Modèle de document pour TALN 2015
Représentation des expressions composées en macédonien en tant qu'entrées lexicales en Unitex

Aneta Rafajlovska[1], Katerina Zdravkova[2]
Université Sts Cyrille et Méthode, Skopje
Faculté de science informatique et de génie informatique
[1]r.aneta@yahoo.com, [2]katerina.zdravkova@finki.ukim.mk

Résumé. Le logiciel de traitement de corpus Unitex 3.0 a été utilisé pour obtenir la flexion automatique de mots simples et des mots composés en langue macédonienne. En utilisant les graphes de flexion des mots simples, nous avons réussi à représenter les expressions composées du corpus en tant qu'entrées lexicales dans un dictionnaire DELAC en Unitex. En outre, nous avons créé des transducteurs à états-finis qui permettent de fléchir les expressions composées et nous avons obtenu automatiquement toutes leurs formes fléchies que nous avons stockées dans un dictionnaire DELACF (DELA de formes Composées Fléchies).

Abstract.

Representation of Multiword Expressions in Macedonian as Lexical Entries in Unitex
The corpus processing system – Unitex 3.0 was used to obtain the automatic inflection of the simple word forms and the multiword expressions in Macedonian. Based on the inflection graphs of the simple word forms we managed to represent the multiword expressions retrieved from the corpus as lexical entries in a DELAC dictionary in Unitex. We also created inflection finite-state transducers for the multiword expressions and as a result we managed to obtain automatically all the inflected forms of the multiword expressions in the form of a DELACF dictionary of compound inflected forms.

Mots-clés : expressions composées, mots composés, mots simples, flexion automatique, transducteurs à états-finis de flexion, Unitex, Multiflex
Keywords: multiword expressions, compound words, simple word forms, automatic inflection, inflection finite-state transducers, Unitex, Multiflex

1 Expressions composées

Les expressions composées ou les mots composés représentent un problème linguistique assez important, surtout à cause de la difficulté de les définir et de les représenter. Par conséquent, les mots composés sont énumérés parmi les problèmes majeurs en traitement automatique des langues (TAL), générant des ambigüités conséquentes (Sag, Baldwin, Bond, Copestake, Flickinger, 2002). Les études récentes dans le domaine du TAL ont incité les linguistes à abordes de nouvelles théories linguistiques et à développer différentes approches par rapport à la syntaxe et la lexicologie dès les années 1960 (Léon, 2004).

Il existe de nombreuses définitions linguistiques et pragmatiques pour les expressions composées. Toutefois, il est généralement admis que les expressions composées contiennent au moins deux ou plusieurs mots qui représentent un seul lexème. Dans ce sens un lexème signifie une seule unité lexicale. De ce fait, les expressions composées contiennent deux ou plusieurs mots, mais ils représentent un seul ensemble qui peut différer du sens premier de ces mots pris séparément. Elles posent également un problème de représentation, car l'unité représntative dans un lexique linéaire est le mot, ce qui les exclut dans les dictionnaires (Gross, 1986).

D'un point de vue de la nature, les expressions composées en macédonien peuvent être des adverbes composés, des noms composés ou des verbes composés, comme dans les exemples suivants : « под услов да » (*pod uslov da*, ADV) - « sous condition que/de », « во врска со » (*vo vrska so*, ADV) – « à propos de / concernant », « во недостиг на » (*vo nedostig na*, ADV) – « à défaut de / en absence de », « високо друштво » (*visoko društvo*, NOM) – « haute société », « здрав разум » (*zdrav razum*, NOM) – « bon sens / lucidité », « роден крај » (roden kraj, NOM) – « pays natal », « доби премија » (*dobi premija*, V) – « gagner le gros lot », « има предвид » (*ima predvid*, V) « prendre en

considération / tenir compte de », « ги зема работите во свои раце » (*gi zema rabotite vo svoi race*, V) – « prendre le contrôle de » etc.

