NEOPHILOLOGICA 2017, T. 29
ISSN 0208-5550 (version imprimée)
ISSN 2353-088X (version électronique)

Alicja Hajok
Université Pédagogique de Cracovie
Pologne

Modèle flexionnel des unités polylexicales nominales de la langue polonaise

A flexion model of the nominal units of the Polish language

Abstract

Traditional dictionaries and electronic dictionaries do not contain necessary information enabling automatic processing of idiomatic expressions. Many a time, an entry, even in a monolexical form, must be representative for the idiom (Mejri, 2008). Our aim is to create a dictionary of compound nouns customized for the natural language processing. In the following paper we are considering the ways of creating a dictionary of Polish compound nouns. We first describe morpho-syntactic properties of compound nouns, then we propose a morphological analyzer for Polish which enables declination of (i) indeclinable compound nouns (*jo-jo*, *lelum polelum*), (ii) declinable compound nouns (*ślepa uliczka*, *chude lata*), (iii) compound nouns containing a declinable and indeclinable elements (*herod-baba*, *języczek u wagi, kraina mlekiem i miodem płynąca*).

Keywords

Dictionary, compound nouns, natural language automatic processing

1. Introduction

Cet article constitue une suite à nos précédentes réflexions sur la constitution de ressources numériques en polonais (Hajok, 2015). Auparavant nous avons discuté les principes de constitution d'un moteur de flexion des unités simples et d'un dictionnaire des formes fléchies dans le but de les intégrer au système Unitex[1].

[1] Unitex est un analyseur du corpus qui permet d'appliquer des ressources lexicales sur les textes — http://unitexgramlab.org/ (consulté le 28 avril 2017).

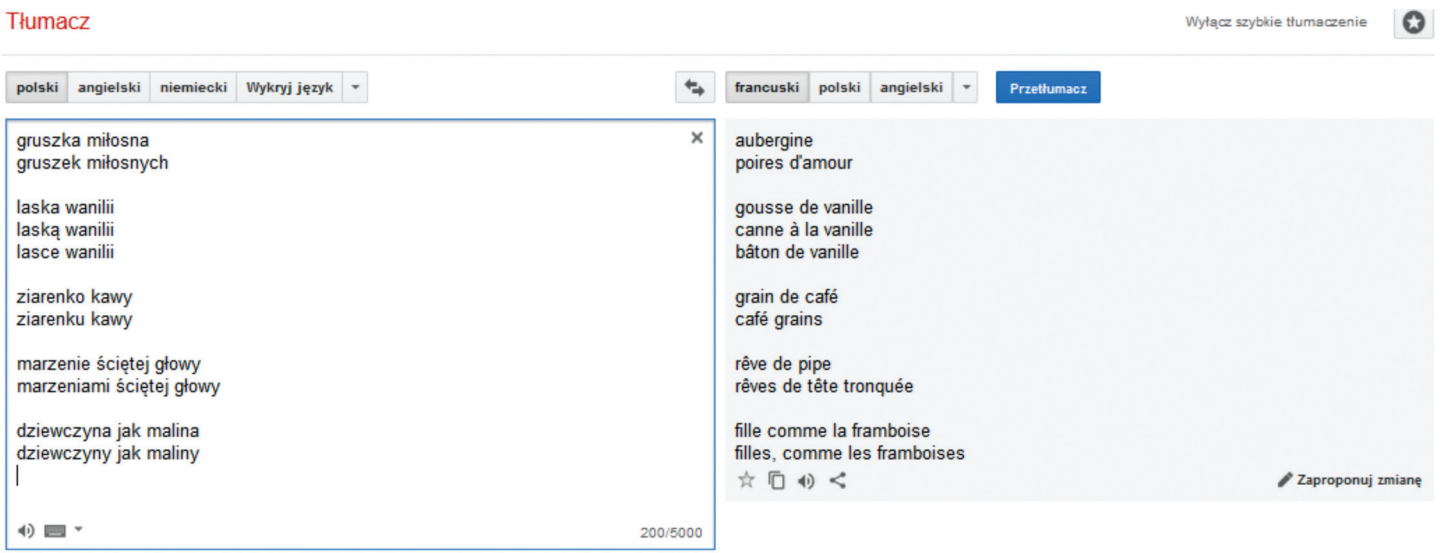

Figure 1. Translator Google, https://translate.google.pl/?hl=pl (consulté le 23 avril 2017)

