Alicja Hajok
Université Pédagogique de Cracovie
Pologne

La constitution de ressources numériques en polonais — les unités simples

Abstract

This article discusses several problems concerning the automatization of declination of simple units in Polish language. Our analysis is based on the Proteus model. The proposed solution contains a module that enables automatic generation of possible inflected variants of the given lexical units.

Keywords

Declination, dictionary, natural language automatic processing

1. Introduction

Le traitement automatique des langues naturelles constitue un domaine à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle. Cette discipline, relativement récente, connaît un développement plus ou moins rapide qui reste toujours en relation avec les besoins des utilisateurs d'une langue donnée. Ainsi le traitement automatique des langues écrites cherche à répondre aux besoins de la traduction automatique, des résumés automatiques ou de l'apprentissage, etc. Ces applications si diverses trouvent leur point commun dans le chaîne de traitement et ils prennent toujours comme point de départ un caractère — un type de donnée informatique permettant de reconnaître non seulement des lettres, mais aussi des signes de ponctuation, les espaces, etc. (Issac, 2009 : 10). Ainsi la première tâche à effectuer consiste donc à dégager de ce flux de caractères un mot.

2. À titre indicatif : quelques particularités morphologiques les plus significatives de la langue polonaise

Selon la grammaire traditionnelle, le polonais distingue dix parties de discours qui se divisent en deux groupes : celles qui se déclinent ou se conjuguent comme le nom, l'adjectif, le numéral, le pronom, le verbe, et celles qui ne se déclinent pas : l'adverbe, la préposition, la conjonction, l'interjection, la particule.

Quatre des dix parties de discours (nom, adjectif, pronom, numéral) se déclinent selon sept cas : *nominatif*, *génitif*, *datif*, *accusatif*, *instrumental*, *locatif*, *vocatif*. La flexion des substantifs polonais est relativement compliquée, car elle doit prendre en compte les modifications casuelles. Théoriquement, chaque substantif polonais posséderait 14 formes nominales, sept pour le singulier et sept pour le pluriel. Contrairement au français qui n'en a que deux. La déclinaison se caractérise, le plus souvent, par un ajout de terminaisons flexionnelles au radical, mais nous observons souvent un changement du radical : *pies* [chien : subst, sg, nom, m2] / *psa* [chien : subst, sg, gén, m2]. Le polonais possède cinq genres : le masculin (qui est divisé en trois sous-types : masculin personnel, masculin animal et masculin inanimé), le féminin et le neutre. Le genre et le nombre se manifestent lors de la déclinaison. Les substantifs sont répartis en groupes de déclinaison selon le genre. Alors, le genre impose au substantif sa règle flexionnelle, par exemple *ryb**ak*** (masculin personnel), *pt**ak*** (masculin animal). Nous notons que ces deux substantifs ont la même terminaison ***-ak***. Cependant, leurs déclinaisons, et aussi leurs relations avec les autres éléments de la phrase, varient en fonction du genre du substantif : *ryb**ak*** (Nhsn) — *ryb**acy*** (Nhpn) et *pt**ak*** (Nasn) — *pt**aki*** (Napn). La grammaire traditionnelle généralise la déclinaison des substantifs polonais en proposant un tableau 1 (Nagórko, 2006 : 142). Une description traditionnelle de la flexion polonaise n'est pas applicable au traitement automatique des langues naturelles.

Tableau 1

Les terminaisons des substantifs singuliers

Genre	-Ø	-o	-ʻo	-e	-ę	-a	-ʻa	-i
masculin personnel	*chłop*	*Moniuszko*	*dziadunio*		*książę*	*poeta*		
masculin animal	*kot*		*piesio*					
masculin inanimé	*dom*	*okno*						
neutre				*morze*	*imię*			
féminin	*wieś*					*mama*	*ciocia*	*pani*

Les adjectifs qualificatifs se déclinent suivant le cas, le nombre et le genre. Les deux langues donnent la possibilité de graduer les adjectifs qualificatifs, sauf qu'en français, l'expression du degré est analytique (*heureux*, *plus heureux*, *le*

plus heureux) et en polonais, elle peut être synthétique (*szczęśliwy*, *szczęśliwszy*, *najszczęśliwszy*) ou analytique (*szczęśliwy*, *bardzo szczęśliwy*). La même remarque concerne la gradation des adverbes en français : *facilement*, *plus facilement*, *le plus facilement* et en polonais : *łatwo*, *łatwiej*, *najłatwiej.* Par contre, en polonais, très souvent les formes synthétiques ont leurs équivalents sous une forme analytique : *wesoło*, *bardziej wesoło* (*weselej*), *najbardziej wesoło* (*najweselej*).