2 Méthodologie

Notre objectif premier est de représenter les expressions composées en tant qu'entrées lexicales dans un dictionnaire morphologique pour pouvoir obtenir automatiquement leur flexion. Notre travail consiste en quatre étapes essentielles :

- Extraction et annotation des expressions composées macédoniennes du corpus
- Création du dictionnaire des mots simples
- Flexion automatique des mots simples
- Création du dictionnaire des mots composés
- Flexion automatique des mots composés

La création de cette ressource Unitex a pris un peu plus de trois mois. Le dictionnaire des mots simples (Figure 1) contient 280 entrées lexicales, ce qui permet d'obtenir automatiquement le dictionnaire des formes fléchies avec, environ 3 762 entrées (Figure 4). Le dictionnaire des mots composés contient 184 mots composés (Figure 5). Il permet d'obtenir automatiquement le dictionnaire des formes fléchies qui contient 1 454 entrées (Figure 8).

3 Particularités morphologiques du macédonien

Les catégories grammaticales des noms et des adjectifs en macédonien sont le genre, le nombre et la définitude qui s'exprime par un suffixe appelé « article défini ». Les étiquettes de ces catégories sont les suivantes :

1. Genre
- Masculin m
- Féminin f
- Neutre n

2. Nombre
- Singulier s
- Pluriel simple p
- Pluriel compté i
- Pluriel collectif z

3. Article définie
- Indéterminé U
- Proximal (objets qui se trouvent à côté de celui qui parle) P
- Distal (objets qui se trouvent loin de celui qui parle) D

4 Dictionnaire et flexion automatique des mots simples

Pour pouvoir aborder le problème linguistique en question et pour obtenir la flexion automatique des mots simples de même que des mots composés, nous avons utilisé Unitex 3.0. Les dictionnaires électroniques distribués en Unitex utilisent la structure DELA (Dictionnaires Electroniques du LADL) (Paumier, 2006). Nous avons créé un dictionnaire macédonien des formes lexicale des mots simples au format DELA en Unitex (Figure 1). Le dictionnaire contient les formes lexicales des mots suivies par une virgule et le nom du graphe de flexion qui sera utilisé pour obtenir toutes les formes fléchies automatiquement.

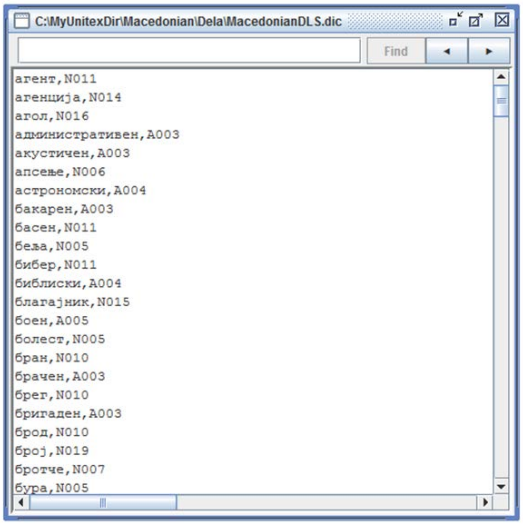

Figure 1 : Extrait du dictionnaire des mots simples dans le format DELA

4.1 Graphes de flexion des mots simples

Les graphes en Unitex compilés sous le format *.fst2* ont été utilisés en vue d'obtenir toutes les formes fléchies des lemmes des mots simples. Pour couvrir toutes les formes lexicales du dictionnaire des mots simples, nous avons créé 19 graphes de flexion (transducteurs à états-finis) pour les noms et 4 graphes pour les adjectifs.

