DOI: 10.15290/baj.2019.19.03

Krzysztof BOGACKI
Uniwersytet Warszawski
kbogacki@gmail.com
http://orcid.org/0000–0003–2755–4276

Agnieszka DRYJAŃSKA
Uniwersytet Warszawski
a.dryjanska@uw.edu.pl
http://orcid.org/0000–0003–1649–8408

LES PRÉNOMS ET LES PATRONYMES DANS LES RESSOURCES DICTIONNAIRIQUES POUR LE TRAITEMENT AUTOMATIQUE DU POLONAIS PAR NooJ

Toute opération de traitement automatique de la langue présuppose une analyse préliminaire du texte qui doit aboutir à une indexation la plus complète possible de ses composantes à différents niveaux : morphologique, sémantique, syntaxique. Le rôle de premier plan dans cette entreprise revient à différents types de dictionnaires formalisés, créés à cet effet. Plus leur nomenclature sera ample et les informations détaillées, moins il restera de lacunes dans la description du texte. Si on ne discute pas de la nécessité de constituer des dictionnaires généraux, on peut rencontrer des avis mettant en doute l'intérêt de créer des dictionnaires des noms propres. Ils forment un champ aux limites floues où sont regroupées les unités très hétérogènes par nature. On y retrouve, en effet, les noms de personnes, ceux de lieu, ceux d'artefacts, d'institutions, d'évènements et de dates. Or dans les textes ils sont omniprésents. Il ne fait aucun doute que ce soit le nom propre de personne qui passe pour prototypique pour l'ensemble des noms propres.

Sur le plan théorique, les noms propres intéressent, depuis des siècles, philosophes, logiciens, psychologues et linguistes. Ceux-ci énumèrent plusieurs propriétés parmi leurs traits définitoires. Ce qui frappe,

c'est le fait que sauf la mono-référentialité[1], les critères discriminant les noms propres semblent apparaître comme dépendants de la langue. Ainsi pour le français, sur le plan morphologique, on relève l'absence de flexion en genre et en nombre : « Les noms propres se définissent selon le fait qu'ils ont une flexion fixe, qu'ils sont invariablement d'un genre donné [...] mais aussi d'un nombre donné » (Togeby 1982 : 120).

Le polonais, quant à lui, ne peut pas faire abstraction de ces catégories qui constituent au contraire un axe d'oppositions central dans le paradigme des patronymes. Seuls quelques prénoms peuvent être considérés comme invariables en genre. Sont minoritaires les prénoms du type *Bronisław – Bronisława*, *Czesław – Czesława*, *Stanisław – Stanisława*, *Wacław – Wacława*, etc., constituant des couples où sont reliées par un lien morphologique une forme masculine et une autre féminine.

D'autres traits sont mentionnés par des auteurs francophones comme étant au centre des débats sur l'essence même des noms propres. Ce sont, sur le plan formel, l'absence de déterminant (inexistant en polonais) en position référentielle et, sur le plan sémantique, le manque de sens lexical[2], ce qui aurait pour conséquence que les noms propres relèvent de l'encyclopédie et non pas du dictionnaire[3] et sont intraduisibles[4]. Le seul contenu sémantique des noms propres se réduirait à l'acte de dénomination (Kleiber 1981 : 329).

Il n'est pas rare de lire que, théoriquement[5], disposant d'un dictionnaire complet des noms communs, il serait plus facile de créer automatiquement une liste de noms propres, en repérant les formes non-reconnues par le dictionnaire de base prenant à l'initiale une majuscule. Cette propriété facilite d'ailleurs leur repérage dans le traitement informatique des textes, où ils sont identifiés par ce trait-là à l'intérieur d'une phrase, non-précédés par un point, un point d'interrogation ou un point d'exclamation. Il est cependant facile d'objecter que nous

[1] Cf. Zeboudj (2011).

[2] Grevisse (1964 : 751) considère que « le nom commun est pourvu de signification, d'une définition, il est utilisé en fonction de cette définition » alors que « le nom propre n'a pas de signification véritable, de définition; il se rattache à ce qu'il désigne par un lien qui n'est pas sémantique, mais par une convention qui lui est particulière ».

[3] Pour une discussion critique voir Vaxélaire (2005) qui combat cette thèse.

[4] « Toute modification aboutit, non à une traduction d'un nom propre, mais à un nouveau nom propre » (Kleiber 1981 : 503).

[5] Deux conditions devraient être satisfaites: le texte soumis à l'analyse ne pourrait contenir aucune faute et le dictionnaire général devrait être complet.

avons là un indice plutôt qu'un critère de définition. Cette propriété est d'ailleurs loin d'être universelle. Il suffit de mentionner d'un côté l'allemand où tous les substantifs (appellatifs et noms propres) prennent toujours une majuscule à l'initiale et, de l'autre, rappeler le cas de certaines langues (p. ex. arabe, chinois, géorgien, thaï ou japonais), utilisant des alphabets où l'opposition entre majuscules et minuscules est inexistante[6]. Il faut se rendre à l'évidence : il serait illusoire de vouloir compiler une liste complète de noms propres. Elle serait infinie car outre ceux qui existent, elle devrait comporter des noms de personnes fictifs.

