La lemmatisation de l'arménien occidental avec NooJ

Anaid Donabedian-Demopoulos, Nisan Boyacioglu

La lemmatisation de l'arménien occidental avec Nooj

Anaïd Donabédian et Nisan Boyacioglu
(INALCO- CNRS FRE2454)

Note des auteurs : Ceci est la version longue de l'article à paraître dans les actes du Workshop Intex-Nooj de Tours (juin 2004). Elle rend compte de l'état d'avancement au 20/02/05 du projet entamé avec Nishan Boyacioglu dans le cadre de son DEA en 2004

1. Introduction

L'arménien occidental : état des lieux

L'arménien moderne s'est constitué au XIXème siècle sous la forme de deux langues littéraires : l'arménien oriental, parlé aujourd'hui en Arménie, en Iran et marginalement en Inde, et l'arménien occidental, né dans l'empire ottoman, qui est la langue de la diaspora issue de cet empire. L'intercompréhension entre locuteurs est relativement aisée, même si elle demande parfois une certaine adaptation. Chacune des variantes a développé sa propre tradition littéraire, sans pour autant rompre une tradition d'enrichissement mutuel, plus ou moins facilitée par les conditions politiques : la période soviétique a favorisé une radicalisation des différences, alors que les contacts développés depuis l'indépendance entraînent une plus grande perméabilité des variantes l'une à l'autre.

En tant que langue de diaspora, l'arménien occidental évolue en effet dans un contexte de contact linguistique accru (Donabedian : 2001), ce qui ne facilite pas la normalisation de la langue parlée. La tradition littéraire est cependant maintenue, grâce à l'activité des écrivains, journalistes et pédagogues, investis du rôle que jouerait une institution de politique linguistique dans un contexte étatique (académie ou autre instance de normalisation). Cependant, les variations stylistiques, qu'elles soient dues à l'évolution rapide de la norme en diaspora, ou à la grande variété dialectale de l'arménien (on compte entre 50 et 70 dialectes arméniens, selon les modes de description) apparaissent même dans les textes écrits, notamment dans les passages dialogués.

Les outils de référence pour l'arménien occidental (dictionnaires, grammaires, corpora, etc.) font largement défaut, enfermant ainsi la langue dans une circularité dangereuse entre le manque d'outils de référence et l'instabilité de la norme, qui elle-même rend de plus en plus difficile l'élaboration d'outils de référence. Ainsi, l'enseignement de l'arménien occidental souffre-t-il d'une sorte de crispation normative, qui ne favorise pas l'acquisition de la langue parlée (cf. Donabedian 2001 b). La nécessité

d'une approche descriptive permettant de constituer un socle fondamental pour des outils de référence plus proches de la réalité semble évidente, et dès 1991, nous avons cherché à appuyer nos travaux de linguistique descriptive sur l'exploitation systématique d'un corpus informatisé. Le manque de moyens dont disposent les études arméniennes en général explique qu'une telle initiative soit si longue à concrétiser. Le corpus réuni, saisi et étiqueté dans Donabedian 1991, constitué d'environ 700 pages (1,4 Mo de texte brut) de textes littéraires publiées au cours du XX^{e} siècle, avait été réalisé sous le logiciel SATO[1] avec des moyens qu'il faut bien qualifier d'artisanaux, et les nombreux aménagements *ad hoc* que l'approche avait exigée[2] ont fait que ce corpus n'a pu être mis à la disposition de la communauté scientifique : il a essentiellement servi à dépouiller le corpus pour émettre des hypothèses sur le fonctionnement de l'article défini en arménien occidental. Depuis, plusieurs projets ont vu le jour, avec des finalités diverses. A notre connaissance, ils sont au nombre de quatre, dont deux concernant l'arménien occidental.

1. Orientée vers l'arménien classique et moyen (textes arméniens du V^{e} siècle au Moyen-Age) la *Leiden Armenian DataBase* (Pays-Bas) coordonnée par J.J. Weitenberg consiste en un corpus de textes annotés utilisables par les linguistes et les historiens. Le projet est le seul abouti et diffusé à ce jour, mais il concerne la langue ancienne.

2. Le projet *Armenian Digital Library*, initié par l'équipe de Meroujan Garabedian à Erevan (Arménie), vise à constituer une bibliothèque numérique arménienne exhaustive, et la démarche étant chronologique, les textes modernes ne sont pas encore traités ; par ailleurs, ce projet a avant tout une visée encyclopédique, d'archivage et de recherche documentaire. Même s'il permet des concordances via un moteur de recherche, l'étiquetage linguistique des textes ou l'élaboration d'un lemmatiseur n'est pas prévue à ce stade du projet. Le projet est diffusé sur www.digilib.am et mis à jour au fur et à mesure de son évolution.

3. *Armenian Lexicon and Digital Library Project*, initié par Michele Sigler (King's College, London), concerne spécifiquement l'arménien occidental. Il s'agit d'un projet ambitieux destiné à constituer à la fois des outils de référence (dictionnaire), des produits destinés au grand public, et un outil de travail pour les recherches linguistiques. Le projet est encore à

[1] Système d'Analyse de Textes par Ordinateur, développé par François Daoust pour l'Université du Québec à Montréal.

[2] Le corpus était saisi sous DOS, avec une interface graphique pour l'arménien, mais la nécessité de supprimer les majuscules pour réduire le nombre de caractères utilisés. Par ailleurs, les limitations dans le nombre de 'mots' que SATO pouvait traiter nous avaient contrainte à segmenter les bases et les désinences, provoquant de nombreux cas d'ambiguité. SATO proposait des procédures pour désambiguiser en fonction du contexte, mais cela rendait la procédure d'interrogation du corpus (concordances portant sur des segments étiquetés) très complexe (cf. Donabedian 1993).

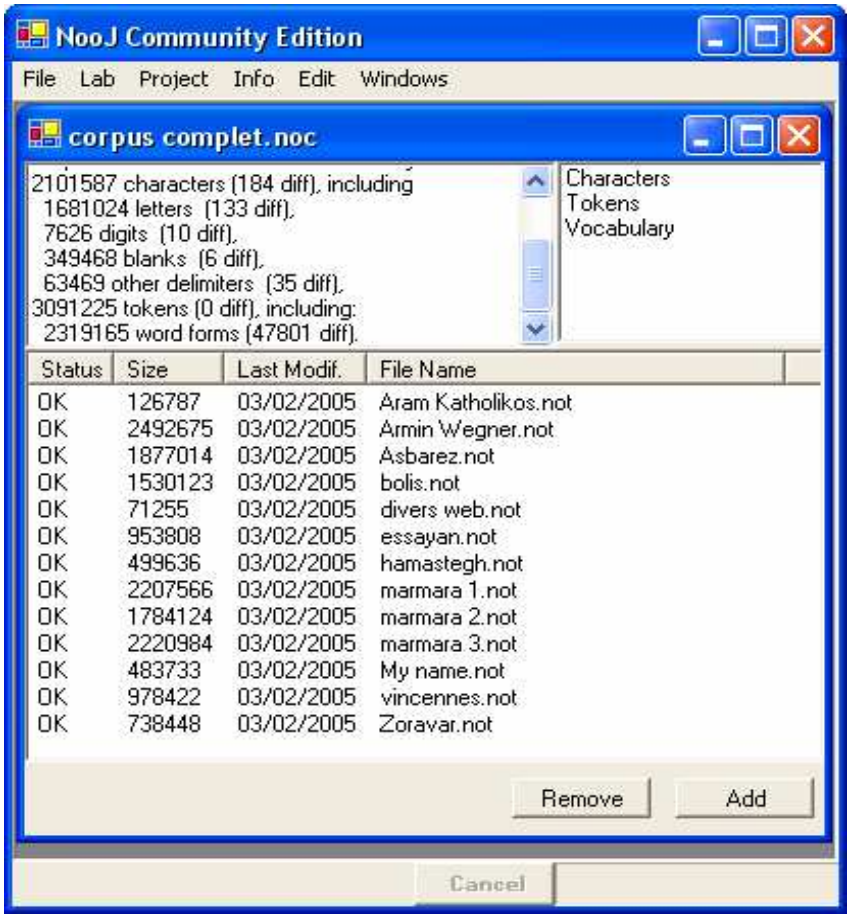
un stade préliminaire, et prévoit que le corpus sera disponible sur le web, avec un moteur de recherche sur une plateforme propre, et une interface web destinée à permettre le travail collaboratif (notamment enrichissement du corpus) est disponible. Le corpus est en cours de constitution, et la lemmatisation n'est pas encore réalisée.

4. D. Lonsdale et Inna Danielyan travaillent sur une implémentation de la morphologie de l'arménien occidental, à base d'un corpus (la Bible, choisie pour sa facilité d'accès) romanisé selon une transcription ad hoc ; ce projet, destiné à fonctionner à la fois en génération et reconnaissance, a déjà fait l'objet de communciations succintes, mais nous ne savons pas s'il est déjà efficient, et il n'est en tout cas pas disponible pour la communauté scientifique.

Un module arménien sous INTEX/Nooj

L'existence d'un module arménien nous Nooj, et donc d'une interface adaptée aux besoins du linguiste, a permis d'engager très rapidement un travail proprement linguistique, sans nécessité de constituer une équipe pluridisciplinaire et de mobiliser des financements lourds comme pour la *Leiden Armenian Database* ou pour le *Armenian Lexicon and Digital Library Project*. Les ressources utilisées ont permis une mise en œuvre rapide du projet :

Le corpus est constitué de certains éléments du corpus Donabedian 1991 (textes littéraires), de textes numérisés mis à notre disposition par le journal Asparez en 1995, ainsi que de divers textes téléchargés sur Internet, de statut divers (presse, traduction, textes littéraires). Il est utilisé uniquement pour des besoins internes à l'équipe, et dans un but opérationnel (vérification des outils linguistiques, des dictionnaires et des grammaires). En vue d'une mise à la disposition du public, le corpus sera ultérieurement enrichi et remanié en tenant compte des droits d'auteur. Les textes recueillis comportent encore un assez grand nombre de fautes (erreurs de segmentation, coquilles, intrusion de courts passages en arménien oriental, notamment dans la presse, etc.), qui sont apurées au fur et à mesure du traitement, et nous estimons que parmi les 47800 mots différents que compte actuellement le corpus, 2000 à 3000 pourraient actuellement résulter de ces fautes.



La description morphologique du verbe et le dictionnaire des verbes avaient été élaborés par N. Boyacioglu dans le cadre de son mémoire de maîtrise, consistant en une description des modèles de flexion verbale, associée à un index de 3250 verbes classés par type flexionnel, et présenté comme un guide de conjugaison de type « Bescherelle » (Boyacioglu : 2003). Le dictionnaire a été construit à partir de l'inventaire des entrées verbales du dictionnaire arménien–français de K. Chahinian (Beyrouth, 1997), légèrement augmenté.