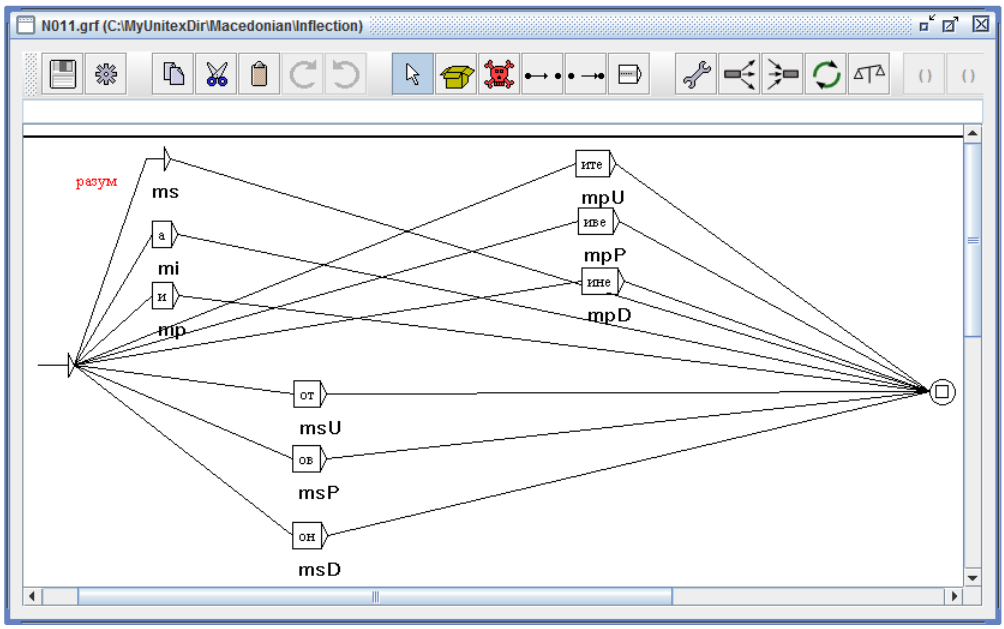

Figure 2 : Un graphe de flexion pour les noms masculins

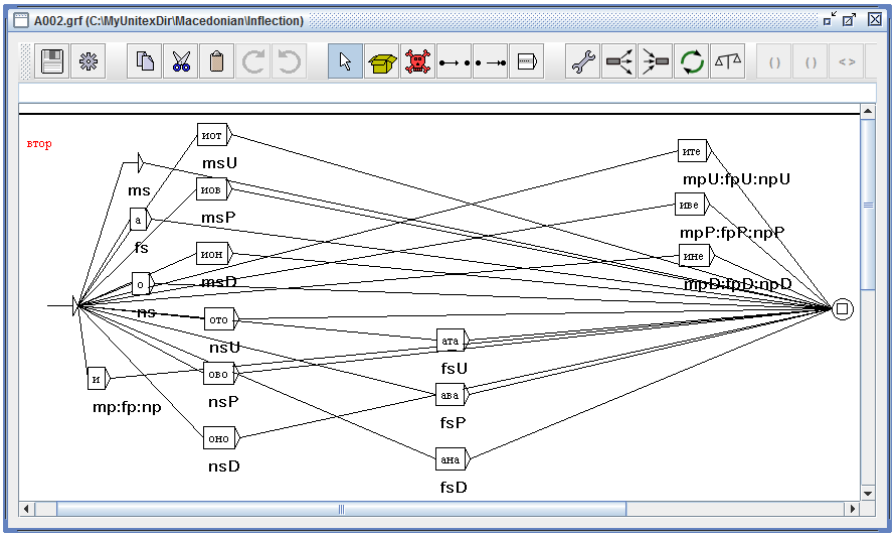

Figure 3 : Un graphe de flexion pour les adjectifs

Dans les Figures 2 et 3, nous pouvons observer les différents suffixes ajoutés à la forme lexicale du mot avec les codes flexionnels pour la catégorie grammaticale. Pour le nom ‘разум’ (razum) ‘sens’ et tous les autres noms qui utilisent le graphe N011 pour la flexion automatique, les suffixes –a et –i seront ajoutés à la forme lexicale pour former les deux formes du pluriel, ce qui donne dans le cas de ‘razum’ - ‘razuma’ et ‘razumi’. Pour l’article défini indéterminé le suffixe –ot sera ajouté ce qui donne ‘razumot’, le suffixe –ov pour les objets proches ‘razumov’ ce qui peut être traduit par ‘celui-ci’, et le suffixe –on ‘razumon’ pour les objets lointains ce qui peut être traduit par ‘celui-là’. Au pluriel les articles définis seront exprimés par les suffixes -ite ‘razumite’ (indéterminé), -ive ‘razumive’ (ceux-ci) et –ine ‘razumine’ (ceux-là). En revanche, les adjectifs varient aussi selon le genre (Figure 3), la forme lexicale est au masculin singulier ‘vtor’. Le suffixe –a est ajouté pour créer la forme féminine ‘vtora’, et le suffixe –o est ajouté pour créer la forme neutre ‘vtoro’. La forme plurielle est la même pour tous les trois genres formée par le suffixe –i ‘vtori’. Tous les trois genres prennent les trois formes de l’article défini, de même que la forme plurielle, pour le masculin ‘vtoriot’, ‘vtoriov’ et ‘vtorion’ ; pour le féminin ‘vtorata’, ‘vtorava’ et ‘vtorana’ ; pour le neutre ‘vtoroto’, ‘vtorovo’ et ‘vtorono’ et pour le pluriel ‘vtorite’, ‘vtorive’ et ‘vtorine’.