Selon Silberztein (2015 : 283), l'intérêt de créer un dictionnaire des noms propres serait faible « puisqu'on comprend que dans la phrase suivante : *X est parti de Y pour acheter le dernier modèle de Z* X est un nom de personne, Y un nom de lieu et Z un nom de produit, quelle que soit la forme de X, Y et Z. Autrement dit, lorsque nous aurons une description suffisamment complète de la langue, nous pourrons reconnaître les noms propres sans avoir à les recenser. »

Or, à ce raisonnement on peut opposer en polonais différentes situations où mention doit être faite du caractère proprial ou appellatif d'un substantif. En effet, certains mots (tels *biczyk*, *pieniążek*, *zmywak*, *żarnik*, etc.) ont une double flexion : soit comme noms communs masculins désignant des objets soit comme patronymes (avec une majuscule à l'initiale !), auquel cas ils s'opposent systématiquement à l'accusatif singulier :

Widzę biczyk (nom commun) – 'je vois un fouet' vs
Widzę Biczyka (patronyme) – 'je vois M. Biczyk'.

Même à l'intérieur des noms propres des subdivisions sont nécessaires pour rendre compte des différences morphologiques. P. ex. *Radom* employé comme nom de famille ou comme toponyme se fléchissent différemment :

Jadę do Radoma (patronyme) – 'je vais chez Radom'
Jadę do Radomia (nom d'une ville) – 'je vais à Radom'.

Dans la suite de ce texte, nous rendrons compte d'une recherche entreprise il y a plusieurs années et qui était inscrite dans une perspective plus large de mettre à la disposition des chercheurs utilisant NooJ des

[6] Les alphabets de ce type sont d'ailleurs utilisés par la majorité des langues.

ressources linguistiques – dictionnaires et grammaires locales – dédiées au polonais[7]. Ces ressources ne sont pas compatibles directement avec d'autres existant pour le polonais[8] et demanderaient une adaptation parfois onéreuse. En ce qui concerne les dictionnaires, à côté du lexique général comportant les verbes, les mots invariables et les noms communs, nous avons envisagé d'élaborer la morphologie des noms propres. Avec cette division, nous nous sommes inscrits dans une tradition qui selon Quemada (1967) remonte au XVI^{e} siècle. Les noms propres constituent un inventaire d'unités hétérogène dont nous présentons la première partie. Elle décrit les noms de personnes : prénoms et noms de famille et laisse de côté les autres grandes sous-classes de noms propres : les toponymes, les noms de marques et de produits, ceux d'institutions, d'évènements et de dates.

Dans le cadre de cet article, nous allons présenter deux approches : dictionnairique et grammaticale. La première consiste à créer un dictionnaire de noms de famille et de prénoms à partir d'un corpus formé de vedettes, récupérées dans des textes polonais et d'une bibliothèque de grammaires locales, représentant chacune un modèle de flexion pour un groupe de mots du corpus. Par contre, l'approche grammaticale, toujours en cours de développement, ne propose qu'un ensemble de grammaires locales, qui permettent de reconnaître des noms propres qui se terminent par des suffixes caractéristiques des noms de famille polonais sans avoir recours à une longue liste de patronymes. Ainsi donc, cette méthode est ouverte aux mots potentiels qui pourraient être créés un jour et seraient repérés pour la première fois dans un texte.

APPROCHE DICTIONAIRIQUE

Pour la constitution de notre corpus, nous avons sélectionné plusieurs sources contenant des noms de famille et des prénoms. Leur nombre est impressionnant, dépassant sans doute celui des ressources prises en compte pour la constitution du lexique général. Le corpus de noms propres traités a été constitué à partir de plusieurs sources publiées sur Internet. La première utilisée et qui a fourni l'essentiel de

[7] Les résultats, sous forme binaire, seront accessibles sur le site www.nooj4nlp.net.
[8] Cf. http://clarin-pl.eu/en/uslugi/

nos ressources (environ 400.000 patronymes) a été l'archive *nazwiska.zip* trouvée sur le site http://www.futrega.org/etc/nazwiska.zip inaccessible depuis quelque temps. Nous avons exploré d'autre part les listes publiées par Bronisław Wildstein (http://www.listaipn.pl/), celles des noms de la noblesse polonaise (http//szlachtarp.pl/lista-nazwisk.html), le site http://nazwiska-polskie.pl/, ainsi que des listes d'anciens élèves de diverses écoles, membres d'institutions, etc. En ce qui concerne les prénoms, nos données proviennent principalement de listes accessibles sur le site https://pl.wikipedia.org/wiki/Kategoria:Listy_imion. Quant au site http://www.bip19.098.pl/index.php, qui s'enorgueillit de contenir plus de 100.000 prénoms appartenant à une trentaine de langues, il ne permet pas de les sortir sous forme de listes. D'autres sites d'importance mineure ont fourni des prénoms bibliques, grecs, latins, persans, hongrois, lituaniens, germaniques, slaves.

Ce corpus a été alimenté en continu avec des noms entendus pendant des émissions télévisées ou à d'autres occasions, trouvés au cours de la lecture de la presse, lors de l'exploration de différents documents rédigés en polonais trouvés sur Internet et traitant de sujets divers : actualité politique et culturelle, économie, géographie, sport, etc. ce qui explique qu'il est hétérogène quant à l'origine des vedettes. Nous avons pris en compte les vedettes au masculin même si la forme trouvée était féminine. Le but de ce dictionnaire étant d'identifier et de baliser au point de vue de la forme morphologique (cas, genre, nombre) les mots trouvés dans les textes actuels, nous n'avons pas cherché à trancher dans chaque cas la question de savoir si tel patronyme est polonisé ou si au contraire le fait de l'avoir trouvé dans un article écrit en polonais et publié dans un site polonophone fournit plutôt un exemple illustrant les mécanismes utilisés pour incorporer un exonyme. On ne s'étonnera donc pas de trouver dans cette ressource les noms de personnes tels que *Stalin*, *Lenin*, *Putin*, *Hitler*, *Mussolini*, *Bismarck* ainsi que beaucoup d'autres patronymes étrangers : allemands, italiens, français non polonisés, anglais, juifs, russes, lituaniens, arabes, espagnols, etc., donnés à des Polonais ou simplement se référant à des étrangers. Nous n'avons pas voulu les exclure sur le critère de l'origine du patronyme. La conséquence en matière de constitution du corpus est que nous avons élaboré une ressource qui n'est pas un dictionnaire de noms propres polonais à proprement parler – patronymes et prénoms – mais plutôt une base de mots d'origines variées fonctionnant comme noms propres trouvés dans des textes polonais.