Les verbes en polonais sont variables en fonction du nombre, du temps, de l'aspect, du mode, de la voix et de la personne. Traditionnellement, le polonais distingue l'aspect perfectif et l'aspect imperfectif qui est marqué lexicalement, contrairement au français où l'aspect est grammatical. L'infinitif en polonais indique s'il s'agit d'un verbe perfectif ou imperfectif. Le nombre et les valeurs de temps et de mode dans les deux langues sont très différents : le polonais distingue trois temps : le présent, le futur et le passé. La différence se situe aussi dans le fait que le français exprime la personne à l'aide de pronoms qui précèdent le verbe, tandis qu'en polonais, la personne est marquée par les terminaisons. Le résultat c'est que les pronoms personnels sont régulièrement omis.

3. Les analyseurs morphologiques du polonais

Les travaux sur la description systématique de la morphologie polonaise ont été initiés par Jan Tokarski (1958—1969). Tokarski était un des rédacteurs chef du dictionnaire de la langue polonaise qui compte 11 volumes. Il était responsable de la description morphologique. Ces travaux ont été ensuite repris par Zygmunt Saloni (2007) qui a complété les ressources de Tokarski et qui a finalement informatisé les ressources. Le premier analyseur morphologique basé sur ces ressources informatisées a été proposé par Krzysztof Szafran (1993). À l'époque, l'analyseur SAM-95[1] contenait 120 000 lemmes et il suggérait aussi une description morphologique des lemmes absents dans la base de données. L'analyseur SAM-95 était utilisé dans la création d'analyseur syntaxique du polonais basé sur la grammaire de Marek Świdziński (1992)[2]. L'analyseur Morfeusz[3] proposé par Marcin Woliński (2014) est une continuation des recherches dans ce domaine. Morfeusz, tout comme l'analyseur précédent, utilise les ressources linguistiques de Tokarski et de Saloni. La version actuelle de ce programme permet d'analyser seulement les mots qui se trouvent dans la base de données préalablement décrite. L'accès à Morfeusz est libre. Les données linguistiques recueillies par l'Univer-

[1] ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/SAM-95/ (accessible : 27.03.2015).
[2] Pour plus d'informations : http://www.mimuw.edu.pl/polszczyzna/ (accessible : 27.03.2015).
[3] http://sgjp.pl/demo/morfeusz?querytext=dam (accessible : 27.03.2015).

sité de Varsovie (Bień, Szafran — http://www.mimuw.edu.pl/polszczyzna/) ont été utilisées dans le projet du Corpus IPI PAN et dans le projet du Corpus National.

Nous retenons aussi les travaux initiés dans les années '90 du XX[e] siècle par l'Institut de Philologie Romane de l'Université de Varsovie sous l'impulsion des travaux du LADL et du LLI. Selon les données présentées pendant la conférence *NooJ '08* (8—10 juin 2008, Budapest) par Christophe Bogacki[4], les ressources lexicographiques du POLLEX compte presque 140 000 formes canoniques qui sont converties au format NooJ. La base de lemmes contient 75 140 formes simples dont les adverbes (1 126), les adjectifs (18 301), les substantifs (55 081), les prépositions (104), les verbes (208), les conjonctions (65), les interjections (129), les pronoms (26) et les autres (11)[5]. De plus, POLLEX comporte « un module générateur incorporant, d'un côté, une liste de racines (ou de bases de dérivation) et de l'autre, une grammaire qui utilise des schémas de flexion typique » (Bogacki, 1997 : 55).