4.2 Dictionnaire macédonien de formes fléchies des mots simples

Après avoir construit le Dictionnaire macédonien de lemmes de mots simples (Figure 1) nous avons appliqué les graphes de flexion et nous avons obtenu automatiquement le Dictionnaire des formes fléchies (Figure 4).

Le Dictionnaire des formes fléchies (Figure 4) comprend la forme fléchie suivie d’une virgule et de la forme lexicale du mot, puis d’ un point et du code flexionnel décrivant catégorie grammaticale.

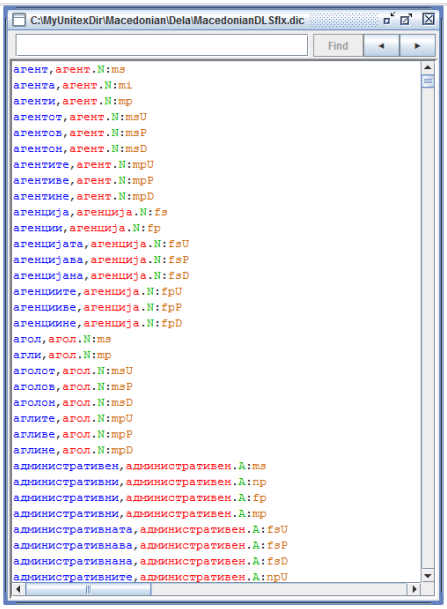

Figure 4 : Extrait du dictionnaire de formes fléchies des mots simples

5 Dictionnaire et flexion automatique des mots composés

L'objectif principal étant la flexion automatique des mots composés, nous avons pris la liste des mots composés du corpus (la traduction macédonienne du roman *Tour du monde en quatre-vingts jours* de Jules Verne) et nous avons créé un dictionnaire DELAC (DELA de formes composées) sous Unitex (Figure 5). Unitex utilise le formalisme Multiflex (Savary, 2008), qui représente une approche graphique pour représenter la flexion des expressions composées. Les graphes de flexion des expressions composées réutilisent les graphes de flexion de ces composants - les mots simples.

Le dictionnaire contient la forme fléchie du premier constituant de l'expression composée, suivie de la forme lexicale entre parenthèses, puis du graphe de flexion du mot, du deuxième constituant et de sa forme lexicale, puis du graphe de flexion du deuxième mot, d'une virgule et du nom du graphe de flexion de l'expression composée.