Cette décision et la technique utilisée pour collecter les vedettes de notre dictionnaire explique la présence de mots commençant par Ç (*Çaliş*, *Crişan*), X (*Xiong* à côté de *Xsiężopolski*, *Xsiężyński*), Y (*Yacoubi*, *Yagel*, *Yahalom*, *Yakar*), V (*Vacqueret*, *Vacula*, *Vahrenholt*, *Vaillant*, etc.), Ž (*Žáková*, *Žaloudik*, *Živojinović*) ou comportant à l'intérieur du mot un caractère absent en polonais : ä (*Brämer*, *Järvinen*), á (*Vargová*), č (*Havličková*, *Jovičić*), ç (*Açikgöz*), ö (*Akerström*), ř (*Jeřábek*, *Kovář*), ü (*Müller*), ş (*Altaş*), ů (*Brůna*, *Růžičková*), v (*Havlin*, *Havryshchuk*, *Devechy*), ž (*Hadžić*, *Možilis*).

Au total, nos archives contiennent plus de 466.000 vedettes dont 7.586 prénoms et 458.244 patronymes. Certaines apparaissent sur les deux listes. En effet, on relève des cas où une même forme peut fonctionner soit comme patronyme, soit comme prénom (cf. Ada Jerzy vs. Jerzy Stanisławski, Wacław Marek vs. Marek Bielecki) auquel cas ils sont fléchis selon deux modèles différents. Ce phénomène est plus fréquent avec les formes ayant la valeur de prénoms à l'origine mais qui sont utilisées comme patronymes en polonais (cf. Anna Klaus, Jerzy Klaus, Zygfryd Klaus, Anna Zygfryd, etc.) Le nombre total de formes flexionnelles reconnues dépasse 33 millions (33.611.781). Les prénoms génèrent 118.000 formes.

L'élaboration de cette ressource n'a pas pu se faire sans tenir compte du *Słownik gramatyczny języka polskiego* (SGJP) dans ses deux versions sur CD-ROM : 1 et 2. La troisième version est accessible sur Internet (http://sgjp.pl/o-slowniku/#liczby). A titre de comparaison, disons qu'elle ne tient compte que d'environ 4.000 noms propres. Le nombre total de vedettes s'élève à 334 845 dont 334 733 lexèmes. La description de l'ensemble a exigé la création de 1 116 schèmes dont 222 pour les verbes, 767 pour les substantifs, 77 pour les adjectifs et 47 pour les numéraux.

LES CATÉGORIES GRAMMATICALES DANS LA DESCRIPTION DES PATRONYMES[9]

Les catégories grammaticales retenues pour l'élaboration de cette ressource exigent un commentaire approfondi. Les entrées de notre dictionnaire sont toutes des substantifs, même si leur forme au départ

[9] On consultera aussi Rzetelska-Feleszko (2005), Woliński (2003), Przepiórkowski & Woliński (2003).

évoque la catégorie adjectivale canonique ou régime (cf. toutes les vedettes en *cki*, *-dzki*, *-ski*), adverbiale (*Zasadniej*), verbale (*Wędzisz*, *Wiej*, *Wypij*, *Zachwiej*, *Zawitasz*) ou phrastique (*Tumidaj*). De ce fait, ne contenant que des substantifs, ce dictionnaire ne comporte aucune catégorie verbale : temps, mode, personne, etc. Nous sommes dispensés d'établir la distinction de 3 ou 5 pour le masculin de même que de tenir compte de la division des substantifs neutres en 2 classes. Les noms et les prénoms attribués à des personnes sont du genre masculin humain (mo) ou féminin (f). Une autre distinction par contre s'est imposée. Elle concerne les féminins qui sont distingués en féminins désignant des personnes non-mariées (fnm) ou mariées (fm). Dans certains cas, elle est marquée par les formes spécialisées en ***-owa*** (fm – *Nowakowa*) et ***-ówna*** (fnm – *Nowakówna*) ou ***-anka*** (fnm – *Skarga* – *Skarżanka*). Cette distinction ne se retrouve que dans la classe des patronymes et n'existe pas parmi les prénoms. On note de plus en plus souvent la disparition de cette forme au profit de l'emploi de forme non-fléchie, en dépit de la recommandation réitérée de fléchir les noms de personnes dans la mesure du possible.