De fait, grâce à ces ressources disponibles, le projet de lemmatisation sous Nooj, initié en décembre 2003, a permis d'ores et déjà d'aboutir en septembre 2004 à la lemmatisation des formes verbales[3] d'un corpus de 6 millions de mots avec un taux d'erreur d'environ 3%, et ce taux d'erreur a été abaissé à moins de 1% en janvier 2005.

[3] La notion de formes verbales couvre ici toutes les variations régulières des entrées lexicales verbales , c'est-à-dire les formes finies (conjuguées) et les formes nominales du verbe avec leur flexion (de type nominal).

2. L'arménien occidental : spécificités de traitement

L'alphabet arménien

- Lettres

L'alphabet arménien, dont la création est attribuée au moine Mesrop Machtots en 401 de notre ère, comptait originellement 36 lettres, deux autres lettres lui ayant été ajoutées au XII[e] siècle : օ [o] permettant de noter la diphtongue *aw* devenue [o] long, et ֆ [f], transcrivant un phonème étranger à l'arménien, mais apparu dans la langue par le biais d'emprunts étrangers. Les 38 lettres actuelles (31 consonnes et 7 voyelles) de l'arménien sont codées de 1329 à 1414 dans Unicode :

majuscule, minuscule	Translittération	majuscule, minuscule	Translittération
Ա ա	a	Մ մ	m
Բ բ	b	Յ յ	y
Գ գ	g	Ն ն	n
Դ դ	d	Շ շ	š
Ե ե	e	Ո ո	o
Զ զ	z	Չ չ	č`
Է է	ē	Պ պ	p
Ը ը	ə	Ջ ջ	ĵ
Թ թ	t`	Ռ ռ	r pointé
Ժ ժ	ž	Ս ս	s
Ի ի	i	Վ վ	v
Լ լ	l	Տ տ	t
Խ խ	x	Ր ր	r
Ծ ծ	c	Ց ց	c`
Կ կ	k	Ւ ւ	w
Հ հ	h	Փ փ	p`
Ձ ձ	j	Ք ք	k`
Ղ ղ	ł	Օ օ	ō
Ճ ճ	č	Ֆ ֆ	f

Il faut ajouter à cet inventaire le digramme և, qui n'existe qu'en minuscule et correspond à une ligature de ե (e) et de ւ (w). En tant que mot indépendant, c'est une conjonction de coordination ('et'). Ailleurs, son usage peut varier selon les normes en vigueur : en effet, en arménien occidental (qui est l'objet de notre travail), il n'apparaît jamais hormis dans le cas de la conjonction de coordination, et donc a fortiori à l'intérieur des mots. En arménien oriental (la norme en vigueur notamment en Arménie), la ligature remplace librement les deux lettres qui la composent dans des positions diverses, y compris à l'intérieur des mots.

Cette lettre n'est donc jamais oligatoire, et toujours substituable par la séquence եւ.On pourrait donc procéder à un remplacement automatique à l'entrée (dans la macrocommande initiale de lissage) ou recourir à une grammaire dans NooJ posant l'équivalence : ու = ե + ւ. A ce stade du projet, le choix n'a pas été fait, et le critère déterminant sera d'évaluer si l'usage de ce digramme peut, lui-même, constituer un objet d'étude. Auquel cas, on pourrait, par exemple envisager d'introduire ce digramme directement dans l'alphabet.[4]

- Ponctuation (délimitateurs)

Les signes de ponctuation, même s'ils ont pour la plupart la même forme que des signes de ponctuation français ou anglais, peuvent avoir un emploi différent :

désignation	**graphie**	**équivalent fonctionnel français**
virgule	,	virgule
point médian	.	point-virgule, deux points ou point
point final	:	point
apostrophe	՝	apostrophe
բունթ (*bouth*) 'pouce'	՝	pas d'équivalent

La virgule sépare les unités de même rang (syntagmes nominaux lors d'énumérations, propositions), comme en français. Elle est cependant utilisée moins systématiquement qu'en français pour délimiter les adverbes de phrase situés en début d'énoncé. Il n'existe pas de virgule spécifique à l'arménien dans unicode, et le signe utilisé est donc le même qu'en latin standard (délimiteur 44).

Le point médian peut séparer deux propositions indépendantes, qui seront ainsi présentées comme plus liées que si elles étaient séparées par un point final. Il peut aussi introduire une énumération, une explication, une citation. Il s'agit donc d'un délimiteur plus faible qu'en français, mais il n'a pas de

[4] Il existe un autre digramme en arménien : n + ւ valant [u]. De fait, en Arménie, du fait de la réforme orthographique, ce digramme est considéré comme une lettre à part entière et prend la place de ւ dans l'alphabet (ւ n'étant pas utilisé en-dehors de ces digrammes en orthographe réformée). Cependant, il s'agit là d'un simple digramme, et le fait de le représenter comme un seul acaractère ou non relève d'un simple choix technique et ne peut pas présenter un intérêt en soi pour l'analyse lingusitique. La transformation automatique ne présente donc aucun inconvénient.

signe spécifique à l'arménien dans unicode, où il est représenté par le délimiteur 46, comme le point final du latin standard[5].

Le point final (deux points) est la ponctuation la plus forte. Elle achève un énoncé. Dans ce cas, la différence de valeur a été prise en considération dans l'encodage unicode, puisque le signe utilisé (del 1417) est spécifique à l'arménien, les deux points latin standard étant codés 58[6].

L'apostrophe marque, comme en français, l'élision d'une voyelle. L'apostrophe arménienne (del 1370) pourrait cependant facilement être confondue avec le guillemet américain simple de fin de citation courbe (del 8217) ou droit (del 39), et tout autant avec le signe d'intonation arménien d'injonction (let 1371) que nous évoquons plus bas.

Le « bouth » (del 1373) combine les usages français des deux points, de la virgule, et d'un signe de ponctuation à valeur syntaxique (évoquant l'emploi du tiret en russe) notamment pour introduire une proposition dont le verbe a été élidé (dans des énoncés du type : '*J'ai acheté une montre, et lui ՝ un chapeau*'). Formellement, il peut être confondu avec le guillemet américain simple de début de citation courbe (del 8216) ou droit (del 96).

Les guillemets et traits d'union sont identiques à ceux utilisés en français. L'arménien, comme le français utilise en principe les guillemets en chevrons (« » , del 1741 et 187), cependant, notamment dans des textes édités dans des pays anglophones, on rencontre les guillemets américains.

Les ressemblances formelles des signes de ponctuation évoquées ci-dessous se reflètent dans les textes de notre corpus qui sont hétérogènes dans l'emploi des signes de ponctuation. Pour des raisons probables de commodité de clavier, certains textes utilisent les apostrophes ou les points latins au lieu des caractères à codage spécifiquement arménien. On observe même des pratiques hétérogènes au sein d'un même texte.

A ce stade du projet, nous avons seulement identifié ces ambiguités. Une macrocommande destinée à un lissage des textes avant même leur incorporation au corpus reste à élaborer.

---

[5] Cela signifie qu'en cas d'utilisation de ressources non-spécifiques à l'arménien incluant des critères de ponctuation, ou en cas d'utilisation de corpus bilingues, il faudrait tenir compte de cette différence de valeur selon la langue.

[6] Les normes typographiques les distinguent également : contrairement à *del 58*, *del 1417* n'est pas précédé d'un blanc typographique. Cela pourrait être une des raisons pour lesquelles ce signe, contrairement au point *del 46*, bénéficie d'un code spécifique en arménien, bien qu'il ait la même forme que les deux points français.

- Une spécificité arménienne : les signes d'intonation

Une dernière catégorie de signes, que les grammaires traditionnelles classent parmi les signes de ponctuation, a un fonctionnement spécifique à l'arménien, de nature à perturber la lemmatisation. En effet, il sont placés à l'intérieur du mot, immédiatement après la voyelle accentuée du mot (soit la dernière voyelle hors clitiques). De ce fait, ils sont traités dans unicode comme des lettres. Leur fonction est de marquer la modalité assertive de l'énoncé, et ils ne dispensent pas de clore la phrase avec un signe de ponctuation à proprement parler (cf. ci-dessus) :

désignation	**graphie**	**fonction**
պարույկ (paruyk) 'cercle'	՞	interrogation
շեշտ 'šešt' 'accent'	՛	injonction ou contraste
երկար (*yerkar*) 'long'	՜	exclamation vraie

Le signe šešt (let 1371) marque une hausse de l'intensité, et une légère hausse du fondamental, mais sans allongement. Il est un signe d'injonction ou de contraste (aussi appelée emphase). On le rencontre donc avec les impératifs, mais aussi dans les cas d'emphase (focalisation ou contraste) : Ե՛ս եկայ (moi venu) C'est **moi** qui suis venu. Mais aussi : Այո՛ : *Oui !*, - Ո՛չ : *Non !*. Il peut être traduit par un point d'exclamation en français.

Le yerkar (let 1372) traduit un allongement de la syllabe qui le porte et correspond à une modalité d'admiration, d'affection, de tristesse, de regret, de surprise, etc. C'est ce que nous appelons l'exclamation vraie, et bien qu'en français, ce signe se traduise égamlement par un point d'exclamation, il ne peut en aucun cas être confondu avec le yerkar vu ci-dessus.

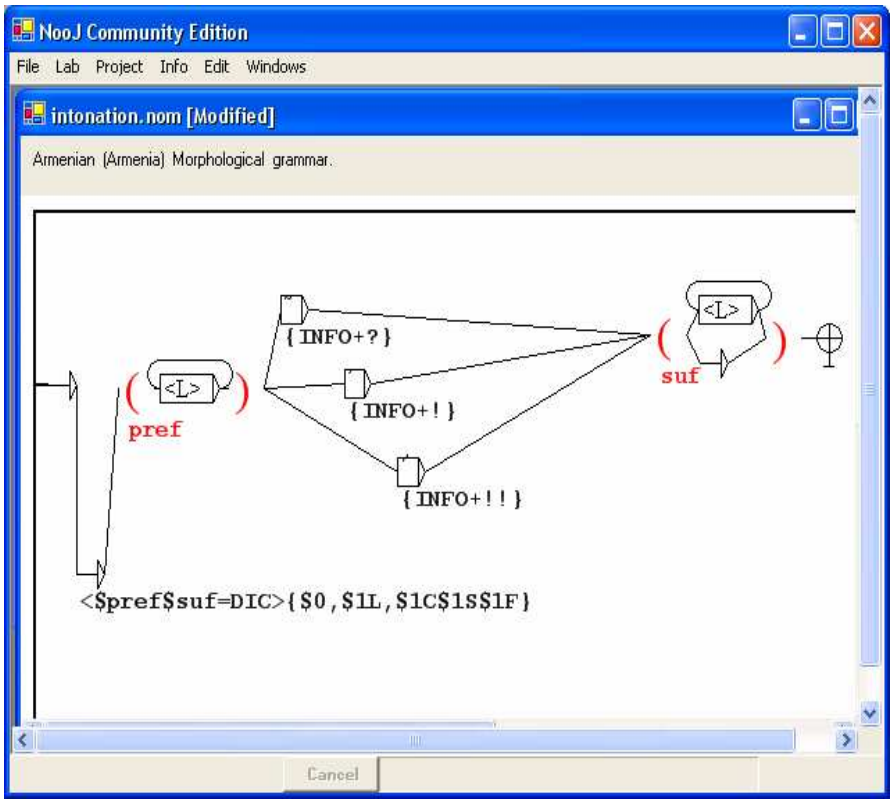
Ex. : Խե՜ղճ բատեկամ : ***Pau**vre ami !*

Le paruyk (let 1374) est un signe d'interrogation, il correspond à une élévation du fondamental : Գո՞հ ես: (content es) *Es-tu content ?*. Il est équivalent au point d'interrogation du français (*?*) hormis le fait que contrairement à ce dernier, il ne clot pas la phrase, qui doit toujours se terminer par un point final (:).