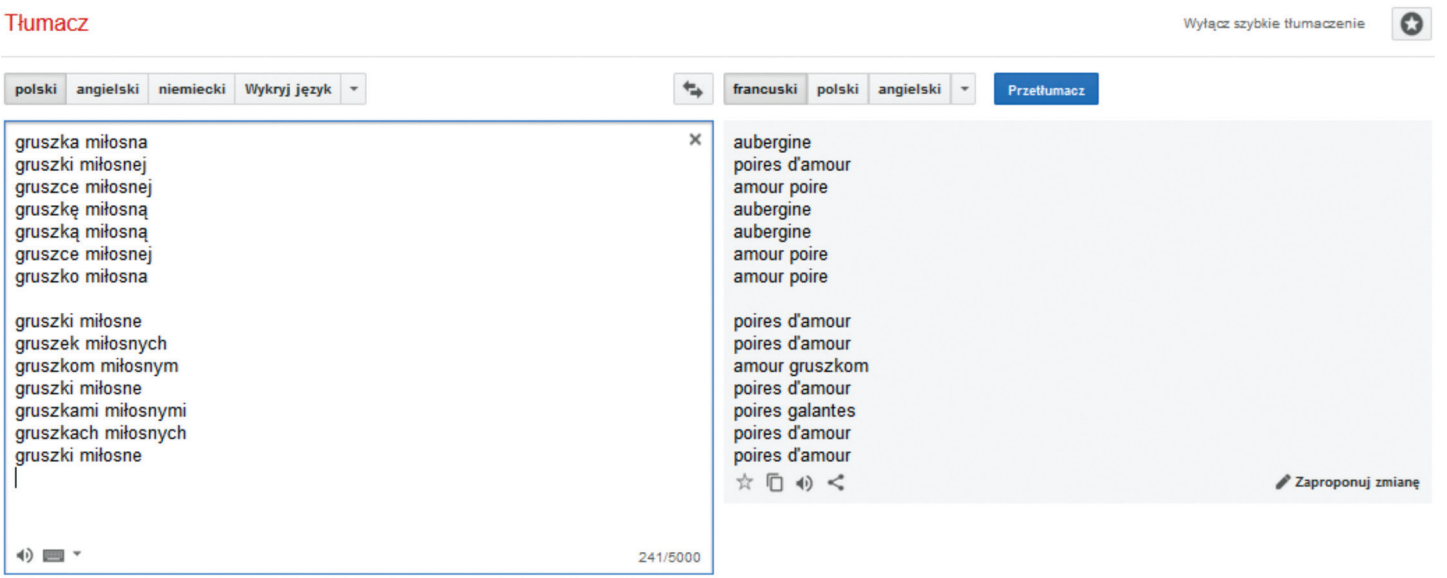

Figure 2. Translator Google, https://translate.google.pl/?hl=pl (consulté le 23 avril 2017)

Le présent travail poursuit ces réflexions, mais cette fois-ci dans le contexte des unités polylexicales.

L'objectif de notre travail est de proposer un dictionnaire applicable au traitement automatique des langues naturelles. Un tel dictionnaire doit prendre en compte les aspects morphologiques, syntaxiques et sémantiques des unités monolexicales et des unités polylexicales. Dans le cas des unités polylexicales, l'entrée du dictionnaire ne doit pas être présentée sous forme d'un mot vedette, mais elle doit être constituée d'une suite figée ou semi-figée.

La nécessité de telles études et de telles ressources se confirme lors de l'emploi des translators polonais-français. Pour illustrer ce problème, nous avons lancé une requête sur le Google translator (http://translate.google.com/ #), nous avons proposé de traduire du polonais vers le français quatre types de constructions nominales et leurs variantes flexionnelles (*cf.* figure 1) : (i) un nom composé : *gruszka miłosna* [trad. lit. 'poire amour'] (une aubergine), (ii) un substantif actualisé à l'aide du déterminant nominal : *laska wanilii* [trad. lit. 'bâton-vanille'] (une gousse de vanille), *ziarenko kawy* (grain de café), (iii) une construction à modifieur : *marzenie ściętej głowy* [trad. lit. 'rêve-tête-coupée'] (un rêve irréel), (iv) une construction comparative : *dziewczyna jak malina* [trad. lit. 'fille comme framboise'] (une très belle fille). Les traductions proposées sont dans la plupart des cas erronées ou aléatoires ce qui s'explique par la non-complémentarité des bases de données, ainsi que par l'insuffisance de fléchisseur automatique des unités complexes (*cf.* figure 2).

2. Les ressources linguistiques en polonais

Les travaux sur les ressources linguistiques de la langue polonaise se basent sur les recherches de Jan Tokarski (1973) qui a initié la description systématique de la morphologie polonaise. Sa conception a été ensuite reprise par Zygmunt Saloni (1988), Marek Świdziński (1992), Janusz Bień (1991, 2001), Krzysztof Szafran (1993 ; Bień, Szafran, 2001) et Zygmunt Vetulani (1998) et cela principalement dans leurs travaux sur le traitement informatique de la langue polonaise. Nous rappelons ci-dessous quelques traits caractéristiques des outils suivants : (i) les ressources présentées sous forme de liste des mots (*Słownik gramatyczny języka polskiego*), (ii) l'analyseurs de flexion (*Morfeusz*), (iii) l'environnement complet (*Corpus National de la langue polonaise*). Néanmoins ces outils ne contiennent pas de bases flexionnels des unités polylexicales.

(i) *Słownik gramatyczny języka polskiego*

Słownik gramatyczny języka polskiego (Saloni *et al*., 2007) contient environ 245 000 lemmes polonais dotés de leurs variantes flexionnelles. Les unités monolexicales constituent les entrées du dictionnaire. Le bloc erratique *Lelum polelum* [trad. lit. : 'un lambin'] est un seul nom composé retenu par ce dictionnaire, par contre il est divisé 'en deux' : *lelum* constitue une entrée et *polelum* une autre.

(ii) *Morfeusz SIaT*

L'analyseur *Morfeusz*[2] proposé par Marcin Woliński (2006) qui est une continuation des recherches de Tokarski (1973 : 158—169), Szafran (1993), Świdziński (1992), ne permet pas de désambiguïser la forme flexionnelle des unités polylexicales. Il propose une annotation morphologique pour chaque unité constitutive de la séquence figée.