LES GRAMMAIRES LOCALES

La description de ce corpus a nécessité la construction de 46 grammaires locales pour les patronymes. Voici leur liste avec indication du nombre de vedettes fléchies selon chaque modèle. Le nom du modèle est emprunté au patronyme ou au prénom qui se fléchit selon les exigences formelles de la grammaire en question.
ADAMKA (9831), BABEL (4286), BABIARZ (11466), BAFTA (3613), BARACK (523), BERKOP (1078), BERUS (14823), BONDARA (3966), BUKŁAHA (1586), CHAMIEC (971), CHAPAŁA (4429), DROŃCZUK (60118), DRYJA (654), DUDA (3480), DUSŁA (51), FREDRO (33), GACA (2434), GIEDROYĆ (1382), GRUZIEC (38), GUSIEC (20), HEJKE (22658), HLED (3116), KAKIET (8245), KIEC (12778), KIERBEDŹ (355), KORENKO (11680), KOTULA (4138), KOWALSKI (127312), LUBAŃ (3944), MUNDZIA (1441), PACJA (75), POŁANIEC (1401), RELIGA (2207), ROGAŚ (3065), SCHUBERT (2149), SUCHOWOLEC (1283), SZLABA (9780), TUROWICZ (76867), TYCEL (4355), WETRÓW (2663), WIECZOREK (553), WIKTOROW (9429), WUNDER (1289), ZAWUŁ (2653), ZELMAN (20605).

En ce qui concerne les prénoms, les 71 schèmes flexionnels que nous avons construits sous forme de grammaires locales sont les suivants: ABEL (35), ADA (544), ADAŚ (401), ADO (1), AGRYKOLA (17), AGUSIA (24), AL (479), ALBERT (183), ALDERYK (100), ALFRED (193), ANDROMACHA (6), ANIA (10), ANIRUDH (16), ANTONI (31), ARTUR (456), BASIEŃKA (1), BAŚKA (10), BEATA (119), BOGUMIŁ (59), BOGUTA (5), BOGUWŁOŚĆ (4), BONAWENTURA (13), BRONISŁAW (1800), CHWALIBÓG (1), CIESZYMYSŁ (20), CZESŁAWA (591), DOBROŻYŹŃ (2), DOLORES (308), ELWIRA (102), FILIP (10), GAWEŁ (3), GORAZD (1), HENIEK (1), HENRYKA (60), IGA (23), IWO (10), IZABELA (54), JAN (370), JAREMA (11), JOKASTA (11), JOLKA (52), KAZIA (4), KAZIK (37), KRZYSIĄTKO (1), KRZYSIEK (7), KRZYSIO (15), KSENIA (5), LUDMIŁA (13), ŁUCJA (235), MARCELI (4), MAREK (981), MARIUSZ (632), MARYNA (3), MODEST (33), NATASZA (21), NERO (70), NIEDALIC (1), NIELUBIEC (1), NOE (5), PERPETUA (1), PROKOP (4), ROCH (24), SOFOKLES (56), SZCZĘSNY (1), TEKLA (12), UDU (26), ULA (53), WAWRZYNIEC (3), WERCIA (35), ZACHARY (75), ZAWISZA (1), ZBYSZKO (5).

La classe des prénoms comprend 1518 diminutifs. Nous n'avons pas procédé de façon systématique pour les créer, nous bornant à 3 ou 4 formes qui nous ont paru les plus fréquentes, correspondant aux prénoms les plus répandus. Le site http://www.imiona.net en offre des listes bien plus longues, p. ex. pour *Agnieszka* on trouve 58 diminutifs, pour 8 autres prénoms testés, les données ont été les suivantes: *Anna* 103, *Barbara – 61, Blanka – 47, Bohdan – 24, Elżbieta – 55, Konstanty – 14, Krzysztof – 75, Leon – 12*. Nous avons allongé sans difficulté chacune de ces listes d'autres formes. Nombreux sont les patronymes provenant d'une même racine, parfois apparaissant dans des versions différentes, combinée avec des suffixes différents (p. ex. *Abracham, Abrachamczyk, Abrachamiak, Abrachamicz, Abrachamik, Abrachamowicz, Abrachimowski, Abragimowicz, Abraham, Abrahamczik, Abrahamczyk, Abrahamej, Abrahamicz, Abrahamik, Abrahamow, Abrahamowicz, Abrahamowicz, Abrahamowski, Abrahamów, Abrahams, Abrahamsson, Abrahanik, Abrahimowicz, Abrahimowicz*).

La flexion des noms de famille pose problème non seulement à cause de la multitude des formes et des règles, mais aussi à cause des usages divergents.

La recommandation générale en ce qui concerne la flexion des noms de famille polonais et étrangers est de les fléchir dans la mesure du pos-

sible, c'est-à-dire s'il existe dans la langue un schéma de flexion. En cas de besoin, celui-ci devrait être modifié. Notre dictionnaire rendant compte du fonctionnement des noms propres dans des textes polonais, nous avons tenu compte des modèles flexionnels de cette langue, ce qui, dans le cas des noms propres d'origines étrangères, nous a conduits à aller souvent contre la réalité linguistique de la langue originelle. Ainsi, dans la nomenclature de notre dictionnaire, on trouvera environ 400 patronymes d'origine lituanienne. En les incorporant dans notre dictionnaire, nous avons pris soin de remplacer les noms se référant à des femmes célibataires terminés en *-aite* (tels que *Priszmontaite, Duczmanaite*) et ceux utilisés pour les femmes mariées (*Satkuniene, Rimaviciene, Tumoniene, Saboniene*) par les masculins les plus fréquents correspondants en *-as*, *-us*, *-is*, *-ys*: *Priszmantas, Duczmanas, Satkunas, Rimavicius, Tumonis, Sabonys.* Les seules exceptions à cette règle sont les patronymes des personnalités connues retenus sous la forme féminine correspondante (*Grybauskaite, Ramoniene*). La flexion décrite pour ces mots ne tient pas compte des particularités lituaniennes. Ainsi, les patronymes masculins en *-as*, *-us*, *-is*, *-ys* ont été reliés au paradigme BERUS ou laissés invariables, abandonnant ainsi le paradigme d'origine.