Les signes d'intonation sont encodés par Unicode comme séparateurs, ce qui est une erreur, car cela conduit à segmenter le mot qui les porte en deux. Le module arménien de Nooj a corrigé cette erreur en en faisant des lettres. Cependant, en vue de la lemmatisation, ils démultiplient potentiellement non seulement le lexique, mais aussi toutes les formes

fléchies (les désinences casuelles ou verbales qui ne sont pas clitiques peuvent porter ces signes).

Pour permettre au lemmatiseur d'identifier une forme avec signe d'intonation à la même forme sans signe d'intonation, en récupérant toutes ses propriétés et en ajoutant une propriété correspondant à sa valeur, nous avons construit le graphe suivant (où les valeurs inhérentes à ces signes sont représentées par un point d'interrogation (*paruyk*), un point d'exclamation (*šešt*) ou par une double point d'exclamation (*yerkar*) :



Ainsi, une forme comme : միաւորէի՞նք 'unissiez ?' est identifiée à une forme présente dans le dictionnaire fléchi (<$pref$suf=DIC>), միաւորէինք,միաւորել,V+FLX=V2+I+1+p dont les propriétés sont récupérées par la commande {$O,$1L,$1C$1S$1F, et la propriété additionnelle **+ ?** est appliquée par {INFO+ ?}.

La morphologie de l'arménien

L'arménien présente une morphologie grammaticale de type mixte, agglutinant et flexionnel. Le type agglutinant, qui concerne essentiellement les substantifs, est foisonnant (36 formes par lemme) mais relativement régulier (avec cependant quelques modèles irréguliers hérités d'un état plus

ancien de la langue). Le type flexionnel concerne les pronoms et les verbes. La morphologie y est moins riche en termes de nombre de morphèmes (les morphèmes flexionnel sont fusionnels), mais beaucoup moins régulière, et la flexion verbale présente de nombreux cas d'irrégularités, de défectivité, et de supplétivité (la description compte actuellement 52 types flexionnels). Cependant, certains éléments agglutinants sont représentés dans le verbe, notamment les infixes de causativation et de passivation, qui multiplient par quatre les tiroirs verbaux (actif, passif, causatif-actif, causatif-passif).

La morphologie lexicale (dérivation et composition) est également très riche, avec un grand nombre de procédés productifs.

La labilité des catégories

Un autre phénomène marquant pour le traitement de l'arménien est la perméabilité des catégories (parties du discours). Selon les critères adoptés, l'inventaire obtenu et le classement d'un mot donné seront très variables. Si on s'appuie uniquement sur le type flexionnel adopté, on peut en principe distinguer assez clairement noms, verbes et pronoms. Cependant, cela ne signifie pas que les catégories grammaticales représentées par ces lemmes sont totalement imperméables. Outre la catégorie du nombre, qui est fréquemment commune au nom et au verbe dans les langues, ou la catégorie de la personne, commune au pronom et au verbe, on observe des phénomènes frappants de transcatégorialité. Tout d'abord, la classe adjectivale n'a pas de contours précis. Selon le critère morphologique, l'adjectif entre dans la catégorie des noms, car il n'existe pas de flexion proprement adjectivale comme dans la plupart des langues flexionnelles. Epithète ou attribut, l'adjectif est invariable en arménien, et c'est sa position immédiatement à gauche du nom qui détemine sa fonction. Cependant, la nominalisation est courante, par adjonction d'une désinence flexionnelle nominale. De même, il est possible d'utiliser dans la position adjectivale un nom pour dénoter une propriété. Ainsi, si les adjectifs dénotant primairement une propriété (couleur, dimension, etc.) et des adjectifs relationnels dénominatifs (comme հայր-ական *'paternel'*, ատոմ-ային *'atomique'*, etc.) peuvent être étiquetés aisément, pour la plus grande partie du lexique, il est souvent très difficile de dissocier une classe de noms et une classe d'adjectifs en lexique.

Par ailleurs, des unités appartenant à la même classe distributionnelle peuvent avoir un comportement morphologique très différent : ainsi, les postipositions peuvent être totalement invariables (պէս *'comme'*), avoir une morphologie nominale (վրայ *'sur'*, տակ *'sous'*), ou être combinables avec les seuls articles possessifs (հետ *'avec'*). Il en va de même pour les adverbes : certains sont des formes invariables (notamment les adverbes

dérivés en –պէս, –օրէն '–ement'), d'autres sont une forme fléchie de postposition ou de nom, analysable comme une forme fléchie.

Ainsi, les désinences de type nominal peuvent apparaître sur des types très variés de lexèmes, à l'exception des pronoms et des formes personnelles du verbe. Cela concerne certains mots-outils comme on l'a vu pour les postpositions, les formes adverbiales, les nominalisations d'adjectifs, mais aussi un champ très large de nominalisations. C'est notamment le cas des formes participiales du verbe : deux d'entre elles sont invariables et entrent seulement en composition dans des formes verbales analytiques, mais quatre autres (le participe agentif en –ող, le participe passé en –ած, le participe prospectif en -իք et l'infinitif) ont une combinatoire nominale complète, ce qui augmente de 144 (4x36) le nombre de formes pour un lemme verbal.

L'ensemble de ces combinatoires peut être prévu par les grammaires flexionnelles, même si elles en sont considérablement alourdies (ainsi, du fait notamment des quatre diathèses et des formes fléchies des participes, on atteint 808 formes par verbe régulier). Mais d'autres cas sont plus difficiles à traiter, comme la surdéclinaison et/ou l'emploi autonymique.

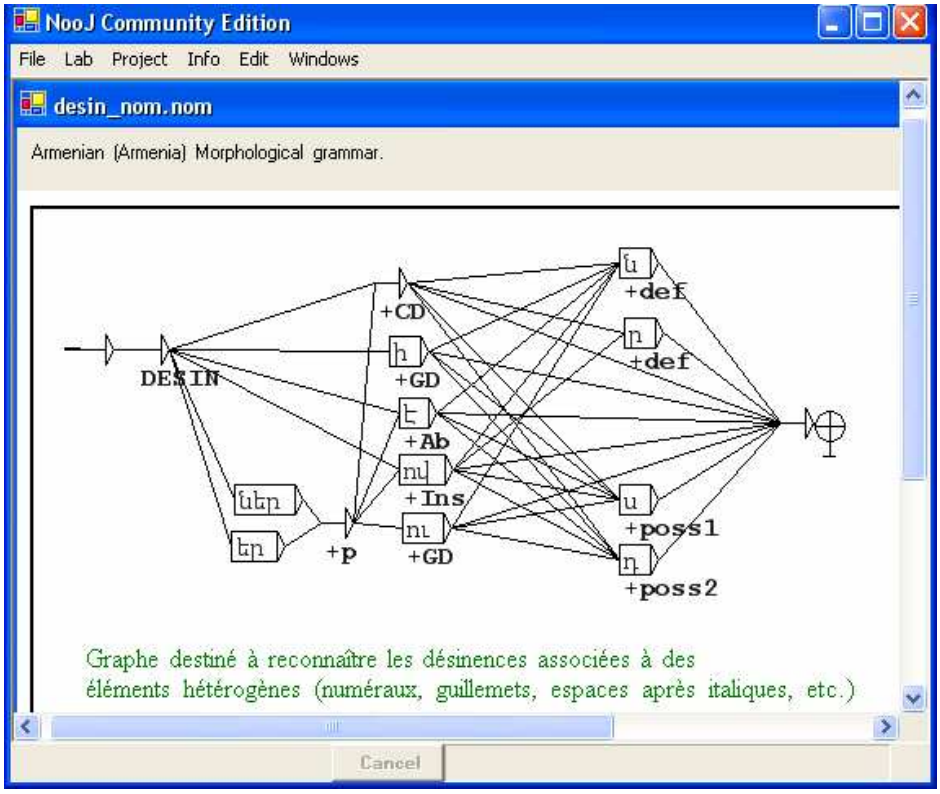
La surdéclinaison permet de former une forme fléchie non pas sur une base nue, mais sur une forme elle-même fléchie, dont elle résulte de la nominalisation : ex. Դիմաց – ի-ն – ներ - ը
en-face G-déf plur déf
Ceux d'en face. (litt. Les d'en face)

Ce phénomène peut aussi résulter d'un emploi autonymique :

երեկ-ներ-դ
hier plur poss2
tes 'hier'…

Dans cet exemple, la base est nue (l'adverbe de temps signifiant 'hier'), mais il peut aussi bien s'agir d'une forme quelconque, y compris une forme personnelle du verbe ('tes « je ne veux pas »'). Ces exemples ne peuvent être traitées dans la morphologie flexionelle, et cela ne serait d'ailleurs pas souhaitable compte tenu de leur probabilité d'occurrence. Cependant, les combinaisons possibles sont infinies.

Dans certains cas, les nominalisations de type autonymique, ou bien concernant un syntagme nominal, ou encore des numéraux, sont typographiquement marquées par des délimitateurs : guillements, trait d'union, etc. Ce cas peut être traité de manière économique par un graphe morphologique qui évite que ces segments ne soient pas reconnus. Le graphe ci-dessous permet de les faire apparaître dans le vocabulaire associées à la catégorie DESIN et aux propriétés grammaticales correspondantes :



Une grammaire syntaxique pourra être développée ultérieurement afin d'imputer les propriétés grammaticales (cas, nombre, détermination) au syntagme constitué par la désinence et l'élément qui la précède.

Quels arbitrages ?