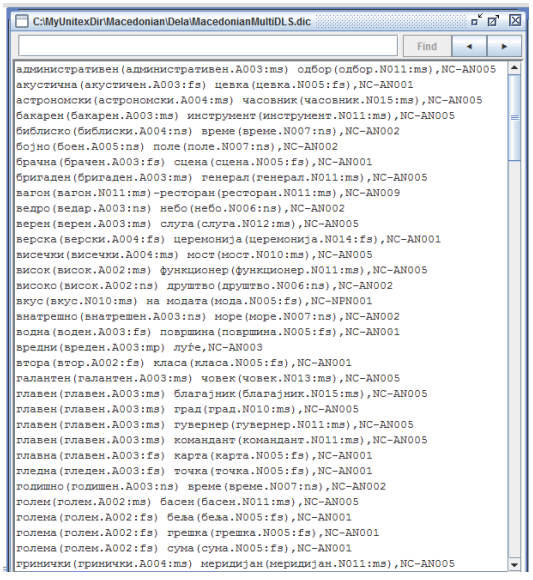

Figure 5 : Extrait du dictionnaire des mots composés

5.1 Graphes de flexion de mots composés

Le graphe de flexion de mots composés (Figure 6) est utilisé pour la flexion de mots composés du type Adjectif-Nom (AN). Les composants du mot composé sont définis par le caractère $1 qui est le premier composant, l'adjectif 'млади' (mladi) 'jeune'; le caractère $2 représente l'espace, et le caractère $3 représente le nom 'години' (godini) 'âge'. Le nom, soit le troisième élément est un féminin pluriel, donc l'adjectif ou le premier élément ne peut être qu'au pluriel Nb=p et au féminin, ce qui est déterminée par Gen=f. En outre, l'adjectif peut prendre le suffixe des trois articles définis qui est déterminé par Def=$d.

Tout d'abord, l'adjectif 'млади' (mladi) 'jeune' est inclus dans le dictionnaire des mots simples et un graphe de flexion lui est associé. Le même principe s'applique pour le nom 'години' (godini) 'âge'. L'entrée dans le dictionnaire des mots simples pour l'adjectif est le suivant : млад,A002 ; et pour le nom : година,N005. L'entrée dans le dictionnaire des mots contenus dans l'expression composée est la suivante : млади(млад.A002:fp) години,NC-AN006. Nous pouvons observer sur le graphe de la Figure 6 que le troisième élément (le nom) ne change pas et qu'il ne faut pas substituer l'entrée le graphe qui lui est associé. En revanche, l'adjectif peut prendre différentes formes et donc il est obligatoire de citer le graphe de flexion de ce mot. Enfin, après l'application du graphe de flexion de l'expression composée Figure 6, le logiciel produit toutes les formes fléchies de l'expression composée, sans produire les formes qui ne sont pas permises, comme, par exemple, la forme du nom au singulier, ni les formes avec l'article défini pour le nom. Dans le dictionnaire des formes fléchies des expressions composées apparaissent seulement les formes au pluriel :

младите години, млади години.N:pf
младиве години, млади години.N:pf
младине години, млади години.N:pf
млади години, млади години.N:pf

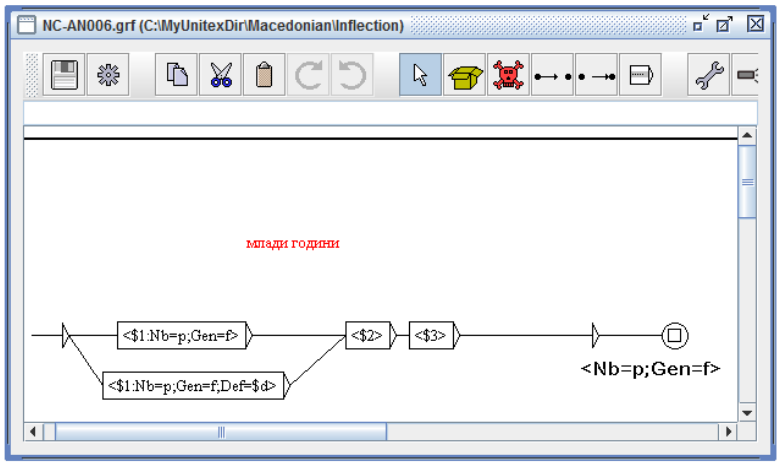

Figure 6 : Un graphe de flexion de mots composés du type AN

Le graphe représenté à la Figure 7 est utilisé pour la flexion des mots composés du type Nom Préposition Nom (NPN), dans le cas où les noms peuvent être fléchis tous les deux. Le caractère $1 représente le premier composant, qui est en fait le nom 'куќичка' (kukjichka) 'coquille', le caractère $2 représente l'espace, le caractère $3 représente la préposition 'на' (na) 'de', le caractère $4 représente l'espace, et le caractère $5 représente le nom 'полжав' (polzhav) 'colimaçon'. Les deux noms peuvent être en n'importe quel nombre Nb=$n et ils peuvent prendre tous les suffixes de l'article défini Def=$d.