Face aux hésitations quant à l'observation des règles restrictives, mais souvent non observées, nous avons décidé d'adopter une position libérale qui n'exclut pas certaines formes, mêmes si elles sont considérées comme erronées par les puristes. Ainsi, on observe que les locuteurs rejettent souvent la flexion de certaines formes au pluriel et ne tiennent pas compte du paramètre [mariée]//[célibataire]. Nous avons donc Mi+f+p: *Panie Nowak* au lieu de Mi+fm+p: *Panie Nowakowe* ou de Mi+fnm+p: *Panny Nowakówny* ou encore *Państwo Nowak* au lieu de *Państwo Nowakowie*.

> *... co wykorzystano do podstawienia Banków i przeniesienia do Zakopanego dla pana Nowak sklep PIK pozorując jednego właściciela komornika Gotfryda Nowak...*
>
> (https://polandtimes.wordpress.com/2016/07/05/planowany-... efekt-domina-finasowego-trzech-krajow/)

> *Jeśli jest tak, to uważam, że jest to niesprawiedliwe dla pana Nowak, który ma rodzinę w Pl, a w Norwegii dostał umowę na rok...*
>
> (http://www.forum-norwegia.pl/viewtopic.php?t=26370&start=108)

Zapamiętałem doskonale ceremoniał parzenia kawy w klasie przez Panią Nowak, *ale to już zupełnie inna historia...*

(https://nk.pl/szkola/3541/forum/44)

...szanuję panią Kowalczyk *jako sportowca ale nie jako człowieka...*

(https: //eurosport.interia.pl/justyna-kowalczyk/news-justyna-kowalczyk-ujawnia-przezylam-zalamanie-nerwowe,nId,1436878)

... pan z nickiem GOŚĆ FORUM:D pisze 2009 07 13 o godz 20:47 do pana Suseł, który również negatywnie komentuje...

(https: //forum.trojmiasto.pl/WYDAJE-MI-SIE-RYWALIZACJA-PONIEWAZ-t166780,1,170.html)

Opter pour cet usage conduirait à exclure le paradigme suivant: *Nowak* (Mi+mo+s), *Nowaka* (Do+mo+s), *Nowakowi* (Ce+mo+s), *Nowaka* (Bi+mo+s), *Nowakiem* (In+mo+s), *Nowaku* (Lo+mo+s), *Nowaku* (Wo+mo+s), *Nowakowa* (Mi+fm+s), *Nowakowej* (Do+fm+s), *Nowakowej* (Ce+fm+s), *Nowakową* (Bi+fm+s), *Nowakową* (In+fm+s), *Nowakowej* (Lo+fm+s), *Nowakowa* (Wo+fm+s), *Nowakówna* (Mi+fnm+s), *Nowakówny* (Do+fnm+s), *Nowakównie* (Ce+fnm+s), *Nowakównę* (Bi+fnm+s), *Nowakówną* (In+fnm+s), *Nowakównie* (Lo+fnm+s), *Nowakówno* (Wo+fnm+s).

Pour les mots de ce type, nous avons décidé, lors de la création de l'algorithme flexionnel, de donner à la grammaire locale la forme modulaire:

NOWAK = :NOWAK1 + :HEJKE;

NOWAK1 = <E>/Mi+mo+s + a/Do+mo+s + owi/Ce+mo+s + a/Bi+mo+s + iem/In+mo+s + u/Lo+mo+s + u/Wo+mo+s + owie/Mi+mo+p + i/Mi+mo+p + ów/Do+mo+p + om/Ce+mo+p + ów/Bi+mo+p + ami/In+mo+p + ach/Lo+mo+p + i/Wo+mo+p + owie/Wo+mo+p + owa/Mi+fm+s + owej/Do+fm+s + owej/Ce+fm+s + ową/Bi+fm+s + ową/In+fm+s + owej/Lo+fm+s + owa/Wo+fm+s + owe/Mi+fm+p + owych/Do+fm+p + owym/Ce+fm+p + owe/Bi+fm+p + owymi/In+fm+p + owych/Lo+fm+p + owe/Wo+fm+p + ówna/Mi+fnm+s + ówny/Do+fnm+s + ównym/Ce+fm+p + owe/Bi+fm+p + ównie/Ce+fnm+s + ównę/Bi+fnm+s + ówną/In+fnm+s + ównie/Lo+fnm+s + ówno/Wo+fnm+s + ówny/Mi+fnm+p + ównych/Do+fnm+p + ównym/Ce+fnm+p + ówny/Bi+fnm+p + ównymi/In+fnm+p + ównych/Lo+fnm+p +