Pour ce qui est des **catégories lexicales**, cette labilité nous met donc devant un choix difficile, puisque les étiquettes sont différentes selon que l'on aborde la question d'un point de vue flexionnel, distributionnel ou sémantique. En vue de la lemmatisation, il est essentiel d'identifier avant tout les classes flexionnelles (verbe : **V,** substantif : **S,** pronom : **PRO,** invariables : **INV**). Cependant, cette information est déjà portée par le type flexionnel affecté à un lexème donné (FLX=V2, FLX=S1, FLX=PRO3, etc.), et il est souhaitable que l'étiquette 'catégorie' permette d'encoder des informations distributionnelles et/ou sémantiques (NOM vs ADJ, CONJ vs PARTIC, etc.). Ainsi, hormis la catégorie du verbe (V), l'état d'avancement actuel du projet n'a pas encore permis d'aboutir à une classification satisfaisante des unités lexicales, cette dernière devra être élaborée en tenant compte d'impératifs théoriques et pratiques à la fois (en accord avec l'usage auquel est destiné le corpus).

La question peut se poser également de savoir à partir de quel moment une forme régulièrement fléchie à partir d'un lemme pourrait constituer elle-

même un lemme indépendant. C'est là une des manifestations de la question des limites entre flexion et dérivation. Les **participes** sont des formes non finies de lexèmes verbaux, et à ce titre, ils ont des propriétés verbales et nominales à la fois. En particulier, pour la plupart d'entre eux, ils sont soumis à la flexion nominale, ce qui signifie qu'ils connaissent les catégories de cas, nombre, détermination, comme un nom. Cependant, pour l'économie de la description, il est peu satisfaisant de les traiter autrement que comme une forme verbale : tous les verbes ont des participes, leur construction suit des alternances de bases qui sont représentées également dans la conjugaison (certains participes se construisent sur une base passé, d'autres sur une base présent), ils sont compatibles avec le préfixe de négation, contairement aux noms, et ils présentent également les infixes de causatif et de passif. Aussi, il s'est avéré préférable d'introduire des sous-graphes de déclinaison nominale à l'intérieur de la grammaire verbale (quitte à inscrire dans les définitions de propriétés le cas, le nombre et la détermination comme compatibles avec la catégorie V). L'autre option, qui aurait consisté à créer un dictionnaire de participes étiquetés comme nominaux, aurait permis de faire l'économie de ce sous-graphe nominal dans la flexion verbale, mais n'aurait pas permis la reconnaissance automatique des formes négatives, ni la prise en compte de la diathèse (passif ou causatif). Elle aurait en outre rendu plus complexe la maintenance du dictionnaire verbal. Cependant, le choix que nous avons opéré a porté le nombre de formes possibles par lemme verbal à 808, et a considérablement alourdi le dictionnaire fléchi des verbes. La situation est quelque peu différente, comme on le verra plus bas, en ce qui concerne les variations de diathèse des verbes (passif, causatif, causatif-passif), car on doit distinguer les formes lexicalisées (sens non-calculable) de celles qui sont obtenues par dérivation productive (sens calculable).

Les **surdéclinaisons**, statistiquement peu nombreuses, mais virtuellement infinies, n'ont pas encore trouvé de solution, desin_nom.nom ne permettant de traiter que les formes 'balisées' par des guillements ou un trait d'union.

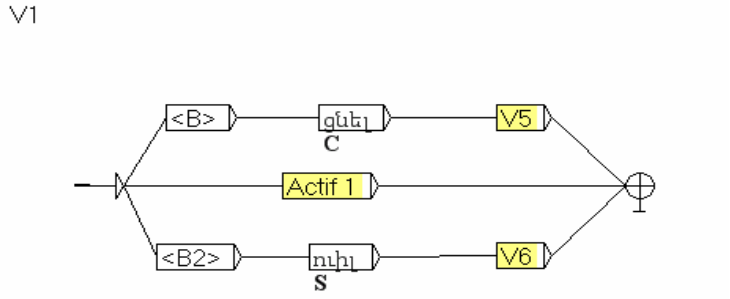
3. La lemmatisation du verbe :

Il s'agit de la partie achevée du projet. Le module qui fonctionne dès à présent consiste en grammaires flexionnelles et morphologiques :

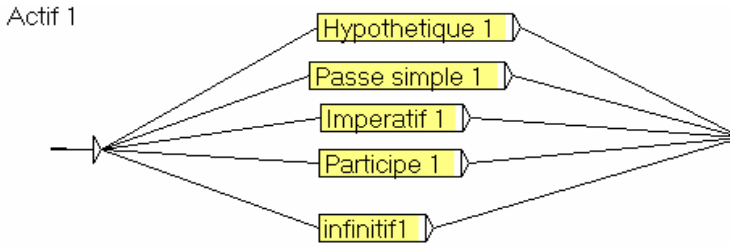
Grammaire flexionnelle

La grammaire verbale hyverb.nof comporte 86 graphes, dont 52 types flexionnels (graphes V1 à V52) et 34 sous-graphes représentant des modules qui concernent plusieurs types flexionnels, ce qui permet une certaine

économie de description. Le type V1 (verbes réguliers en –el) est ainsi présenté comme ayant trois formes, active, causative et passive, ces deux dernières étant conjuguées respectivement selon le type V5 et V6 comme le montre le graphe principal du modèle[7] :

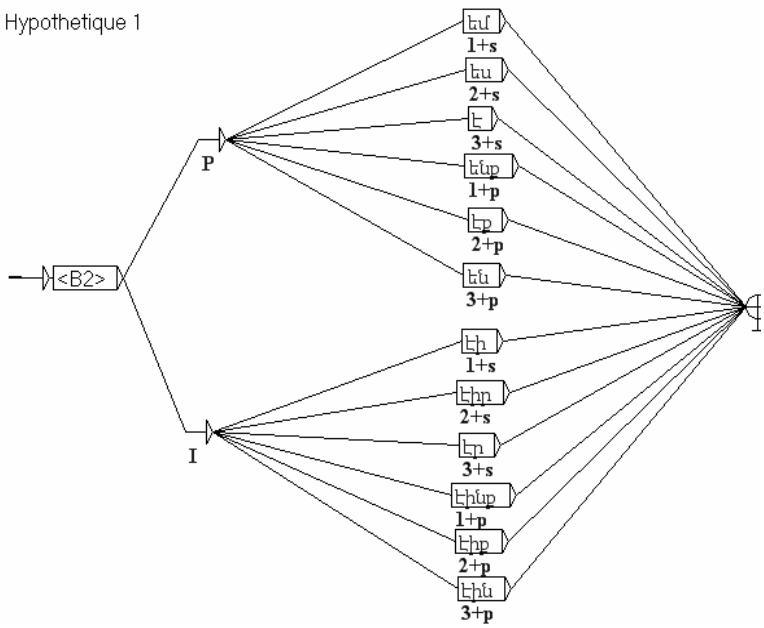


La forme active mobilise 10 graphes successivement emboîtés :

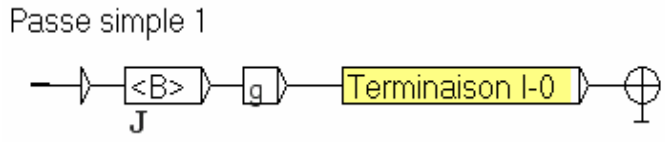


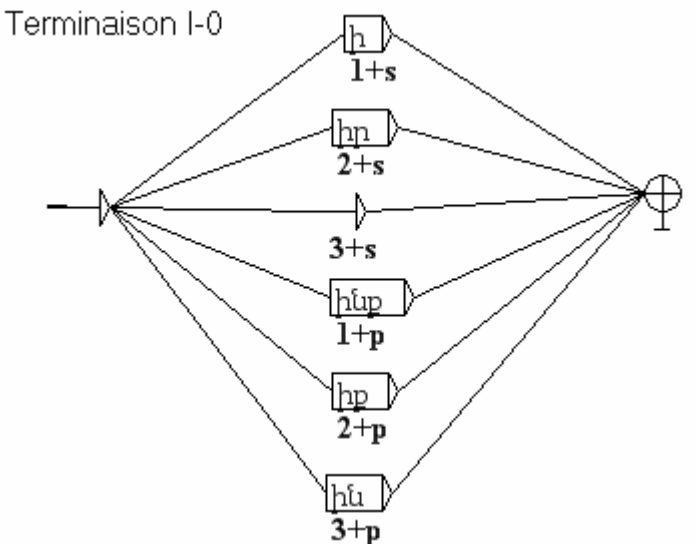
L'hypothétique est la forme permettant de construire le présent et l'imparfait par adjonction d'une particule de validation կը/կ'/կու. Il n'y a donc pas d'indicatif présent dans le graphe flexionnel, cette forme sera décrite dans une grammaire syntaxique (կը + hypothétique présent = présent de l'indicatif). Le graphe suivant décrit l'un des 4 modèles d'hypothétique possibles en arménien :

[7] Il est à noter que la flexion des causatifs intègre à son tour une possibilté de passivation, comme le permet la langue.

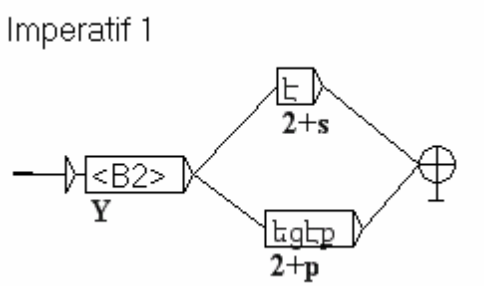


Le graphe ‘Passé simple 1’ construit la base passé à partir de l’infinitif (lemme), et renvoie à l’un des 4 modèles de conjugaison disponibles pour le passé simple en arménien :

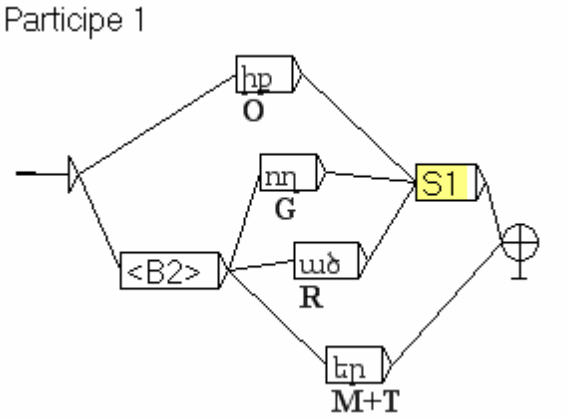




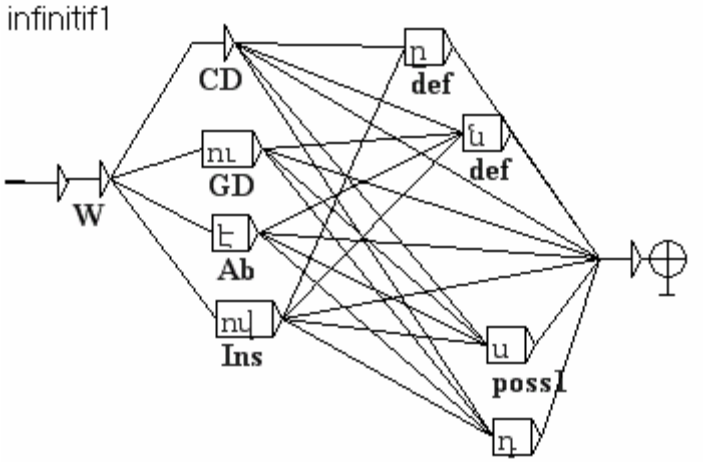
De même le modèle de l'impératif est sélectionné parmi 4 graphes d'impératif possibles :