4. La flexion des unités simples — modèle Proteus[6]

L'objectif est de proposer un analyseur morphosyntaxique du polonais qui associe le moteur de flexion et un dictionnaire de formes fléchies. Contrairement aux analyseurs précédents, l'étiquetage proposé devrait être compatible avec les étiqueteurs de la langue française. Alors, la description morphosyntaxique repose sur le modèle de description de la langue française appliquée au Morfetik (Mathieu-Colas, 2009). Certaines catégories sont propres à une langue analysée. Pour cela le système d'encodage est préalablement défini pour chaque langue et il se compose des éléments analogues et des éléments propres à une langue donnée. Par exemple, le lexème *okien* décliné au génitif du pluriel et son équivalent français *fenêtres* se caractériseront par encodage suivant : *Okien* — Nnpg et *Fenêtres* — Nfp.

La description des ressources linguistiques doit répondre aux besoins du TAL, alors nous étions amené à proposer une nouvelle répartition des catégories grammaticales. Les modifications concernent la suppression de la catégorie *liczebnik / numéral* qui n'est pas considérée comme une catégorie grammaticale, mais comme une sous catégorie grammaticale de certaines parties du discours. Nous avons

[4] *Extensions des ressources polonaises*, http://www.nytud.hu/nooj08/program/bogacki.pdf (accessible : 27.03.2015).

[5] Les données viennent du site internet http://www.nytud.hu/nooj08/program/bogacki.pdf (accessible : 27.03.2015).

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

ajouté une catégorie *determinant / déterminant* (Hajok, 2010). Cependant, nous avons retenu deux catégories propres au polonais : *odsłownik / substantif déverbal* et *partykuła / particule*, nous avons proposé respectivement les codes suivants : G et B.

Les catégories retenues sont présentées dans le tableau 2.

Tableau 2

Les catégories grammaticales

La catégorie en polonais	Le correspondant en français	Code
Czasownik	verbe	V
Determinant	déterminant	D
Odsłownik	substantif déverbal	G
Partykuła	particule	B
Przyimek	préposition	S
Przymiotnik	adjectif	Q
Przysłówek	adverbe	R
Rzeczownik	substantif	N
Spójnik	conjonction	C
Wykrzyknik	interjection	I
Zaimek	pronom	P
Znak interpunkcyjny	ponctuation	F

Pour rendre les étiquettes de substantifs opérationnelles pour le polonais, nous avons procédé à réaliser les modifications suivantes :

— nous avons ajouté les codes pour le cas : *n*, *g*, *d*, *a*, *i*, *l*, *v* ;
— nous avons ajouté les codes pour le genre[7] : *h*, *a*, *i*, *f*, *n*[8].

Les tableaux 3—5 illustrent les différences dans l'étiquetage des substantifs français (cf. tab. 3) et dans l'étiquetage des substantifs polonais (cf. tab. 4—5). Nous observons une complexité de la description de ces unités en polonais.

[7] 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.

[8] 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.

Tableau 3

Encodage des substantifs français

Attribut	Valeur	Exemple	Code
Catégorie	substantif	*garçon*	N
Genre	masculin	*garçon*	m
	féminin	*fille*	f
Nombre	singulier	*garçon*	s
	pluriel	*garçons*	p

Tableau 4

Encodage des substantifs polonais

Attribut	Valeur	Exemple	Code
Catégorie	substantif	*chłopiec*	N
Genre	masculin personnel	*chłopiec*	h
	masculin animal	*pies*	a
	masculin inanimé	*zeszyt*	i
	féminin	*dziewczynka*	f
	neutre	*dziecko*	n
Nombre	singulier	*dziecko*	s
	pluriel	*dzieci*	p
Cas	nominatif	*dziecko*	n
	génitif	*dziecka*	g
	datif	*dziecku*	d
	accusatif	*dziecko*	a
	instrumental	*dzieckiem*	i
	locatif	*dziecku*	l
	vocatif	*dziecko*	v