ówny/Wo+fnm+p + ówny/Wo+fnm+p + owie/Mi+fmo+p + ów/Do+fmo+p + om/Ce+fmo+p + ów/Bi+fmo+p + mi/In+fmo+p + ach/Lo+fmo+p + owie/Wo+fmo+p;
HEJKE = <E>/Mi+mo+s + <E>/Do+mo+s + <E>/Ce+mo+s <E>/Bi+mo+s + <E>/In+mo+s + <E>/Lo+mo+s + <E>/Lo+mo+s + <E>/Mi+mo+p + <E>/Do+mo+p + <E>/Ce+mo+p + <E>/Bi+mo+p + <E>/In+mo+p + <E>/Lo+mo+p + <E>/Lo+mo+p + <E>/Wo+mo+p + <E>/Mi+fm+s + <E>/Mi+fm+s + <E>/Do+fm+s + <E>/Ce+fm+s + <E>/Bi+fm+s + <E>/In+fm+s + ach/Lo+fm+s + <E>/Wo+fm+s + <E>/Mi+fm+p + <E>/Mi+fm+p + <E>/Do+fm+p + <E>/Do+fm+p + <E>/Ce+fm+p + <E>/Bi+fm+p + <E>/In+fm+p + <E>/Lo+fm+p + <E>/Wo+fm+p + <E>/Mi+fnm+s + <E>/Do+fnm+s + <E>/Ce+fm+p + <E>/Bi+fm+p + <E>/In+fm+p + <E>/Do+fm+p + <E>/Ce+fnm+s + <E>/Bi+fnm+s + <E>/In+fnm+s + <E>/Lo+fnm+s + <E>/Wo+fnm+s + <E>/Mi+fnm+p + <E>/fnm+p + <E>/Ce+fnm+p + <E>/Bi+fnm+p + <E>/In+fnm+p + <E>/Lo+fnm+p + <E>/Wo+fnm+p + <E>/Wo+fnm+p + <E>/Mi+mf+p + <E>/Do+mf+p + <E>/Ce+mf+p + <E>/Bi+mf+p + <E>/In+mf+p + <E>/Lo+fnm+p + <E>/Wo+fnm+p;

Le premier module (NOWAK1) est responsable de la flexion « recommandée élargie », tenant compte des formes variables selon le sexe et l'état civil, le second (HEJKE, utilisé d'ailleurs largement dans notre dictionnaire pour d'autres patronymes et prénoms) rend compte des formes invariables. Vu l'usage qui se répand, la double modularité se retrouve dans la majorité des grammaires. La modularité simple est le fait entre autres de la grammaire « adjectivale » représentée par KOWALSKI et HEJKE qui décrit les noms de famille invariables.

L'APPROCHE GRAMMATICALE

Cette approche, qui constitue une alternative par rapport à l'approche dictionnairique, est inspirée d'une méthode de reconnaissance des noms propres à partir du suffixe *-isme*, proposée par M. Silberztein dans les ressources françaises de NooJ[10] et exploite le fait qu'en polonais ils commencent par une majuscule en toute position dans la phrase. Cette méthode de repérage qui permet de discriminer dans un texte polonais toutes les formes se terminant par les suffixes caractéristiques des patro-

[10] Cf. aussi Daille & Morin (2000).

nymes polonais, est fondée sur la conception des grammaires locales dont l'exemple est présenté ci-dessous sous forme graphique. Contrairement à l'approche dictionnairique, l'approche grammaticale permet de prendre en charge les formes potentielles, jamais rencontrées dans un texte.

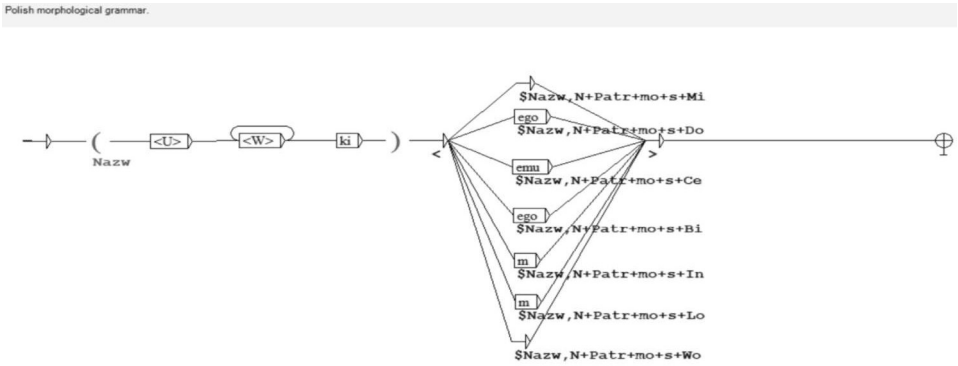

Figure 1 : Grammaire morphologique – noms de famille polonais en *ki*

Cette grammaire reconnaît tous les noms de famille non composés (littéralement toutes les formes en majuscule) qui se terminent en *ski*, en *cki* et en *dzki* et toutes leurs formes fléchies au singulier. Dans ce cas, elle identifie tous les mots qui commencent par une lettre majuscule (<U>), suivie d'un nombre quelconque de lettres minuscules (<W>) et terminés par le suffixe *c/dz/ski*. Cette séquence de caractères est ensuite suivie d'une désinence flexionnelle correspondant au paradigme de flexion des noms de famille en *ki*. Elle devra être complétée par les règles prenant en charge les formes féminines au singulier et toutes les formes au pluriel.

Figure 2 : Texte annoté à l'aide de la grammaire morphologique Nazw_ki.nom

Ces formes seront annotées dans un texte de façon suivante : le nominatif du mot dans le texte, N (nom) + Patr (patronyme) + mo (mas-

culin) + s (singulier) + nom de cas du mot dans le texte. Nous présentons ci-dessous un exemple de texte annoté avec cette grammaire (fig. 2).