Figure 3. *Morfeusz SGJP*, http://sgjp.pl/morfeusz/demo/?text=ziarenko+kawy (consulté le 27 avril 2017)

[2] http://sgjp.pl/morfeusz/demo (consulté le 28 avril 2017).

(iii) *Corpus National de la langue polonaise*

Le corpus d'IPI PAN[3] créé par l'Institut d'Informatique de l'Académie des Sciences[4] est le premier grand corpus de la langue polonaise. Les travaux sur le Corpus de la langue polonaise ont débuté en avril 2001 (Przepiórkowski, 2004 : 5) et leur objectif était de surmonter les insuffisances dans le domaine de la linguistique de corpus en polonais. Tous les segments du corpus sont annotés morpho-syntaxiquement. Le fonctionnement du corpus IPI PAN se résume en trois points[5] :

1. La segmentation — les étiquettes morphosyntaxiques sont ajoutées aux segments. Dans l'approche d'IPI PAN, un segment ne peut pas être plus long qu'un mot, compris comme une suite de caractères séparés par les deux blancs, mais il peut être plus court qu'un mot. Une telle segmentation ne permet pas de distinguer les unités monolexicales des unités polylexicales ; par exemple la locution verbale *wziąć byka za rogi* [prendre le taureau par les cornes] n'est pas analysée comme une séquence figée, mais comme une séquence libre — *wziąć* [wziąć:inf:perf] *byka* [byk:subst:sg:acc:m2] *za* [za:prep:acc] *rogi* [róg:subst:pl:acc:m3].
2. L'indexation morpho-syntaxique consiste à ajouter à chaque segment décliné — un lemme et une partie de discours puis des traits grammaticaux, par ex. : *okno* [okno:subst:sg:nom:n].
3. Le langage des requêtes est basé sur la syntaxe employée par le programme Corpus Query Processor (CQP), créé à l'Université de Stuttgart, Allemagne. La syntaxe des requêtes n'opère pas seulement sur les formes fléchies du mot, sur les lemmes, mais aussi sur les classes grammaticales et sur les catégories grammaticales attribuées à ces classes. Une requête [pos=adj & number=pl & case=»nom» & gender=n] nous permet de dégager tous les adjectifs neutres, pluriels au nominatifs.

3. La représentation et la structuration flexionnelle des unités polylexicales

Une des étapes de la création du dictionnaire des noms composés consiste à proposer un fléchisseur morphosyntaxique qui associe un moteur de flexion,

[3] http://korpus.pl/ (consulté le 23 avril 2017).

[4] Zespół Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki PAN.

[5] Il s'agit de la traduction des informations publiées sur le site d'internet de NKJP — Ściągawka do Narodowego Korpusu Języka Polskiego ; Adam Przepiórkowski, Aleksander Buczyński, Jakub Wilk, http://nkjp.pl/poliqarp/help/pl.html (consulté le 28 avril 2017).

un dictionnaire de formes fléchies et une interface de consultation. La description morphosyntaxique que nous avons utilisée repose sur le standard de description appliquée au Morfetik de la langue française (Mathieu-Colas, 2009 ; Buvet *et al.*, 2007 ; Issac, 2009 ; Hajok, 2015). Cependant, le système d'encodage préalablement défini se compose d'éléments analogues et d'éléments propres à chaque langue. Pour substantifs, nous avons repris les étiquettes retenues dans le Morfetik français. Cependant, pour rendre ces étiquettes opératoires pour le polonais, nous avons procédé aux modifications suivantes (Hajok, 2015 : 127) :

— nous avons ajouté les codes supplémentaires qui renvoient aux cas : n, g, d, a, i, l, v ;
— nous avons ajouté les codes supplémentaires qui renvoient aux genres : h, a, i, f, n[6].

Le tableau 1 illustre les différences dans l'étiquetage des substantifs français et dans l'étiquetage des substantifs polonais.

Tableau 1

Encodage des substantifs en français et en polonais

Attribut	Valeur en français	Exemple en français	Code en français	Valeur en polonais	Exemple en polonais	Code en polonais
Catégorie	substantif	*Garçon*	N	substantif	*Chłopiec*	N
Genre	masculin	*Garçon*	M	masculin personnel	*Chłopiec*	h
	féminin	*Fille*	F	masculin animal	*Pies*	A
				masculin inanimé	*Zeszyt*	I
				féminin	*dziewczynka*	F
				neutre	*dziecko*	N
Nombre	singulier	*garçon*	S	singulier	*chłopiec*	S
	pluriel	*garçons*	P	pluriel	*chłopcy*	P
Cas				nominatif	*chłopiec*	N
				génitif	*chłopca*	G
				datif	*chłopcu*	D
				accusatif	*chłopca*	A
				instrumental	*chłopcem*	I
				locatif	*chłopcu*	L
				vocatif	*chłopcze*	V

[6] « La distinction de trois types de masculin n'entre pas dans le cadre de l'étiquetage morphologique, mais dans le cadre de l'étiquetage syntaxique. Cependant, pour faciliter la description des relations entre les éléments de la phrase, il nous semble indispensable de noter régulièrement cette information qui est pertinente non seulement pour les substantifs, mais aussi pour les adjectifs, les pronoms et les déterminants. Pour rendre l'encodage polonais compatible avec l'encodage appliqué aux autres langues, nous avons remplacé les abréviations traditionnelles m1/ m2/ m3 respectivement par h/ a/ i » (Hajok, 2015 : 127).