Les noms 'куќичка' (kukjichka) 'coquille' et 'полжав' (polzhav) 'colimaçon' sont inclus dans le dictionnaire des mots simples et un graphe de flexion leur est associé. L'entrée dans le dictionnaire des mots simples pour les deux noms est la suivante : куќичка,N005; et полжав,N011. L'entrée dans le dictionnaire des mots composés de l'expression composée est le suivant : куќичка(куќичка.N005:fs) на полжав(полжав.N011:ms),NC-NPN003. Nous pouvons observer que les deux graphes de flexion des noms sont cités, ainsi que le graphe de flexion de l'expression composée Figure 7. Enfin, après l'application du graphe de flexion de l'expression composée le logiciel produit toutes les formes fléchies de l'expression composée. Dans le dictionnaire des formes fléchies des expressions composées apparaissent les formes suivantes:

куќичката на полжав, куќичка на полжав.N:s

куќичката на полжавот, куќичка на полжав.N:s
куќичкава на полжав, куќичка на полжав.N:s
куќичкава на полжавов, куќичка на полжав.N:s
куќичкана на полжав, куќичка на полжав.N:s
куќичкана на полжавон, куќичка на полжав.N:s
куќичките на полжави, куќичка на полжав.N:p
куќичките на полжавите, куќичка на полжав.N:p
куќичките на полжавиве, куќичка на полжав.N:p
куќичкине на полжави, куќичка на полжав.N:p
куќичкине на полжавине, куќичка на полжав.N:p
куќичка на полжав, куќичка на полжав.N:s
куќичка на полжавот, куќичка на полжав.N:s
куќичка на полжавов, куќичка на полжав.N:s
куќичка на полжавон, куќичка на полжав.N:s
куќички на полжави, куќичка на полжав.N:p
куќички на полжавите, куќичка на полжав.N:p
куќички на полжавиве, куќичка на полжав.N:p
куќички на полжавине, куќичка на полжав.N:p

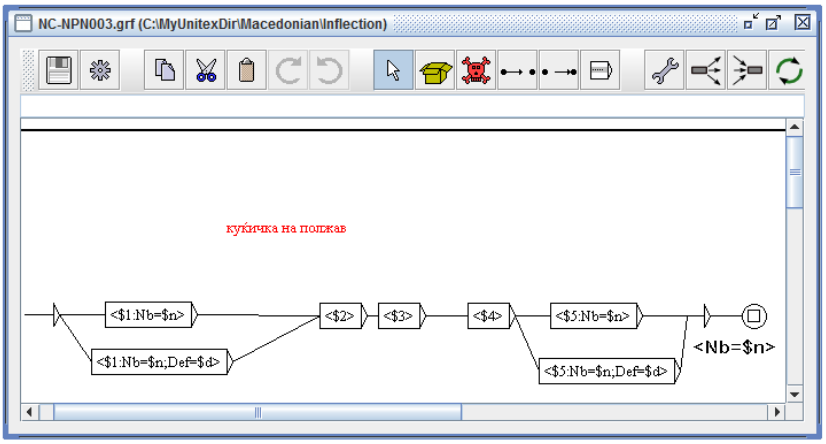

Figure 7 : Un graphe de flexion de mots composés du type NPN

5.2 Dictionnaire macédonien de formes fléchies des mots composés

La Figure 8 montre une partie du dictionnaire DELACF (DELA de formes Composées Fléchies) obtenu automatiquement par l'application des graphes au dictionnaire des mots composés Figure 5. À la gauche, toutes les formes fléchies des expressions composées sont citées, suivies de la forme lexicale, et des codes flexionnels.

