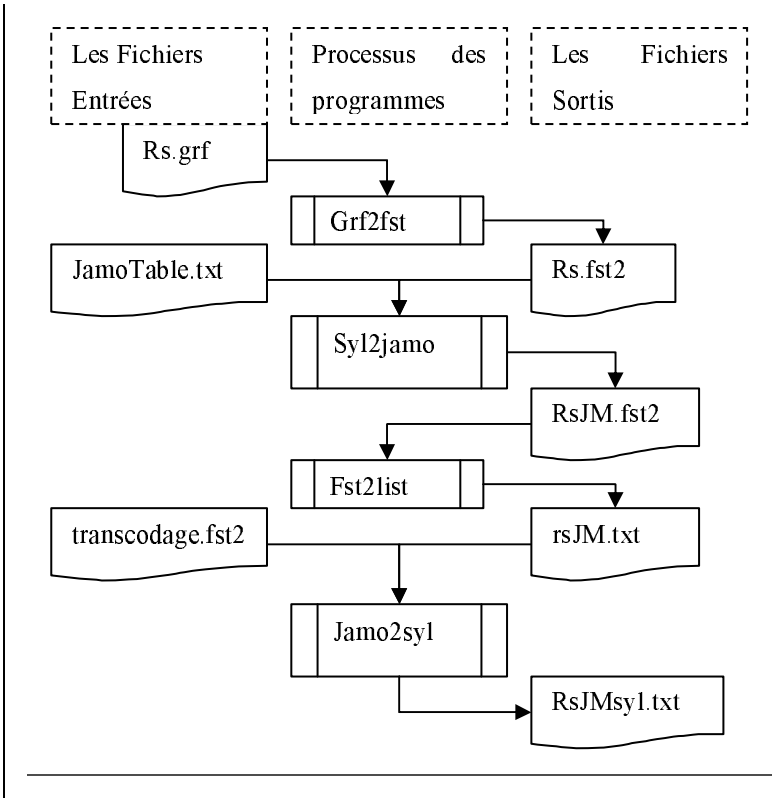


Figure H Une racine "mog_da" et les séquences de suffixes
Le résultat d'un processus d'exécution des programmes ci-dessous.



syl2jamo -j -o t0.fst2 %1.fst2

fst2list -m -v -t s -f s -ss "#" -s0 "," -rl ":" -c SS=0x318d -c ss=0x318d -i Cida_C -i Cida_V -i Cver_SAN_V_CJ -i Cver_HA0_V_H -i Cadj_HA0_V_H -i NV -i NC -i sfx_adj_se_leb t0.fst2

jamo2syl -c SS=0x318d -c SS=0x318d -o %1.txt D:\myUx\Korean\decodage\uneSyljm.fst2 t0.txt

Une partie du fichier des listes est le suivant :

먹었더니라,<먹,manger><었,Morph+Time+Pass><더,Morph+asp+de><느니,St+dec+Hge>:

먹었더니,<먹,manger><었,Morph+Time+Pass><더,Morph+asp+de><느니,St+dec+Hge>:

먹었어야,<먹,manger><었,Morph+Time+Pass><어야,Sc+cond>:

먹었어야만은,<먹,manger><었,Morph+Time+Pass><어야,Sc+cond><만,Post+spc+only><는,Post+spc>:

먹었어야만요,<먹,manger><었,Morph+Time+Pass><어야,Sc+cond><만,Post+spc+only><요,Morph>:

먹었어야만,<먹,manger><었,Morph+Time+Pass><어야,Sc+cond><만,Post+spc+only>:

먹었어야,<먹,manger><었,Morph+Time+Pass><어야,Sc+cond><하,V+aux>:Cver_HA0_V_H

먹었어도,<먹,manger><었,Morph+Time+Pass><어도,Sc+subor>:

먹었어,<먹,manger><었,Morph+Time+Pass><어,St+dec+Hla>:

Nous mettons l'option « -i », elle indique l'arrêt du parcours, s'il y a appellation de sous-graphe qui est défini avec cette option.

L'option « -ss » indique une chaîne de caractères qui se situe après cette option et entre les symboles « <> » pour être utilisée comme arrêt de parcours au milieu d'un parcours.

L'option « -c » est le changement d'une chaîne de caractères à un caractère en UNICODE. Nous l'utilisons pour exprimer la marque du bord de syllabe. La marque de bord de syllabe n'existe pas dans l'alphabet coréen, nous utilisons la chaîne de caractères « <SS> ».

Il y a des chemins cycliques à cause des suffixes de transformation ou de contraction. Il y a trois méthodes pour traiter la condition cyclique où le parcours de l'automate rencontre l'état qui est l'état passé.

- Arrêter le parcours
- Présenter par les symboles des chemins cycliques.
- Indiquer la marque cyclique

Avec le schéma ci-dessous, on peut obtenir les affichages des chemins cycliques selon les options.

Figure I Chemin cyclique

Tableau E Résultats selon les options

	Les options de la ligne de commande	
a)		ilfaittrèstrès ilfaittrèsbeau
b)	-s ";" -rx "<,>"	;il;fait<;très;très> ;il;fait;très;beau
c)	-s ";" -rs "[,]*"	;il;fait;[très|C0]*;beau C0[;très
d)	-s ";"-rl ">,:"	;il;fait;très;>Loc0 Loc0:;très;>Loc0 Loc0:;beau
e)	-t s -s0 ",,," -s ";" -rx "<,>"	;il;fait<;très;très>,,,;A;<;B;B> ;il;fait;très;beau,,,;A;;B;C
f)	-t s -s0 ",,," -s ";" -rs "[,]*"	;il;fait;[très|C0]*;beau,,,;A;;[B|C0]*;C C0[;très,,,;B
g)	-t s -s0 ",,," -s ";" -rl "\,,:"	;il;fait;très;,,,;A;;B;,Loc0 Loc0:;très;,,,;B;,Loc0 Loc0:;beau,,,;C

h)	-t s -f a -s0 ",,," -s ";" -rx "<,>"	;il,,,A;fait,,,<;très,,,B;très,,,B> ;il,,,A;fait,,,;très,,,B;beau,,,C
i)	-t s -s0 ",,," -s ";" -rs "[,]*"	;il,,,A;fait,,,;[très|C0]*,,,[B|C0]*;beau,,,C C0[;très,,,B
j)	-t s -s0 ",,," -s ";" -rl "\,,:"	;il,,,A,;fait,,,,;très,,,B,;,,,,Loc0 Loc0:;très,,,B,;,,,,Loc0 Loc0:;beau,,,C,

Ce programme a la limitation de traiter le nombre des chemins cycliques. Nous vous conseillons d'utiliser l'automate qui a un peu de nombre de chemins cycliques.

Nous utilisons le programme « fst2list » pour fabriquer les fichiers entrés de la compression.

Utiliser la vérification des chemins des graphes des transducteurs ou des automates.

Il peut aussi être utilisé pour montrer les mots de la combinaison de la racine spécifique et les séquences de suffixes.