Le modèle *Proteus*[7] « s'appuie sur un ensemble d'opérateurs sur les caractères et d'un objet que nous appelons 'pile' capable de recevoir des caractères que l'on 'met de côté'. L'utilisation dans un certain ordre de ces opérateurs constitue un 'code' correspondant à une fonction de transformation : mot + code = mot fléchi » (Issac, 2009 : 14). Prenons un exemple de génération du génitif singulier du substantif *Krasnystaw* [trad. : 'nom d'une ville en Pologne'] qui s'effectue en plusieurs étapes (*cf.* tableau 2). À la forme canonique du substantif, nous appliquons un code 4P\y\/ego/4D/u/. Les étapes 1—6 présentent progressivement les opérations effectuées par *Proteus* : il faut mettre du côté les 4 caractères (4P), ensuite il faut remplacer un caractère *y* par un caractère *ego (\y\/ego/)*, puis il est nécessaire de déplacer les 4 caractères mis de côté (4D) et finalement ajouter un caractère (*u*).

Tableau 2

Modèle *Proteus*

N°	Mot	Pile	Code
1	Krasn**y**staw[1]		4P**y**\/**ego**/4D/**u**/
2	Krasn**y**	staw	**y**\/**ego**/4D/**u**/
3	Krasn	staw	/**ego**/4D/**u**/
4	Krasn**ego**	staw	4D/**u**/
5	Krasn**ego**		/**u**/
6	Krasn**ego**staw**u**		

Les résultats obtenus sont au format XML et sont présentés sous forme de quatre tableaux : forme fléchie / lemme / étiquette / code de déclinaison :

rzęsa	rzęsa	Nfsn	S_001
rzęsy	rzęsa	Nfsg/fpn/fpa/fpv	S_001
rzęsę	rzęsa	Nfsa	S_001
etc.			

En proposant la déclinaison des suites figées, nous avons profité de la déclinaison des unités simples. À l'aide des règles de concaténation[8], nous avons reconstitué la déclinaison des suites figées. Pour ce faire, nous avons constitué un patron flexionnel qui montre les dépendances flexionnelles entre les éléments du nom composé. Un nom composé peut contenir des éléments variables *'T'*, *'C'* et des éléments invariables *'I'*. Les variations flexionnelles dépendent du degré de figement des noms composés, autrement dit la flexion doit tenir compte de l'éventuelle autonomie des éléments constitutifs. Cependant, un grand nombre de noms composés se comportent comme des groupes nominaux libres. Cela s'explique par le fait que « le figement est un processus qui doit son existence au temps et au fonctionnement normal de la langue » (Mejri, 1997 : 135). Alors les noms

[7] *Proteus* est un outil constitué par Fabrice Issac, Université Paris 13 (Issac, 2009).

[8] L'idée de règle de concaténation a été retenue lors de longues discussions avec Fabrice Issac.

composés du type AN et du type NA ne posent aucun problème flexionnel car ils reposent sur le même principe flexionnel que les groupes nominaux libres. Autrement dit la combinatoire interne qui régit les formes flexionnelles est la même dans le cas des séquences libres et des séquences figées. Mais, il serait trop rapide de classifier ces constructions comme flexionnellement régulières. Il est indispensable de tenir compte non seulement des variations casuelles, mais aussi des variations du nombre (*chudy rok = chude lata* [trad. lit. : 'maigre' : adj, sg, nom, m1 ; 'année' : subst, sg, nom, m1/ 'l'année maigre' // trad. lit. : 'maigre' : adj, pl, nom, m1 ; 'année' : subst, pl, nom, m1/ 'les années maigres']) et du genre (*stary malutki = stara malutka* [trad. lit. : 'vieux' : adj, sg, nom, m ; 'petit' : adj, sg, nom, m // trad. lit. : 'vieux' : adj, pl, nom, f ; 'petit' : subst, pl, nom, f/ 'un garçon ou une fille précoce <trop mûre> pour son âge']).

La typologie des noms composés permet de dégager plusieurs types morphologiques, ce qui nécessite la création de différents patrons flexionnels. Alors, un nom composé se compose d'un **élément-tête** *'T'* qui impose à l'**élément-complément** *'C'* : le cas, le genre et le nombre [gnc] et des **éléments invariables** *'I'*. Le tableau 3 présente quelques types de patrons flexionnels.

Tableau 3

Patrons flexionnels

Exemple	Moule locutionnel	Élément variable	Patron flexionnel	Traduction
mocna ***karta***	AN	**A[v]N[v]**	C[gnc]T[gnc]	une carte forte
błędne ***koło***	AN	**A[v]N[v]**	C[gnc]T[gnc]	un cercle vicieux
ostatnia deska *ratunku*	ANN	**A[v]N[v]**N	C[gnc]T[gnc]I	une planche de salut
wikt *i* ***opierunek***	NconjN	**N[v]**conj**N[v]**	(T[nc]IT[nc])s	logé et blanchi
obiecanka cacanka	NN	**N[v]N[v]**	T[nc]T[nc]	une belle promesse
anioł stróż	NN	**N[v]N[v]**	T[nc]T[nc]	un ange gardien
woda *ognista*	NA	**N[v]A[v]**	(T[gnc]C[gnc])s	l'eau de vie
karta *przetargowa*	NA	**N[v]A[v]**	T[gnc]C[gnc]	une carte forte
kraina *mlekiem i miodem* ***płynąca***	NnconjNP	**N[v]**NconjN**P[v]**	T[gnc]IIIC[gnc]	un pays de cocagne
burza *w szklance wody*	NprépNN	**N[v]**prépNN	TIII	une tempête dans un verre d'eau
kolos *na glinianych nogach*	NprépAN	**N[v]**prépAN	TIII	un colosse aux pieds d'argile
pan *życia i śmierci*	NNconjN	**N[v]**NconjN	TIII	le maître de la vie et de la mort
A — adjectif, c — cas, C — élément complément, conj — conjonction, g — genre, I — élément invariable, N — nom, n — nombre, P — participe, prép — préposition, T — élément-tête, v — élément variable				