Il est à noter que cette grammaire devra être appliquée avec la priorité la plus faible pour n'étiqueter que les formes inconnues, et non pas toutes les formes en majuscule (comme celles en début de phrase) (fig. 3), ou qu'il conviendra d'y ajouter une formule qui ne détecte que les noms propres non précédés d'un point, point d'interrogation ou point d'exclamation.[11]

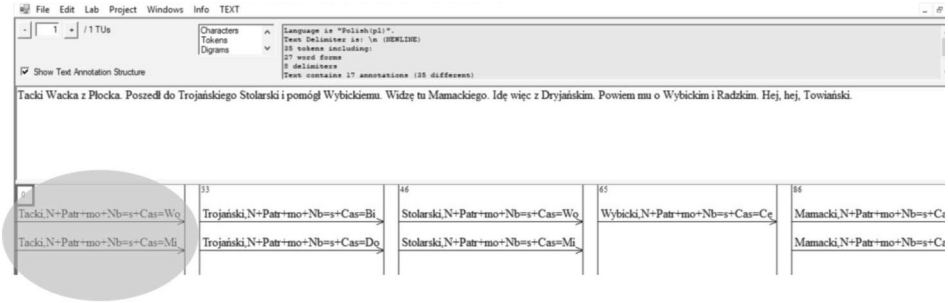

Figure 3 : Exemple d'une annotation erronée

La plateforme NooJ dispose aussi de la fonctionnalité « Générer une langue » pour créer un dictionnaire à partir d'une grammaire morphologique. Le résultat de cette opération est présenté sur la figure 3.

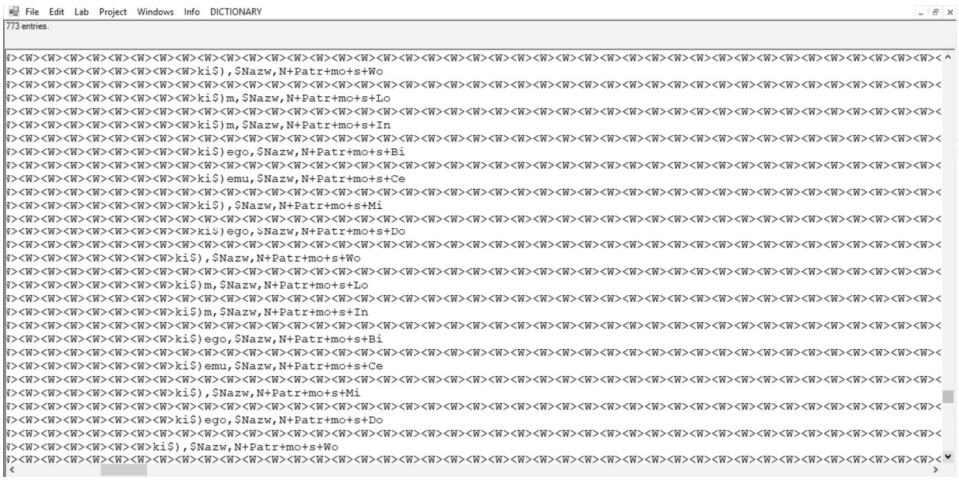

Figure 4 : Dictionnaire généré par la grammaire Nazw_ski.nom (fin du fichier)

[11] En début de phrase, on peut avoir aussi des noms de famille. Cette solution limite le nombre d'ambiguïtés mais ne les élimine pas toutes.

Afin de concevoir des grammaires locales reconnaissant les formes se terminant par tous les autres suffixes caractéristiques des patronymes polonais, il conviendrait de suivre le modèle présenté par la figure 1 pour le suffixe *ki* en recourant aux paradigmes flexionnels employés dans l'approche dictionnairique.

DICTIONNAIRES, HOMONYMIE ET AMBIGUÏTÉS OU L'HORIZON FUYANT ?

Si la première méthode exploitant un corpus extensif existant ou facile à constituer à partir des sources accessibles sur Internet permet un démarrage rapide, la méthode « grammaticale » a l'avantage de prendre en compte les vedettes potentielles, jamais trouvées dans des textes existants. L'approche grammaticale présente une alternative à l'approche dictionnairique et s'en distingue à l'étape préliminaire de la création d'un dictionnaire de noms propres. Exploitant la propriété formelle de ceux-ci, elle permet le repérage des lexèmes avec une majuscule à l'initiale à l'intérieur de la phrase et ouvre ainsi une voie à l'identification de noms propres potentiels. Les deux méthodes font appel ensuite à une bibliothèque de grammaires locales décrivant le mécanisme de flexion de différents types de noms propres.

Il est indéniable que l'utilisation d'un dictionnaire de noms propres ou de la technique « grammaticale » du repérage et de description de ceux-ci, de concours avec un dictionnaire général de grande couverture, contribue à diminuer le nombre de termes inconnus dans un texte soumis à l'analyse. Cependant, force est de reconnaître qu'en dépit des dimensions imposantes du dictionnaire, son usager n'est pas à l'abri des ambiguïtés qui peuvent se présenter dans nombre de cas. Elles découlent de l'homonymie, largement représentée et bien décrite pour le polonais[12], qui est de deux types : interne (observée à l'intérieur du paradigme d'un même mot) et externe (entre les paradigmes des vedettes différentes). Dans le module NOWAK1 que nous avons évoqué plus haut, l'homonymie interne concerne les 11 formes suivantes sur un total de 19 reconnues dans le paradigme, atteignant parfois 4 valeurs grammaticales pour une forme (cf. *Nowaki*, *Nowaków* et *Nowakówny*) :

[12] Pour une description plus détaillée voir en particulier Awramiuk (1999) et Buttler, Branicka & Tokarski (1984).