Le graphe de formation des participes est l'un des 5 graphes de participes disponibles. On constate que hormis le participe du médiatif et de la négation analytique (codés M et T), toutes les formes participiales (G=agentif, R=résultatif, O=prospectif) se déclinent selon le modèle S1 (cf. infra flexion nominale) :



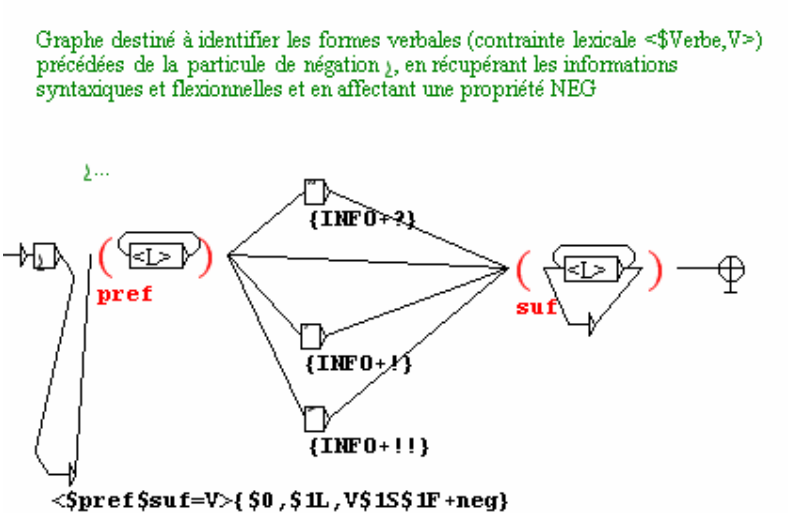
Le graphe de l'infinitif (sélectionné parmi deux modèles possibles) est aussi un graphe de déclinaison, mais à la différence des participes, l'infinitif ne suit pas le modèle de déclinaison régulier, la déclinaison est donc ici représentée intégralement,et non par un sous-graphe :



Grammaires morphologiques

- Negation verbale :

Le préfixe de négation étant combinable avec toutes les formes verbales, il a été possible de le traiter par une grammaire flexionnelle, pour ne pas alourdir inutilement les dictionnaires fléchis.



Le graphe prévoit que toute forme constituée du préfixe négatif չ suivi d'une forme identifiée comme verbale (ce qui confirme l'intérêt de considérer les participes comme des formes verbales) est interprétable comme չ + V, et se voit associer les propriétés de V. Une version initiale de ce graphe n'intégrait pas le module concernant les signes d'intonation entre (pref) et (suf), cependant, en l'absence d'articulation entre la grammaire intonation.nom et negation.nom, les formes relevant des deux grammaires à la fois n'étaient pas reconnues. La priorité de cette grammaire doit être inférieure aux grammaires flexionnelles afin qu'elle ne soit appliquée qu'aux mots pour lesquels aucune autre analyse n'est possible. Cependant, actuellement, nous n'avons pas identifié de doublons pour ce schème.

Résultats

La description de la flexion verbale a été réalisée avec le souci de privilégier la reconnaissance (pour la lemmatisation). Ainsi, nous avons prévu la génération de toutes les formes possibles, même si elles ne sont pas possibles pour tous les verbes de ce modèle flexionnel, notamment pour des raisons sémantiques. C'est le cas par exemple des dérivations passives ou causatives. Ainsi, si les verbes մտնել 'entrer', et գտնել 'trouver' suivent le même modèle flexionnel, en réalité, le causatif de գտնել 'trouver' est totalement théorique, et n'est pas attesté, à l'exception de la 3ème personne

du singulier de l'indicatif présent[8]. Pour affiner la représentation, il faudrait donc introduire des variables sémantiques pour prédire le comportement des verbes vis-à-vis de la dérivation causative ou passive.

Le deuxième problème posé par les causatifs et les passifs ramène au choix descriptif entre dérivation et flexion. Dans la mesure où il s'agit de formations productives, le choix d'un bon taux de reconnaissance a conduit à traiter la diathèse comme un mécanisme flexionnel et non dérivationnel, afin de ne pas négliger les verbes dont les formes causatives et passives sont ignorées dans les entrées des dictionnaires. Cependant, dans le dictionnaire qui a servi de base à la construction du dictionnaire des verbes, certains passifs ou causatifs plus ou moins lexicalisés figurent comme entrées lexicales. Ils ont été maintenus en tant que tels dans notre dictionnaire hyverb.dic, ce qui génère des doublons. On peut penser qu'ils sont lexicalement justifiés, l'une des formes étant un passif lexicalisé, l'autre un passif productif. Cependant, cette question, encore mal décrite en arménien, pourrait justement faire l'objet d'une étude sur corpus, destinée à vérifier la répartition entre formes lexicalisées (sens non calculable à partir de la forme active) et non lexicalisées (sens non calculable) pour ces doublons.

Taux de reconnaissance

Avec les outils décrits ici, nous avons atteint un taux d'erreur en reconnaissance d'environ 1%, ce qui est assez satisfaisant. Les formes verbales inconnues par Nooj s'expliquent par plusieurs facteurs :

Le dictionnaire se limite aux quelque 3251 verbes de la deuxième partie arménien-français, due à K. Chahinian, du *Dictionnaire pratique français-arménien arménien-français* (2ème impression, Beyrouth, éditions Chirak) le plus récent (2000) et le plus couramment utilisé par les apprenants francophones ; il s'agit d'un dictionnaire de poche d'environ 20000 mots, ce qui n'est qu'une partie du stock lexical de la langue (à titre indicatif, le dictionnaire en 4 volumes de Malkhasiantz compte environ 145000 mots). Le corpus, de taille modeste, couvre néanmoins un large champs lexical, du littéraire à la langue de la presse, en passant par des dialogues populaires.

[8] La question du causatif et du passif est rendue complexe par la possibilté idiomatique d'employer des tournures quasi-causatives et quasi-passives à la troisième personne du singulier (à valeur impersonnelle) signifiant 'cela se fait' ou 'cela ne se fait pas' : (1) կ'ուտուի 'cela se mange (quasi-passif)', (2) կը կերցուի 'cela se fait de faire manger cette chose (causatif+quasi-passif)' (3) կը գիտցուի 'cela se sait (quasi-causatif+quasi-passif)', չի բարձրացուիր 'cela ne se soulève pas', cela n'est pas possible à soulever' (quasi-causatif+quasi-passif). On constate que des marqueurs de quasi-diathèse peuvent se combiner avec des marqueurs de diathèse véritables (2)

L'arménien connaît des **variations de norme** importantes, et la grammaire flexionnelle hyverb.nof ne rend compte que de la norme standard. Les formes non-standard sont néanmoins présentes dans le corpus (dialogues, récits populaires, tradition orale des Arméniens de Constantinople). Le texte bolis.not, qui contient le recueil des traditions orales des Arméniens de Constantinople est à l'origine de la plupart d'entre eux, comme par exemple les causatifs en –c`unel au lieu de –c`nel (անցունել 'faire passer', est attesté dans le corpus en face du standard անցնել). De même, au présent et à l'imparfait de l'indicatif, la particule de validation կու, qui s'écrit séparément de la forme verbale en arménien standard (կու գամ), peut s'écrire dans certains usages accolée au verbe (կուգամ). Ce dernier point n'est cependant pas coûteux à résoudre (par le biais d'une grammaire morphologique sur le même principe que negation.nom).

4. La lemmatisation du nom

Choix préliminaires

Les outils destinés à lemmatiser les formes nominales et pronominales ne sont pas achevés, mais d'ores et déjà certains choix ont été nécessaires

- Les syncrétismes

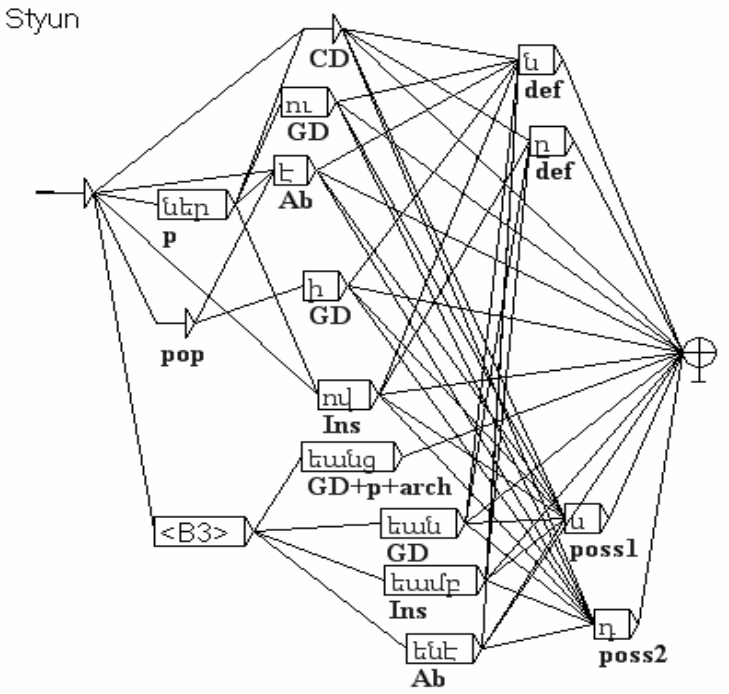
L'inventaire des cas de l'arménien varie selon que l'on s'intéresse au nom ou au pronom. Les pronoms ont 6 formes différentes (Nominatif, Accusatif, Datif, Ablatif, Instrumental, auxquels s'ajoute le Génitif, qui est généralement étiqueté comme un possessif). Les substantifs, eux, n'ont que 4 formes casuelles, le nominatif et l'accusatif ayant la même forme (Cas Direct, CD), ainsi que le Génitif et le Datif (GD). Nous avons choisi pour le moment de refléter cette disparité dans notre description, plutôt que de dédoubler artificiellement les formes de CD et GD des substantifs.