Pour décliner *ostatnia deska ratunku* [trad. lit. : ‘dernier’ : adj, sg, nom, f ; ‘planche’ : subst, sg, nom, f ; ‘secours’ : subst, sg, gén, m3 / ‘une planche de salut’], dont le patron flexionnel est **C[gnc]T[gnc]I**, nous avons décliné le substantif-tête *deska* [‘planche’ : subst, sg, nom, f] qui a imposé sa valeur flexionnelle au complément *ostatni* [‘dernier’ : adj, sg, nom, m] et finalement nous avons ajouté un élément invariable *ratunku* [‘secours’ : subst, sg, gén, m]. Ce mécanisme se présente comme suit :

1) la déclinaison d’élément-tête *deska* T[gnc]

deska	deska	Nfsn S_025
deski	deska	Nfsg/fpn/fpa/fpv S_025
desce	deska	Nfsd/fsl S_025
deskę	deska	Nfsa S_025
deską	deska	Nfsi S_025
desko	deska	Nfsv S_025
desek	deska	Nfpg S_025
deskom	deska	Nfpd S_025
deskami	deska	Nfpi S_025
deskach	deska	Nfpl S_025

2) la déclinaison de l’élément-complément *ostatni* C[gnc]

ostatni	ostatni	Aphsn/pasn/pisn/pisa/phsv/pasv/pisv/phpn/papn/pipn/phpv/papv/pipv A_002
ostatnie	ostatni	Apnsn/pnsa/pnsv/pfpn/pnpn/pnpv/pnpv/pfpa/pnpa A_002
ostatnia	ostatni	Apfsn/pfsa/pfsv A_002
ostatniego	ostatni	Aphsg/pasg/pisg/pnsg/phsa/pasa A_002
ostatniej	ostatni	Apfsg/pfsd/pfsl A_002
ostatniemu	ostatni	Aphsd/pasd/pisd/pnsd A_002
ostatnią	ostatni	Apfsa/pfsi A_002
ostatnim	ostatni	Aphsi/pasi/pisi/pnsi/phsl/pasl/pisl/pnsl/phpd/papd/pipd/pfpd/pnpd A_002
ostatnich	ostatni	Aphpg/papg/pipg/pfpg/pnpg/phpa/papa/pipa/phpl/papl/pipl/pfpl/pnpl A_002
ostatnimi	ostatni	Aphpi/papi/pipi/pfpi/pnpi A_002
ostatnio-	ostatni	Acomplexe A_002

3) l’accord de l’élément-complément **C** avec l’élément-tête **T**

À cette étape, il s’agit de croiser les étiquettes flexionnelles de l’élément-tête **T** avec celles de l’élément-complément **C** et de garder seulement les étiquettes

dont les informations morphologiques sont communes : genre, nombre et cas. Ainsi nous retenons les constructions suivantes :

(i)

deski	N**fsg**/**fpn**/**fpa**
ostatniej	Ap**fsg**
ostatnie	Ap**fpn**/p**fpa**

(ii)

ostatniej	Ap**fsg**	deski	N**fsg**
ostatnie	Ap**fpn**/p**fpa**	deski	N**fpn**/**fpa**

4) l'ajout d'élément invariable **I** *ratunku* :

ostatniej	deski	**ratunku** Nfsg
ostatnie	deski	**ratunku** Nfpn/fpa

4. La classification des noms composés selon leur flexion interne

Selon les changements casuels, nous avons retenu trois types des noms composés en polonais[9] :

(i) les noms composés invariables (*jo-jo* [un yo-yo], *lelum polelum* [un lambin]),

(ii) les noms composés où la flexion affecte tous les éléments (*ślepa uliczka* [trad. lit. : 'aveugle' : adj, sg, nom, f ; 'ruelle' : subst, sg, nom, f / 'une impasse'], *chude lata* [trad. lit. : 'maigre' : adj, pl, nom, m1 ; 'année' : subst, pl, nom, m1/ 'les années maigres']),

(iii) les noms composés où un ou plusieurs éléments varient (*herod-baba* [trad. lit. : 'Herode' : subst, sg, nom, m1 ; 'femme' : subst, sg, nom, f / 'une virago'], *kraina mlekiem i miodem płynąca* [trad. lit. : 'pays' : subst, sg, nom, f ; 'lait' : subst, sg, instr, n ; 'miel' : subst, sg, intstr, m3 ; 'nager' : participe, sg, nom, f / 'un pays de cocagne']).

(i) Les noms composés invariables

Parmi les noms composés invariables, nous comptons les substantifs d'origine étrangère, par ex. les noms propres (*Los Angeles*, *Rio de Janeiro*, *Ułan Bator,*

[9] Pour plus de précisions voir en outres : Bańko (2007) ; Jadacka (2007) ; Nagórko (2006).