Tableau 5

Encodage des déterminants polonais

Attribut	Valeur	Exemple	Code
Catégorie	déterminant	*determinant*	D
Type	démonstratif	*ten*	d
	possessif	*mój*	s
	indéfinie	*pewien*	i
	interrogatif — exclamatif	*jaki*	t
	relatif	*który*	r
	numéral	*jeden*	k
	particule	*ale*	b
	nominal	*tona*	n
	adverbial	*dużo*	r
Personne	première	*mój*	1
	deuxième	*twój*	2
	troisième	*jej*	3
Genre	masculin personnel	*mojego*	h
	masculin animal	*mojego*	a
	masculin inanimé	*mój*	i
	féminin	*moją*	f
	neutre	*moje*	n
Nombre	singulier	*mój*	s
	pluriel	*nasz*	p
Possesseur	singulier	*moi*	s
	pluriel	*nasze*	p
Cas	nominatif	*mój*	n
	génitif	*mojego*	g
	datif	*mojemu*	d
	accusatif	*mojego*	a
	instrumental	*moim*	i
	locatif	*moim*	l

Ce système d'encodage est retenu pour toutes les catégories grammaticales, variables et invariables. Lors du fléchissement, les étiquettes morphologiques s'ajoutent automatiquement.

Tableau 6

Table de lemmes

Lemme	Règle
Klasa	S_001
Sroka	S_002

Chaque unité est dotée de deux types de tables : les tables de lemmes (cf. tab. 6) et les tables de flexion (cf. tab. 7—8). Leurs structures diffèrent selon les catégories morphosyntaxiques. Les formes invariables possèdent seulement les tables de lemmes dans lesquelles nous listons les entrées.

Chaque règle est décrite dans Proteus. La constitution de règles demande trois opérations (cf. tab. 7—9). Par exemple, la génération du génitif singulier du substantif *RZĘSA* s'effectue en plusieurs étapes. À la forme canonique du substantif, nous appliquons un code 1P\a\3D/ami/. Le Proteus met de côté un caractère (1P\a\), ensuite il ajoute trois caractères (3D/ami/).

1. La première opération consiste à enlever la terminaison du lemme (SUB1)[9].

Tableau 7

SUB1

```
<?xml version=»1.0» encoding=»UTF-8»?>
<!DOCTYPE proteus SYSTEM «table.dtd»>
                    <proteus>
      <flex id=»sub1» type=»nonterm»>
                  <name>N</name>
                  <info>nom</info>
    <op type=»mask» value=»sub1term»>
        <item value=»sub1radical»/>
                    </op>
        </flex>
        <mask id=»sub1radical»>
      <info>enlever terminaison</info>
     <item ervalue=».»>R/\\a\\/V</item>
     </mask>
</proteus>
```

2. La deuxième opération consiste à ajouter des terminaisons appropriées au cas en question (SUB1TERM). En même temps, nous fournissons les informations sur les étiquettes morphologiques.

[9] Les tableaux 4—6 renvoient à l'analyseur morphologique Proteus proposé par Fabrice Issac de l'Université Paris 13, mais elles intègrent les données permettant fléchir automatique les unités simples de la langue polonaise.

Tableau 8

SUB1TERM

```
<?xml version = “1.0” encoding = “UTF-8”?>
<!DOCTYPE proteus SYSTEM "table.dtd">
<proteus>
        <desc>
                conjugaison des noms
        </desc>
        <flex id=“sub1term” type=“term”>
                <name></name>
        <info>substantifs groupe 1</info>
                <flex id=“fsn”>
                        <name>fsn</name>
                        <code>/a/</code>
                </flex>
                <flex id=“fsg”>
                        <name>fsg/fpn/fpa/fpv</name>
                        <code>/y/</code>
                </flex>
                <flex id=“fsa”>
                        <name>fsa</name>
                        <code>/ę/</code>
                </flex>
                <flex id=“fsd”>
                        <name>fsd/fsl</name>
                        <code>/ie/</code>
</flex>
<flex id="fsi">
                        <name>fsi</name>
                        <code>/ą/</code>
                </flex>
                <flex id=“fsv”>
                        <name>fsv</name>
                        <code>/o/</code>
                </flex>
                        <flex id=“fpg”>
                        <name>fpg</name>
                        <code>/ /</code>
                </flex>
                <flex id=“fpi”>
                        <name>fpi</name>
                        <code>/ami/</code>
                </flex>
                <flex id="fpl">
                        <name>fpl</name>
                        <code>/ach/</code>
                </flex>
                </flex>
</proteus>
```