- Les variations de normes

L'arménien occidental présente une déclinaison régulière, productive, et agglutinante, et plusieurs modèles non productifs (sauf dans le cas de types de dérivés) ayant des types irréguliers de fonctionnement flexionnel (les désinences ne sont pas segmentables en nombre-cas[9]). Cependant, le modèle productif contamine, à des degrés divers, tous les modèles irréguliers, bien que la norme n'admette pas toujours ces formes reconstruites. Cependant, sachant qu'elles peuvent se rencontrer dans les textes, nous avons choisi de les faire figurer dans les graphes, en stipulant cependant qu'elles sont populaires (+pop). De la même façon, des formes

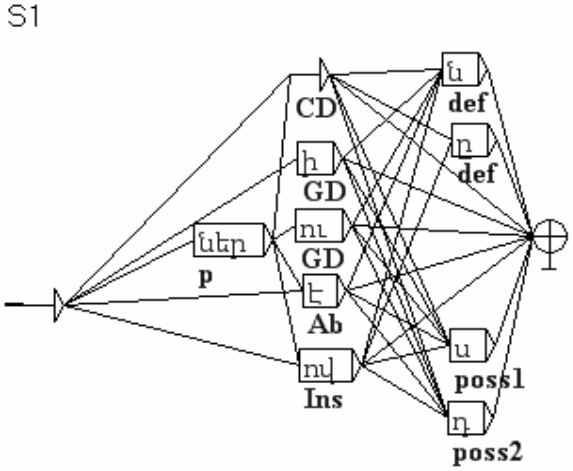
[9] L'arménien ne connaît pas le genre grammatical.

archaïques encore en usage ont été mentionnées comme telles (+arch). Dans le modèle flexionnel Styun (substantifs dérivés en –tyun), le GD et l'Ab singuliers ont une forme standard et une forme populaire, et le GD pluriel a une forme standard et une forme archaïque :



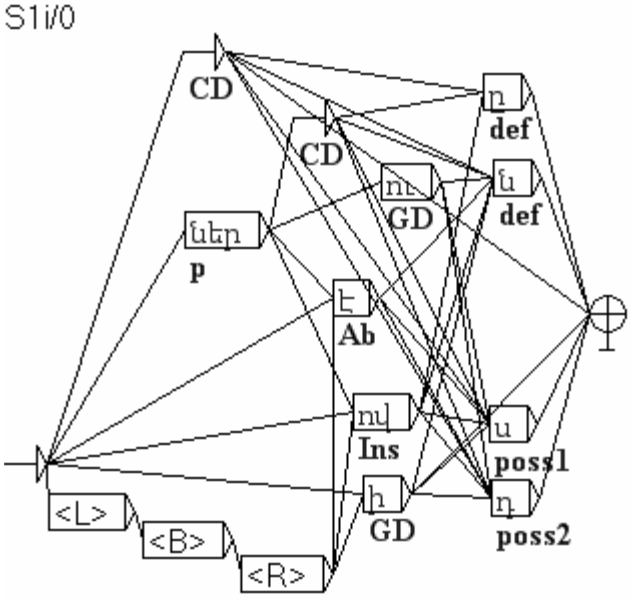
Grammaires flexionnelles et dictionnaires

Les graphes flexionnels de la grammaire du nom hynoun.nof sont donc relativement riches, puisque même hors variation stylistique, un nom régulier aura 36 formes distinctes, compte tenu de la combinaison de 4 cas, 2 nombres et 5 options de détermination :



Outre la description du modèle régulier et productif S1, nous avons d'ores et déjà réalisé les graphes flexionnels des modèles irréguliers productifs : noms monosyllabiques à marque de pluriel différentiel, noms à réduction vocalique, noms à accidents de frontières (finale vocalique, yod, etc.), modèles dérivationnels productifs (en -tyun ci-dessus, en –um). Les classes fermées ne sont pas encore décrites (classes en –an, en –or, etc.).

Le graphe S1i/o concerne par exemple les substantifs où la voyelle antépénultième est réduite dans certaines formes fléchies, comme երկիր, yerkir, 'pays', GD : երկրի, yerkri :



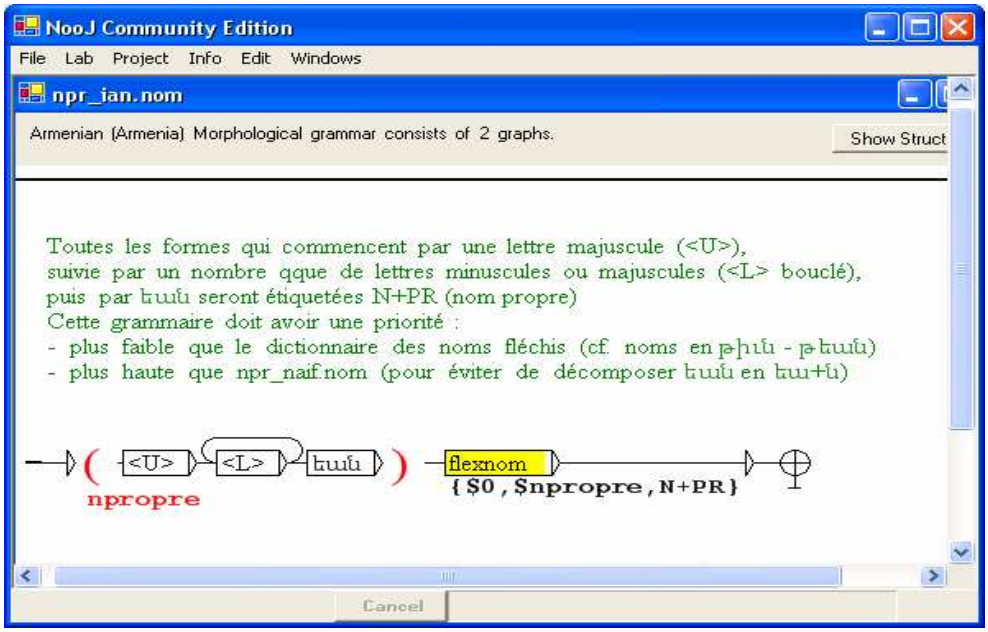
Le dictionnaire des noms n'a pas été constitué de manière systématique, mais une ébauche de dictionnaire opérationnel a été construite à partir d'une exportation du lexique du corpus Donabedian 1991.

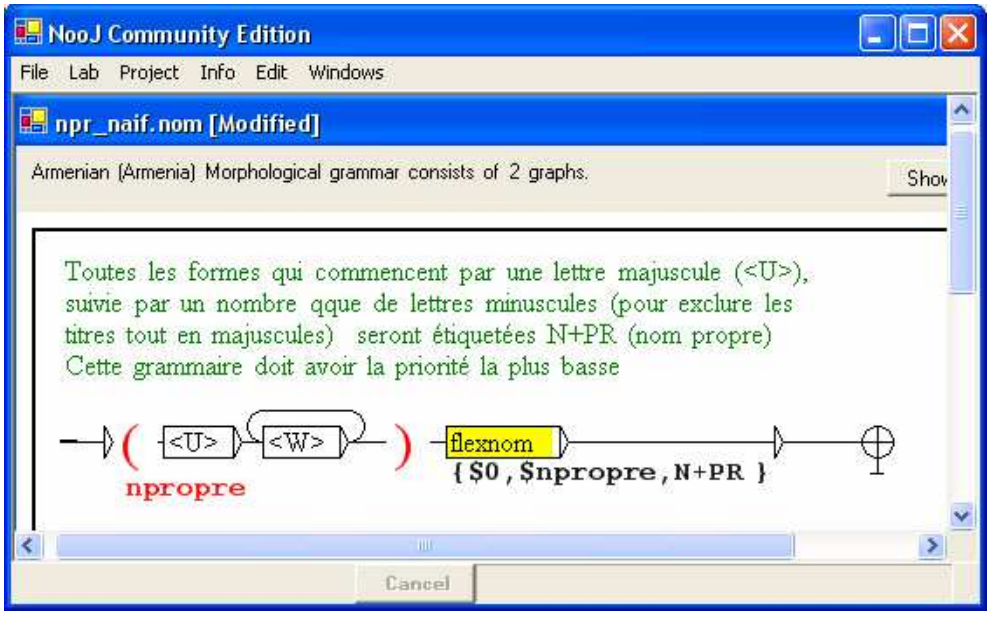
Concernant les pronoms, on peut se demander s'il est préférable de construire un graphe ou directement un dictionnaire des formes fléchies de chaque pronom, compte tenu du fait que les traits morphologiques communs dans la flexion des pronoms sont relatifs, ou plutôt, peu évidents en surface. Nous avons considéré cependant que les graphes avaient l'avantage de la lisibilité (représentation graphique, ainsi que la possibilité de regrouper tous les graphes dans une grammaire hypronoun.nof), et avons provisoirement commencé à travailler dans ce cadre.

Grammaires morphologiques

La lemmatisation du nom a recours aux mêmes graphes morphologiques que le verbe : intonation.nom (cf. supra p. 9), et desin_nom.nom (cf. supra p. 12). Pour le traitement des noms propres, source importante de non-reconnaissance, deux grammaires sont élaborées.

L'une est destinée à identifier comme un nom de famille arménien les formes commençant par une majuscule et se terminant par –ian, et qui ne seraient pas reconnues par les dictionnaires fléchis. L'autre traite le reliquat, et peut être activée optionnellement, compte tenu de son taux d'erreur prévisible. Dans les deux cas, un sous-graphe flexnom prend en charge la déclinaison nominale selon le modèle régulier S1.



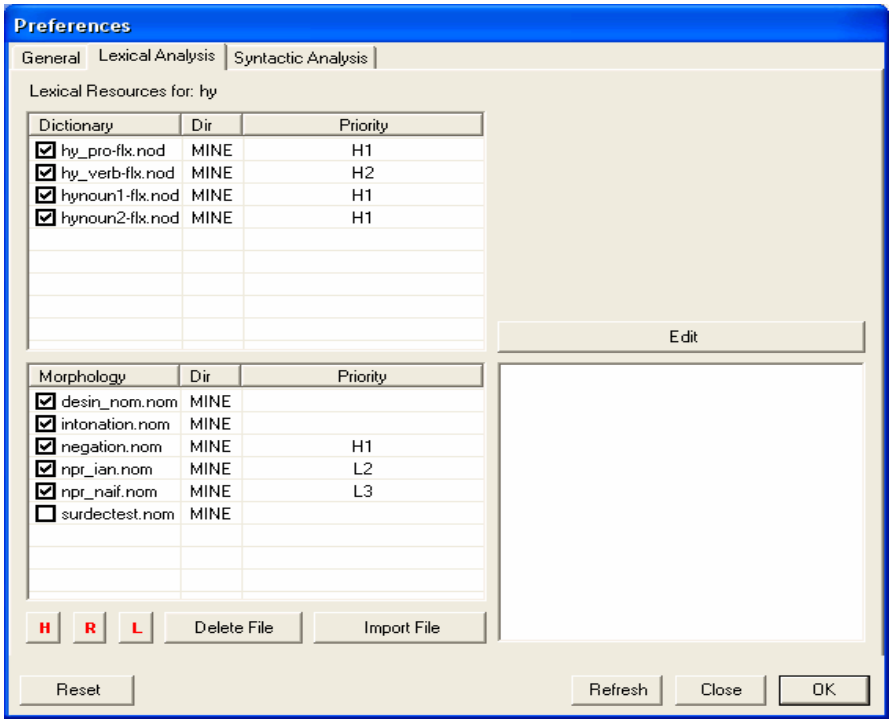


5. Résultats globaux et perspectives

Le corpus comprend actuellement 47 000 formes différentes, parmi lesquelles nous sommes parvenus à 15765 formes inconnues, et 35178 formes reconnues. Cela signifie que 3800 formes (soit environ 4%), donnent lieu à des reconnaissances en doublons, soit du fait d'une ambiguité lexicale (ce qui est normal), soit du fait d'un manque de puissance des outils descriptifs (ce qui doit être amélioré). On trouvera en annexe des captures d'écran montrant les jeux de propriétés utilisés et des exemples de reconnaissance à partir du corpus complet.