Alma Mater) et les noms communs (*absolutum dominium*). Dans ce cas, le simple listage suffit. Le modèle *Proteus* attribue automatiquement à ces substantifs des étiquettes flexionnelles adéquates.

Alma Mater	Alma Mater	Nfsn/fsg/fsd/fsa/fsi/fsl/fsv/	S_000
absolutum dominium	absolutum dominium	Nnsn/nsg/nsd/nsa/nsi/nsl/nsv/	S_000
Los Angeles	Los Angeles	Nisn/isg/isd/isa/isi/isl/isv/	S_000
Rio de Janeiro	Rio de Janeiro	Nnsn/nsg/nsd/nsa/nsi/nsl/nsv/	S_000
Ułan Bator	Ułan Bator	Nisn/isg/isd/isa/isi/isl/isv/	S_000

À noter qu'il existe des noms composés d'origine étrangère où l'un des éléments accepte les modifications casuelles (*Addis Abeba* / *Addis Abeb**ie*** ou *prima sort* / *prima sor**cie*** ou *prima aprilis* / *prima aprilis**ie***) et même la marque du pluriel (*prima aprilis* / *prima aprilis**y***).

À noter que certains emprunts, à force d'être employés ont été polonisés, ils admettent aussi les variantes flexionnelles, par ex. *baby-sitter* dont la forme masculine accepte la déclinaison par le cas et le nombre, comparons :

baby-sitter	baby-sitter	Nfsn/fsg/fsd/fsa/fsi/fsl/fsv/ fpn/fpg/fpd/fpa/fpi/fpl/fpv/	S_000
baby-sitter	baby-sitter	Nhsn	
baby-sittera	baby-sitter	Nhsg/hsa	
baby-sitterowi	baby-sitter	Nhsd	
baby-sitterem	baby-sitter	Nhsi	
baby-sitterze	baby-sitter	Nhsl/hiv	
baby-sitterzy	baby-sitter	Nhpn/hpv	
baby-sittery	baby-sitter	Nhpn/hpv	
baby-sitterów	baby-sitter	Nhpg/hpa	
baby-sitterom	baby-sitter	Nhpd	
baby-sitterami	baby-sitter	Nhpi	
baby-sitterach	baby-sitter	Nhpl	

(ii) Les noms composés variables

Le plus souvent, il s'agit de constructions du type NA ou AN où N est un substantif-tête qui impose le cas, le nombre et le genre à l'adjectif. Les substantifs *woda ognista* ['l'eau de vie'], *karta przetargowa* ['une carte forte'] et *błędne koło* ['un cercle vicieux'] se déclinent selon les règles appropriées aux unités monolexicales. Contrairement aux unités simples, l'étiquetage ne consiste pas à attribuer une étiquette à chaque élément constituant le nom composé (≠*woda*$_{[Nfsn]}$ *ognista*$_{[Apfsn]}$) mais à approprier une seule étiquette à toute la construction (= *woda ognista*$_{[Nfsn]}$).

Dans les construction du type $N_{NOM}N_{NOM}$ (*anioł stróż* [‘un ange gardien’]), il est difficile de dégager le substantif-tête, car la déclinaison de deux éléments est indépendante dans la mesure d’appropriation des règles flexionnelles, mais du point de vue de la combinatoire interne et de la combinatoire externe, les deux éléments doivent accepter le même cas et le même nombre :

(*anioł*$_{[Nasn]}$ *stróż*$_{[Nasn]}$)$_{[Nasn]}$ / (*anioła*$_{[Nasg]}$ *stróża*$_{[Nasg]}$)$_{[Nasg]}$.

(iii) Les noms polylexicaux composés d’éléments variables et d’éléments invariables

Généralement, il s’agit des constructions $N_{NOM}N_{GEN}$ (*człowiek honoru* [‘l’homme d’honneur’]) et des constructions composées de plusieurs éléments, par exemple NprépAN (*kolos na glinianych nogach* [‘un colosse aux pieds d’argile’]) ou ANN_{GEN} *ostatnia deska ratunku* [‘une planche de salut’]). Les patrons flexionnels sont les suivants : TI (T- *człowiek*, I- *honoru*), TIII (T- *kolos,* I- *na,* I- *glinianych,* I- *nogach*), C[gnc]T[gnc]I (C- *ostatnia,* T- *deska,* I- *ratunku*).

Dans les constructions $N_{NOM}N_{GEN}$, nous notons une relation de dépendance entre les deux éléments $N_{NOM}N_{GEN}$. Le substantif-tête impose le cas génitif au deuxième substantif — complément qui restera ainsi invariable (*pranie*$_{[Nisn]}$ *mózgu*$_{[Nisg]}$)$_{[Nisn]}$ / (*prania*$_{[Nisg]}$ *mózgu*$_{[Nisg]}$)$_{[Nisn]}$ [‘un lavage de cerveau’].