Toutes les expressions composées nominales représentées sur Figure 8 ont été extraites du corpus. Leur forme lexicale est représentée sur le côté droit, avec la règle d'annotation utilisée pour obtenir la forme fléchie. Les formes fléchies obtenues en utilisant cette règle sont représentées sur le côté gauche du dictionnaire.

Il est à noter que l'expression composée nominale 'акустична цевка' (akustichna cevka) 'tuyau acoustique', apparaît dans sa forme plurielle 'акустични цевки' (akustichni cevki), 'tuyaux acoustiques', dans le roman de Verne. De même, 'бакарни инструменти' (bakarni instrumenti) 'instruments de cuivre', 'библиски времиња' (bibliski vreminja) 'temps bibliques', etc. y sont au pluriel, mais nous avons utilisé la forme au singulier en tant que forme lexicale citée dans le dictionnaire des mots composés.

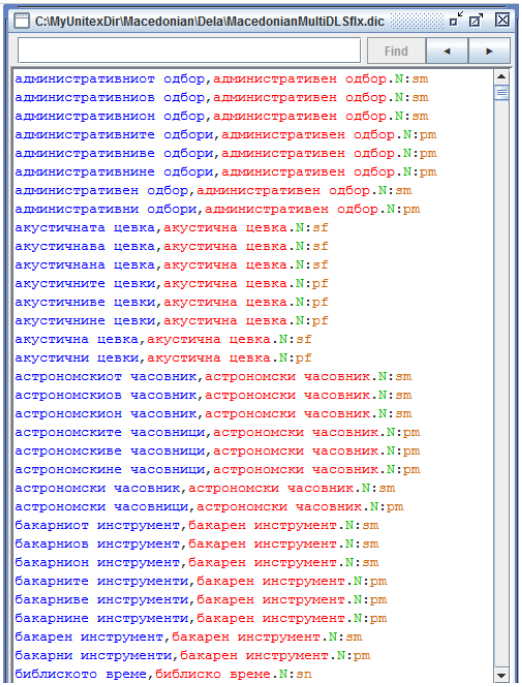

Figure 8 : Extrait du dictionnaire de formes fléchies des mots composés

6 Conclusion

Nous avons utilisé Unitex pour créer un dictionnaire et des graphes de flexion de mots simples, afin de pouvoir utiliser ces ressources pour pouvoir représenter les mots composés et d'obtenir leur flexion automatiquement. Par la suite, nous avons créé le dictionnaire des mots composés du macédonien et grâce aux graphes de flexion nous avons compilé le dictionnaire de toutes les formes fléchies des mots composés. L'examen linguistique montre que les graphes couvrent toutes les formes fléchies des mots simples et composés. Il est assez facile d'étendre cette ressource Unitex à d'autres corpus. Les actions nécessaires consistent à ajouter une annotation manuelle de la forme lexicale des mots simples et de lui associer un graphe de flexion et, si nécessaire, de modifier le graphe ou d'en créer un nouveau. Ensuite, ajouter la forme lexicale du mot composé et de lui associer le graphe correspondant, ou si nécessaire, de le modifier ou d'en créer un nouveau. Ainsi toutes les formes fléchies serraient-elles obtenues automatiquement.

Références

GROSS M.(1986). Lexicon-Grammar. The representation of compound words. Actes de *Eleventh International Conference on Computational Linguistics*, 1-6.

LEON J.(2004). Lexies, synapsies, synthèmes: le renouveau des études lexicales en France au début des années 1960. *History of Linguistics in Texts and Concepts*, 405-418.

PAUMIER, S. (2006). Unitex 3.0 User Manual. *Université Paris-Est*.

SAG I. A., BALDWIN T., BOND F. COPESTAKE A., FLICKINGER D.(2002). Multiword Expressions: A Pain in the Neck for NLP. *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing*.

SAVARY, A. (2008) "Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches", *Linguistic Issues in Language Technology*, 1(2):1–53.