3. La troisième opération met en relation les deux précédentes (SUBTOTAL).

Tableau 9

SUBTOTAL

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE proteus SYSTEM "table.dtd">
<proteus>
        <flex id="S_001" type="final">
        <name></name>
        <info>Substantifs type 1</info>
        <op type="add">
                <item value="sub1"/>
        </op>
    </flex>
    <flex id="S_002" type="final">
        <name></name>
        <info>Substantifs type 2</info>
        <op type="add">
                <item value="sub2"/>
        </op>
    </flex>
    <flex id="S_003" type="final">
        <name></name>
        <info>Substantifs type 3</info>
        <op type="add">
                <item value="sub3"/>
        </op>
    </flex>
```

Les résultats obtenus sont au format XML et ils 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
rzęsie	rzęsa	Nfsd/fsl	S_001
rzęsę	rzęsa	Nfsa	S_001
rzęso	rzęsa	Nfsv	S_001
rzęs	rzęsa	Nfpg	S_001
rzęsom	rzęsa	Nfpd	S_001
rzęsami	rzęsa	Nfpi	S_001
rzęsach	rzęsa	Nfpl	S_001

5. Conclusion et perspectives

Les travaux présentés sont loin d'être terminés. Nous envisageons de proposer :

(a) le fléchisseur automatique de toutes les unités simples du polonais. Il s'agit avant tout d'avancer dans les travaux sur les verbes. La reconnaissance des verbes dans le texte permettra (i) de vérifier le réel pourcentage des unités reconnues dans le texte, (ii) d'avancer dans la description d'une grammaire locale de cette langue, (iii) de générer automatiquement des phrases.

(b) le fléchisseur automatique des unités complexes. Nous avons déjà procédé à la déclinaison des noms composés (Hajok, à paraître). Cette déclinaison demande d'introduction de codes spécifiques permettant de gérer la flexion interne de ces suites. Les variations flexionnelles dépendent du degré de figement des séquences figées, autrement dit la flexion doit tenir compte d'éventuelle autonomie des éléments constitutifs. Un grand nombre d'unités composées se comportent comme des groupes nominaux libres. Alors les structures syntagmatiques ne devront pas poser de problèmes flexionnels car elles reposent sur le même principe flexionnel du syntagme libre. Autrement dit la combinatoire interne qui régit les formes flexionnelles est la même dans le cas des séquences libres et des séquences figées. Mais, les constructions composées ne sont pas flexionnellement régulières. Il est indispensable de prendre en compte non seulement les variations casuelles, mais aussi les variations du nombre (*chudy rok = chude lata*) et du genre (*mały malutki = mała malutka*). Or, les études sur les noms composés et les constructions N_Modif en polonais ont permis de dégager plusieurs types flexionnels (Hajok, à paraître).

Références

Bień Janusz, 1991: *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Bień Janusz, 2001: «Analiza morfologiczna języka polskiego w praktyce». *Bulletin de la société polonaise de linguistiques*, fasc. LVII, 2001, http://bc.klf.uw.edu.pl/88/1/JSB-KS-PTJ01.pdf (accessible : 14.09.2014).

Bogacki Krzysztof, 1997 : « POLLEX — un dictionnaire électronique morphologique du polonais ». *Bulletin de Linguistique Appliquée et Générale, Numéro Spécial Actes FRAC-TAL '97*, 55—63.

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, 2010 : *Étude sémantico-syntaxique de la détermination simple et complexe en français et en polonais. Approche contrastive.* Thèse de doctorat, Université Paris 13.

Hajok Alicja, à paraître : « Structure du dictionnaire électronique des noms composés polonais ». In : *EL DICCIONARIO : neología, lenguaje de especialidad, computación*. Congreso Internacional, Ciudad de México (octobre 2013), Université de Puebla, Mexique.

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

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

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

Saloni Zygmunt, 2007: *Czasownik polski.* Warszawa: Wiedza Powszechna.

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

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

Świdziński Marek, 1993: *Gramatyka formalna języka polskiego.* Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.

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

Woliński Marcin, 2014: "Morfeusz reloaded". In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis, eds.: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014.* Reykjavík, Iceland, ELRA, 1106—1111.