Nous relevons également les composés avec l’interfixe *-o* dont l’emploi est très fréquent. Il s’agit avant tout de la formation des adjectivaux, *biał**o**-czarny* [‘blanc et noir’], *słodk**o**-kwaśny* [trad. lit. : ‘doux et aigre’ / ‘aigre-doux’]. Quant aux substantifs, ils connaissent le plus souvent la soudure des éléments : *beczk**o**-wóz* [trad. lit. : ‘tonneau et véhicule’ / ‘un tonneau’], *dramat**o**pisarz* [trad. lit. : ‘drame et écrivain’ / ‘un auteur dramatique’], *deszcz**o**chron* [trad. lit. : ‘pluie et protection’ / ‘un parapluie’], *piorun**o**chron* [trad. lit. : ‘foudre et protection’ / ‘un parafoudre’]. Mais, il existe aussi des substantifs écrits avec le trait d’union : *chłodziark**o**-zamrażarka* [trad. lit. : ‘réfrigérateur et congélateur’ / ‘un réfrigérateur’], *Austr**o**-Węgry* [‘Autriche et Hongrie’]. Le premier élément est toujours invariable, la déclinaison concerne seulement le deuxième élément et elle est basée sur le principe de la déclinaison des unités monolexicales. Alors, *beczko**wóz*** [trad. lit. : ‘tonneau et véhicule’ / ‘un tonneau’] se décline comme ***wóz*** [‘un véhicule’] et *chłodziarko-**zamrażarka*** [trad. lit. : ‘réfrigérateur et congélateur’ / ‘un réfrigérateur’] se décline comme ***zamrażarka*** [‘un congélateur’].

(iv) Les difficultés flexionnelles

La description morphologique se complique davantage, quand il s'agit de la déclinaison : (a) des éléments qui n'ont pas d'existence autonome, (b) des singulare tantum ou de pluralia tantum, (c) des constructions qui demandent l'attribution de deux règles flexionnelles.

Certains éléments du nom composé (*obiecanki cacanki* ['une belle promesse'], *duby smalone* ['des bêtises'], *lukullusowa uczta* ['un festin de Lucullus'], *insza inszość* ['c'est une autre paire de manche']) n'ont pas d'existence autonome, donc ils ne sont pas retenus dans les listes des lemmes simples, ainsi l'appropriation d'une règle flexionnelle n'est pas possible. Dans certains cas, il s'agit de la déclinaison manuelle (1), mais l'objectif de *Proteus* est de permettre aussi la déclinaison de toutes les unités linguistiques. Ainsi, il est possible d'associer une règle flexionnelle à une unité inconnue en analysant sa terminaison et en la comparant avec celles retenues dans la base des lemmes (2).

(1)

insza inszość	Nfsn
inszej inszości	Nfsg/fsd/fsl
inszą inszością	Nfsa/fsi
insze inszości	Nfpn/fpa
inszych inszości	Nfpg/fpl
inszym inszościom	Nfpd
inszymi inszościami	Nfpi

(2) *lukullusowa uczta — lukullusow**y***, c'est un adjectif qui n'a pas d'existence autonome, mais il se décline comme *luksusow**y*** — règle A_014.

Les difficultés flexionnelles consistent dans le fait que certaines constructions acceptent soit le singulier (*wdowi grosz* ['de la veuve' : adj, sg, nom, m ; 'denier' : subst, sg, nom, m3 / 'un denier de la veuve']) soit le pluriel (*kocie łby* ['de chat' : adj, pl, nom, m ; 'tête' : subst, pl, nom, m3 / 'un pavé']). Pour ne pas générer des formes incorrectes, il est nécessaire de marquer cette exception dans le patron flexionnel correspondant (*cf.* tableau 4).

Certaines constructions demandent l'association de deux règles flexionnelles. Le redoublement flexionnel trouve son explication dans la formation du pluriel. Par exemple, le substantif *rok* se fléchit régulièrement au singulier. Par contre la forme plurielle est orthographiquement différente : *lata*. Pour le besoin de *Proteus*, nous avons dissocié ces deux formes et pour chacune, nous avons proposé une règle appropriée. D'où la nécessité de redoublement des entrées dans la base des lemmes composés. Dans les deux cas de figure, nous notons obligatoirement l'information sur le nombre (s/p) (cf. tableau 5).

Tableau 4

Singulare tantum ou Pluralia tantum

Exemple	Moule locutionnel	Élément variable	Patron flexionnel	Traduction
wdowi ***grosz***	AN	**A[v]N[v]**	(C[g**n**c]T[g**n**c])s	un denier de la veuve
kocie ***łby***	AN	**A[v]N[v]**	(C[g**n**c]T[g**n**c])p	un pavé
duby *smalone*	NA	**N[v]A[v]**	(T[gnc]C[gnc])p	des bêtises
pogoda *w kratkę*	NprépN	**N[v]prépN**	(TII)s	le temps variable (incertain)
oczy *bazyliszka*	NN	**N[v]N**	(TI)p	les yeux de basilic
(…)s construction employée seulement au singulier **(…)p** construction employée seulement au pluriel				

Tableau 5

Redoublement des entrées

Exemple	Moule locutionnel	Élément variable	Patron flexionnel	Traduction
chudy ***rok***	AN	**A[v]N[v]**	(C[g**n**c]T[g**n**c])s	une année maigre
chude ***lata***	AN	**A[v]N[v]**	(C[g**n**c]T[g**n**c])p	des années maigres
człowiek *honoru*	NA	**N[v]A[v]**	(T[g**n**c]C[g**n**c])s	un homme d'honneur
ludzie *honoru*	NA	**N[v]A[v]**	(T[g**n**c]C[g**n**c])p	des hommes d'honneur
(…)s construction employée seulement au singulier **(…)p** construction employée seulement au pluriel				

5. Conclusion

Le travail que nous avons entrepris ne constitue qu'une première réflexion sur la création d'un dictionnaire des noms composés en polonais. Cependant, pour constituer une ressource complète et applicable aux besoins du TAL, il est nécessaire de suivre les indications de Salah Mejri (2008) qui sont les suivantes :

a) isoler toutes les formes complètement figées,
b) constituer un lemmatiseur des noms composés, permettant de générer toutes les variantes du nom composé en question,
c) récupérer toutes les formes transformationnelles des noms composés,
d) décrire la combinatoire interne des noms composés,
e) décrire la combinatoire externe des noms composés,
f) intégrer les informations sémantiques.