Pour le traitement des **doublons**, la hiérarchie des ressources dans les préférences est cruciale, et des choix restent à faire (hiérarchie entre les dictionnaires verbaux et nominaux pour réduire les doublons comme ապրիլ, 'avril' et 'vivre', գործէ 'travail-ABL.' et travailler-1SG-PST, ազատիս 'se-libérer-2SG-PST' et le moins probable 'libre-GD-POSS1). Il semble d'ores et déjà que les ambiguités entre formes adjectivales fléchies et formes verbales conjuguées homonymes penchent statistiquement en faveur du verbe, ce qui pourrait être un argument en faveur d'un étiquetage spécifique des adjectifs, ou d'un dictionnaire qui leur soit réservé. En revanche, pour les ambiguités entre substantif décliné et verbe conjugué, il n'est pas certain qu'un arbitrage puisse être trouvé sur une base statistique.

La hiérarchie la plus efficace s'est révélée la suivante :



La **non-reconnaissance** est imputable à deux causes essentielles :

- l'absence de dictionnaire nominal systématique comme nous en avons un pour le verbe,
- les fautes dans le corpus (absences de blancs, coquilles, etc.).

L'état actuel du travail nous a cependant permis de parvenir à un taux de reconnaissance qui rend possible un travail progressif de correction du corpus et de mise à jour des ressources à partir de la liste des inconnus. D'ores et déjà, elle a pu être utilisée (grâce à l'affichage par ordre alphabétique inverse, puis l'exportation) pour repérer les désinences verbales non reconnues, et donc pour mettre à jour le dictionnaire des verbes.

Par ailleurs, les dérivations productives sont un facteur prévisible de non-reconnaissance, et ne seront pas résolues par un dictionnaire. L'inventaire des procédés de dérivation productifs est en cours dans le cadre du mémoire de DEA de Goar Chobanyan à l'INALCO, il devrait permettre d'aboutir à des graphes morphologiques améliorant grandement la lemmatisation du corpus. Il restera ensuite à traiter également les emprunts, qu'ils soient en caractères latins ou arménien.

L'état auquel est parvenu ce projet en 18 mois semble d'ores et déjà remarquable compte tenu des énergies mobilisées (et ce, grâce à l'aide attentive de Max Silbersztein, qui a mis au point le module arménien de Nooj et nous a accompagnés tout au long de ce travail), bénéficiera grandement d'une synergie avec les autres projets mentionnés en introduction, en vue de développer et d'échanger des ressources.

Références

Boyacioglu, N., 2003, *Guide morphologique de la conjugaison des verbes de l'arménien occidental*, Mémoire de Maîtrise, INALCO, Paris, juillet 2003.

Boyacioglu, N., 2004, *Description linguistique du verbe en arménien occidental pour la lemmatisation*, mémoire de DEA sous la direction d'Anaïd Donabédian, INALCO, Paris, septembre 2004.

Deryle Lonsdale and Inna Danielyan, à paraître, A two-level implementation for Western Armenian morphology; à paraître dans *Annual of Armenian Linguistics* (15 pages)

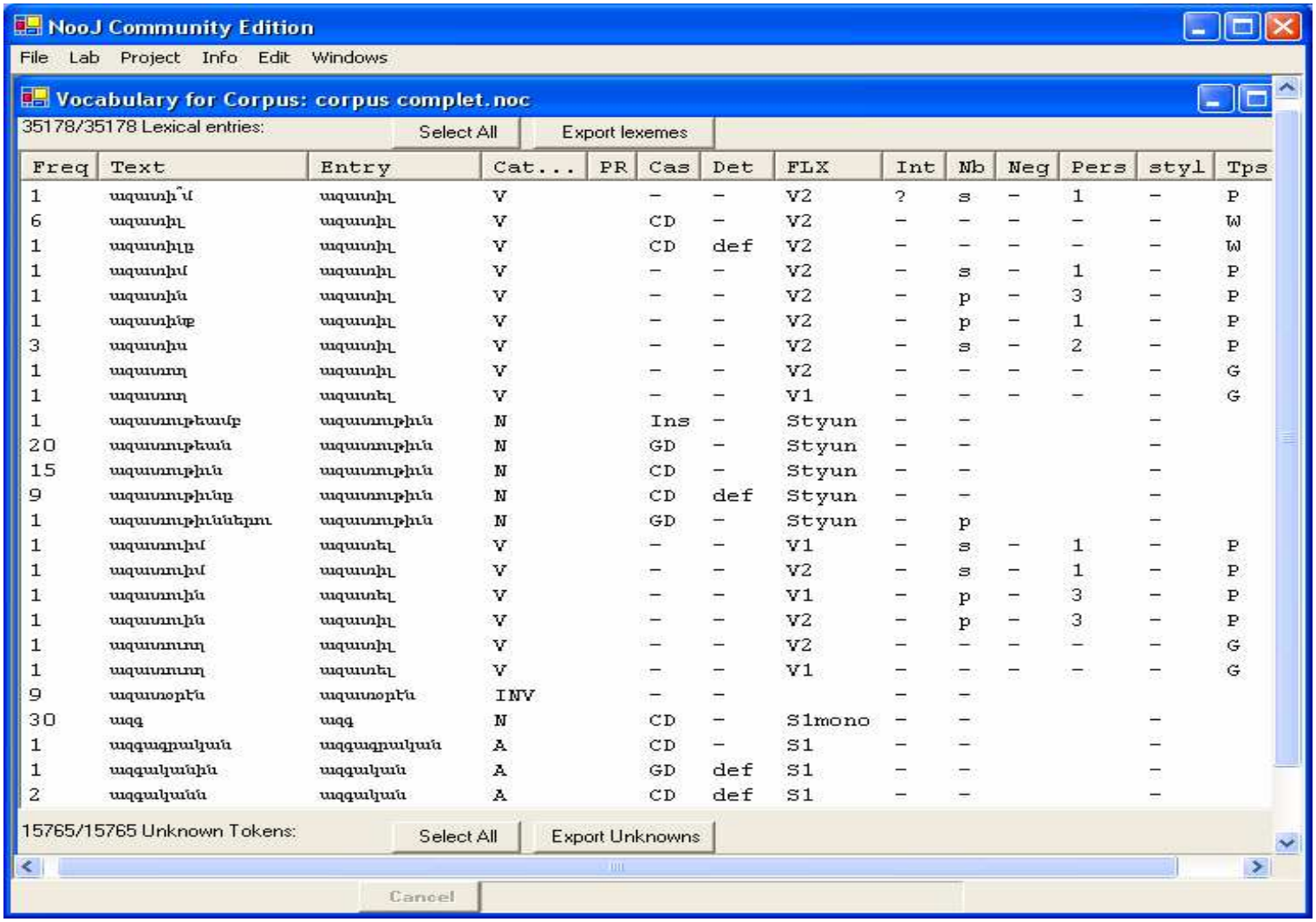
Donabédian, A. (ed.), 2001, *Langues de diaspora, langues en contact*, Faits de Langue, Ophrys, Paris, 18/2001, 282 p.

Donabédian, A., 1993, *L'article dans l'économie des catégories nominales en arménien occidental moderne*, thèse de doctorat sous la direction de Claude Hagège, Université Paris III, Sorbonne-Nouvelle, Février 1991, 350 pages.

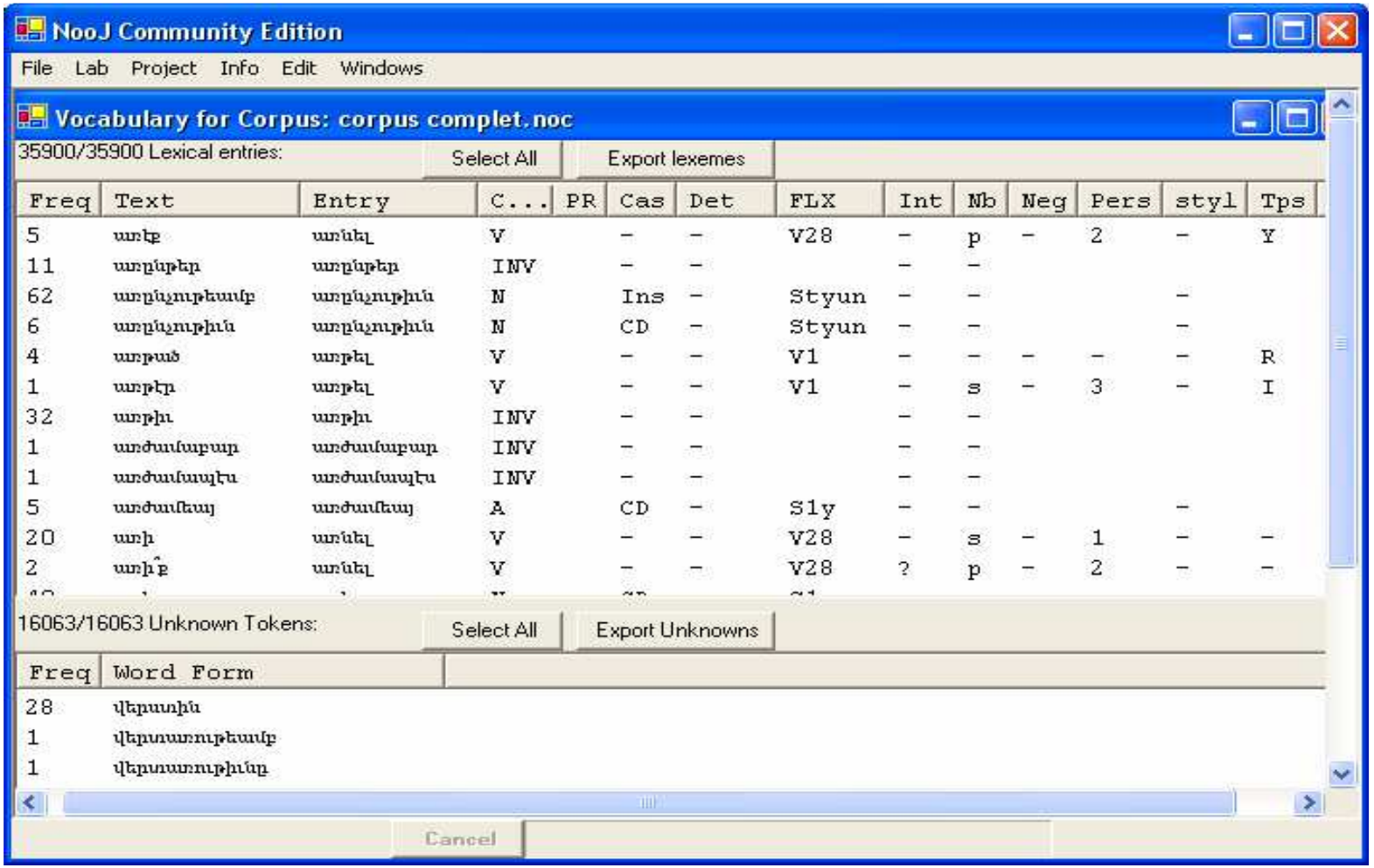
Donabédian, A., 2001b, Tabou linguistique en arménien occidental : 'gor' progressif est-il 'turc'? , in Donabédian, A. (ed.), *Langues de diaspora, langues en contact, Faits de Langue*, Ophrys, Paris, 18/2001, pp. 201-210.