Références

Bańko Mirosław, 2007: *Wykłady z polskiej fleksji*. Warszawa: PWN.

Bień Janusz, 1991: *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*. Rozprawy Uniwersytetu Warszawskiego / Dissertationes Universitatis Varsoviensis. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa. http://bc.klf.uw.edu.pl/12/ (consulté le 28 avril 2017).

Bień Janusz, 2001: „O pojęciu wyrazu morfologicznego". In: Włodzimierz Gruszczyński, Urszula Andrejewicz, Mirosław Bańko, Dorota Kopcińska, red.: *Nie bez znaczenia…* Białystok: Wydawnictwo Uniwersytetu w Białymstoku, 67—77.

Bień Janusz, Szafran Krzysztof, 2001: „Analiza morfologiczna języka polskiego w praktyce". *Bulletin de la société polonaise de linguistiques LVII* : 1—17. URL: http://bc.klf.uw.edu.pl/88/1/JSB-KS-PTJ01.pdf (consulté le 28 avril 2017).

Buvet Pierre-André, Cartier Emmanuel, Issac Fabrice, Mejri Salah, 2007 : « Dictionnaires électroniques et étiquetage syntactico-sémantique ». In : Nabil Hathout, Philippe Muller, éds : *Actes des 14e journées sur le Traitement Automatique des Langues Naturelles*. Toulouse : IRIT Press., 239—248.

Hajok Alicja, 2015 : « La constitution de ressources numériques en polonais — les unités simples ». *Neophilologica*, **27**, 123—134.

Issac Fabrice, 2009 : « Place des ressources lexicales dans l'étiquetage morpho-syntaxique ». *L'Information grammaticale*, **122** : 10—18.

Jadacka Hanna, 2007: *Kultura języka polskiego.* Warszawa: PWN.

Lipińska Halina, Saloni Zygmunt, 1987: „O grupach koniugacyjnych w języku polskim". In: *Studia gramatyczne.* T. 8. Wrocław: Ossolineum, 71—88.

Mathieu-Colas Michel, 2009 : « *Morfetik* : une ressource lexicale pour le TAL ». *Cahiers de lexicologie*, **94**, 1, 137—146.

Mejri Salah, 1997 : *Le figement lexical Description linguistique et structuration sémantique*. Manouba.

Mejri Salah, 2008 : « Vers un dictionnaire électronique des séquences figées ». In : Giovanni Dotoli, Giulia Papoff, éds : *Du sens des mots. Le réseau sémantique du dictionnaire : actes des Journées italiennes des dictionnaires : deuxièmes journées, Benevento 28—29 janvier 2008.* Fasano : Schena Editore (Biblioteca della Ricerca), 117—129.

Mejri Salah, Neveu Franck, éds, 2009 : *L'Information grammaticale*, nº 122 : *Catégories linguistiques et étiquetage de corpus*.

Nagórko Alicja, 2006: *Zarys gramatyki polskiej.* Warszawa: PWN.

Przepiórkowski Adam, 2004: *Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version*. Warszawa: IPI PAN.

Saloni Zygmunt, red., 1988: *Studia z polskiej leksykografii współczesnej.* T. 1. Warszawa: Ossolineum.

Saloni Zygmunt, Gruszczyński Włodzimierz, Woliński Marcin, Wołosz Robert, 2007: *SGJP*: *Słownik gramatyczny języka polskiego*. Warszawa: Wiedza Powszechna, version informatisée sur CD-Rom, wersja 1.0.

Saloni Zygmunt, Woliński Marcin, 2003: "A Computerized Description of Polish Conjugation". In: Peter Kosta *et al.*, ed.: *Investigations into Formal Slavic Linguistics*. Frankfurt-am-Main; part I, 373—384.

Savary Agata, Rabiega-Wiśniewska Joanna, Woliński Marcin, 2009: "Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex". *Aspects of Natural Language Processing*, LNCS 5070, 111—141.

Świdziński Marek, 1992: *Gramatyka formalna języka polskiego*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Szafran Krzysztof, 1993: *Automatyczna analiza fleksyjna tekstu polskiego (na podstawie „Schematycznego indeksu a tergo" Jana Tokarskiego).* Rozprawa doktorska, Wydział Polonistyki UW, Warszawa.

Tokarski Jan, 1973: *Fleksja polska*. Warszawa.

Tokarski Jan, 2001: *Schematyczny indeks a tergo polskich form wyrazowych*. Opracowanie i redakcja Zygmunt Saloni. Wydanie drugie. Warszawa: Wydawnictwo Naukowe PWN.

Vetulani Zygmunt, 1998: *Dictionary based methods and tools for language engineering*. Poznań: Uniwersytet im. Adama Mickiewicza.

Woliński Marcin, 2006: "*Morfeusz* — a Practical Tool for the Morphological Analysis of Polish". In: Mieczysław Kłopotek, Sławomir Wierzchoń, Krzysztof Trojanowski, eds.: *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*, 503—512.