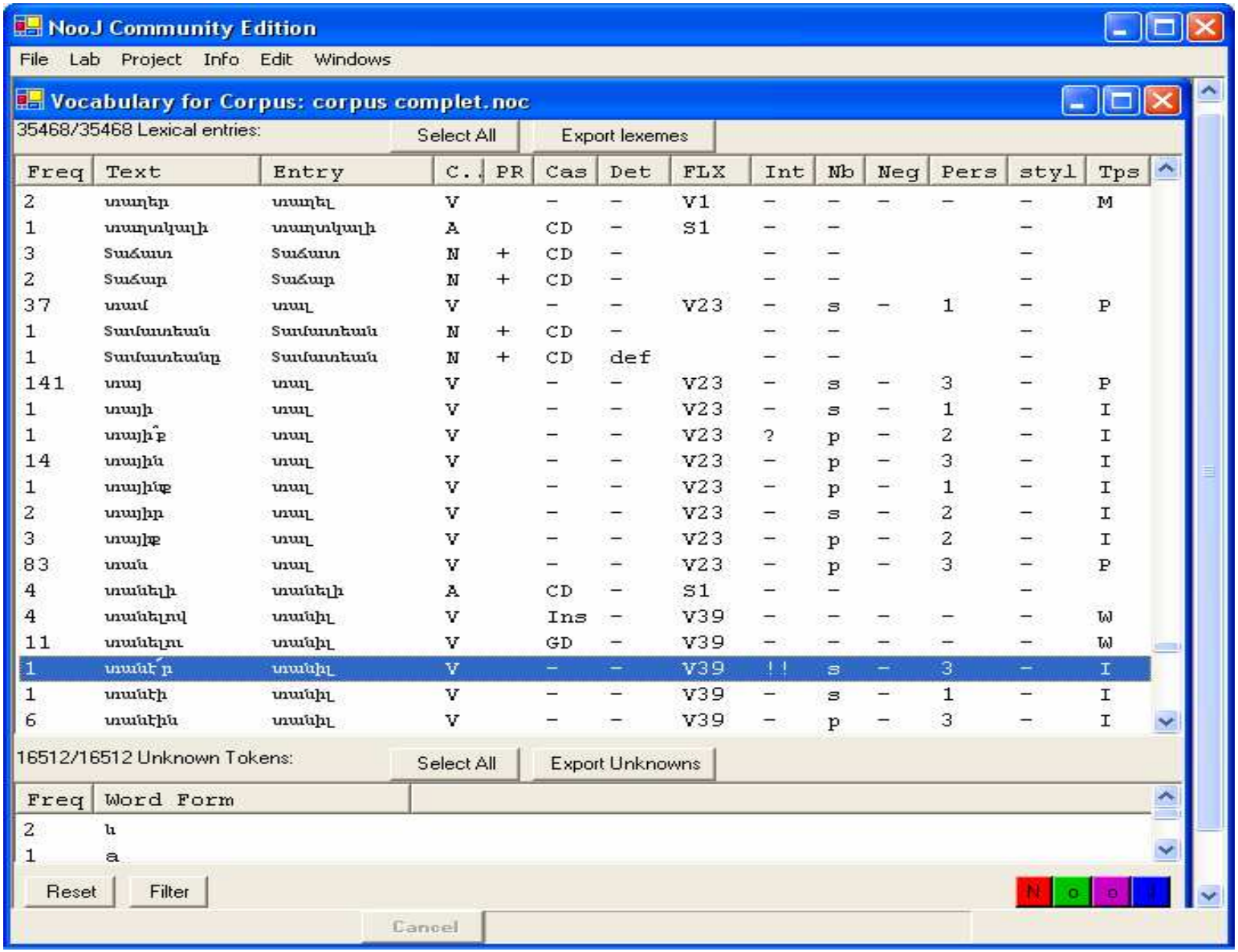
Donabédian, A., Description morphosyntaxique de l'arménien moderne occidental: utilisation d'une base de données textuelle, in *Computers in armenian philology*, Armenian Academy, Yerevan, 1993 , p. 6-14.

Annexes



Freq	Text	Entry	Cat...	PR	Cas	Det	FLX	Int	Nb	Neg	Pers	styl	Tps
1	ազատի՞մ	ազատիլ	V		–	–	V2	?	s	–	1	–	P
6	ազատիլ	ազատիլ	V		CD	–	V2	–	–	–	–	–	W
1	ազատիլը	ազատիլ	V		CD	def	V2	–	–	–	–	–	W
1	ազատիմ	ազատիլ	V		–	–	V2	–	s	–	1	–	P
1	ազատին	ազատիլ	V		–	–	V2	–	p	–	3	–	P
1	ազատինք	ազատիլ	V		–	–	V2	–	p	–	1	–	P
3	ազատիս	ազատիլ	V		–	–	V2	–	s	–	2	–	P
1	ազատող	ազատիլ	V		–	–	V2	–	–	–	–	–	G
1	ազատող	ազատել	V		–	–	V1	–	–	–	–	–	G
1	ազատութեամբ	ազատութիւն	N		Ins	–	Styun	–	–			–	
20	ազատութեան	ազատութիւն	N		GD	–	Styun	–	–			–	
15	ազատութիւն	ազատութիւն	N		CD	–	Styun	–	–			–	
9	ազատութիւնը	ազատութիւն	N		CD	def	Styun	–	–			–	
1	ազատութիւններու	ազատութիւն	N		GD	–	Styun	–	p			–	
1	ազատուիմ	ազատել	V		–	–	V1	–	s	–	1	–	P
1	ազատուիմ	ազատիլ	V		–	–	V2	–	s	–	1	–	P
1	ազատուին	ազատել	V		–	–	V1	–	p	–	3	–	P
1	ազատուին	ազատիլ	V		–	–	V2	–	p	–	3	–	P
1	ազատուող	ազատիլ	V		–	–	V2	–	–	–	–	–	G
1	ազատուող	ազատել	V		–	–	V1	–	–	–	–	–	G
9	ազատօրէն	ազատօրէն	INV		–	–		–	–				
30	ազգ	ազգ	N		CD	–	S1mono	–	–			–	
1	ազգագրական	ազգագրական	A		CD	–	S1	–	–			–	
1	ազգականին	ազգական	A		GD	def	S1	–	–			–	
2	ազգականն	ազգական	A		CD	def	S1	–	–			–	


NooJ Community Edition
File Lab Project Info Edit Windows
Vocabulary for Corpus: corpus complet.noc
35900/35900 Lexical entries:
Select All
Export lexemes
Freq Text Entry C... PR Cas Det FLX Int Nb Neg Pers styl Tps
5 առեք առնել V - - V28 - p - 2 - Y
11 առընթեր առընթեր INV - - - -
62 առընչություններ առընչություն N Ins - Styun - - - -
6 առընչություն առընչություն N CD - Styun - - - -
4 առած առել V - - V1 - - - - - R
1 առեր առել V - - V1 - s - 3 - I
32 առիւ առիւ INV - - - -
1 առժամապար առժամապար INV - - - -
1 առժամապէս առժամապէս INV - - - -
5 առժամեայ առժամեայ A CD - S1y - - - -
20 առի առնել V - - V28 - s - 1 - -
2 առի՞ք առնել V - - V28 ? p - 2 - -
16063/16063 Unknown Tokens:
Select All
Export Unknowns
Freq Word Form
28 վերսւին
1 վերսւոտութեամբ
1 վերսւոտութիւնը
Cancel


NooJ Community Edition
File Lab Project Info Edit Windows
Vocabulary for Corpus: corpus complet.noc
35468/35468 Lexical entries:
Select All
Export lexemes
Freq Text Entry C. PR Cas Det FLX Int Nb Neg Pers styl Tps
2 տաղեր տաղել V – – V1 – – – – – M
1 տաղտկալի տաղտկալի A CD – S1 – –
3 Տաճառ Տաճառ N + CD – – –
2 Տաճար Տաճար N + CD – – –
37 տամ տալ V – – V23 – s – 1 – P
1 Տահառեան Տահառեան N + CD – – –
1 Տահառեանը Տահառեան N + CD def – –
141 տայ տալ V – – V23 – s – 3 – P
1 տայի տալ V – – V23 – s – 1 – I
1 տայի՞ք տալ V – – V23 ? p – 2 – I
14 տային տալ V – – V23 – p – 3 – I
1 տայինք տալ V – – V23 – p – 1 – I
2 տայիր տալ V – – V23 – s – 2 – I
3 տայիք տալ V – – V23 – p – 2 – I
83 տան տալ V – – V23 – p – 3 – P
4 տանելի տանելի A CD – S1 – –
4 տանելով տանիլ V Ins – V39 – – – – – W
11 տանելու տանիլ V GD – V39 – – – – – W
1 տանո՞ւմ տանիլ V – – V39 ! ! s – 3 – I
1 տանուի տանիլ V – – V39 – s – 1 – I
6 տանուին տանիլ V – – V39 – p – 3 – I
16512/16512 Unknown Tokens:
Select All
Export Unknowns
Freq Word Form
2 և
1 a
Reset
Filter
Cancel