Nowaka (Do+mo+s, Bi+mo+s),
Nowaki (Mi+mo+p, Wo+mo+p, Mi+fmmo+p + Wo+fmmo+p),
Nowakowa (Wo+fm+s, Mi+fm+s),
Nowakową (Bi+fm+s, In+fm+s),
Nowakowe (Mi+fm+p, Bi+fm+p, Wo+fm+p),
Nowakowej (Do+fm+s, Ce+fm+s, Lo+fm+s),
Nowakowie (Mi+fmmo+p, Wo+mo+p,),
Nowaków (Do+mo+p, Bi+mo+p, Do+fmo+p, Bi+fmo+p),
Nowakównie (Ce+fnm+s, Lo+fnm+s),
Nowakówny (Mi+fnm+p, Do+fnm+s, Bi+fnm+p, Wo+fnm+p),
Nowaku (Lo+mo+s, Wo+mo+s).

En ce qui concerne l'homonymie externe, dans le domaine des noms propres, elle peut apparaître en particulier dans le cas des patronymes et des toponymes :

– le nom de famille *Drawska* (nominatif) et le nom de lieu *Drawska* (génitif dérivé de *Drawsko*),
– le nom de famille *Terespolska* (nominatif) et le nom de lieu (*rue*) *Terespolska* (nominatif).

Les mots comme *Drawsko*, *Drawski*, *Terespolska* se trouvant dans le dictionnaire de noms de famille et également dans le dictionnaire de noms de lieux, un texte annoté à l'aide de ces deux dictionnaires contiendra des ambiguïtés.

Si la méthode grammaticale permet de repérer au niveau lexical tous les mots susceptibles de fonctionner comme noms propres, elle ne permet pas d'éliminer l'ambiguïté au niveau des valeurs grammaticales ni ne donne pas de précisions sur la classe sémantique des formes identifiées (patronymes ?, prénoms ?, toponymes ? ou autres). En effet, l'ambiguïté grammaticale est importée au sein de la méthode grammaticale par le biais des grammaires locales décrivant le jeu des désinences casuelles des formes identifiées comme noms propres. La réduction de ce type d'ambiguïté passerait par le recours à l'analyse syntaxique du contexte d'apparition du nom propre. Identifiant sa position syntaxique au sein de la phrase, elle fournirait une information sur son cas grammatical. L'élimination de l'homonymie externe ferait appel à l'étiquetage multiple des vedettes introduisant des précisions sur leur type sémantique, ce qui est rendu possible par la structure ouverte d'une ligne de description des mots du dictionnaire autorisant l'adjonction d'informations variées

telles que l'origine du terme (allemand/français/italien, etc.), le registre de la langue (p. ex. diminutif/augmentatif, péjoratif, archaïque), identification de la profession dans le cas des noms propres renvoyant à des personnes (artiste/homme d'affaires/médecin...), etc. Le succès de l'opération dépendrait de la granularité de description des unités lexicales stockées. Elle est imprévisible à priori et dépendrait des besoins particuliers qu'on aurait face à un texte. Ainsi, voulant identifier les entités nommées désignant les machines agricoles, on serait obligé d'associer l'étiquette [+machine agricole] au nom propre *John Deer* qui peut avoir aussi comme référence un chanteur ou un homme d'affaires.

CONCLUSION

La conclusion finale se résumerait en ces mots : Le recours à un dictionnaire de noms propres, même de très large couverture, réduit certes le nombre de mots inconnus dans un texte analysé, sans pourtant éliminer lacunes et ambiguïtés.

BIBLIOGRAPHIE

Awramiuk E., 1999, *Systemowość polskiej homonimii międzyparadygmatycznej*, Białystok, Wydawnictwo Uniwersytetu w Białymstoku.

Buttler D., Branicka T. & Tokarski J. red., 1984, *Słownik polskich form homonimicznych*, Wrocław, Ossolineum.

Constanza J., 2016, *Nom propre et nomination : Etude d'un cas : la nomination des hommes politiques dans la presse écrite française*, thèse de doctorat, Tours.

Daille B. & Morin E., 2000, « Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations », in : *Traitement automatique des langues*, Vol. 41, no 3, pp. 601–621.

Grevisse M., 1964, *Le bon usage – Grammaire française*, Louvain, Duculot, Hatier.

Kleiber G., 1981, *Problèmes de référence : descriptions définies et noms propres*, Metz, Centre d'Analyse Syntaxique.

Przepiórkowski A., Woliński M., 2003, « A Flexemic Tagset for Polish », in: *Proceedings of the Workshop on Morphological Processing of Slavic Languages*, EACL 2003, pp. 33–40.

Quemada B., 1967, *Les dictionnaires du français moderne 1539–1863 – Etude sur leur histoire, leurs types et leurs méthodes*, Paris, Didier.

Rymut K., 2003–2005, *Słownik nazwisk używanych w Polsce na początku XXI wieku*, Kraków–Warszawa, GenPol Tomasz Nitsch.

Rzetelska-Feleszko E., éd., 2005, *Polskie nazwy własne*, Kraków, Instytut Języka Polskiego Polskiej Akademii Nauk.

Silberztein M., 2015, *La formalisation des langues : l'approche de NooJ*, London, ISTE Editions.

Togeby K., 1982, *Grammaire française – Vol. I : le Nom*, Copenhague, Akademisk Forlag.

Vaxélaire J.-L., 2005, *Les noms propres. Une analyse lexicologique et historique*, Paris, Honoré Champion.

Woliński M., 2003, « System znaczników morfosyntaktycznych w korpusie IPI PAN », in : *Polonica*, XXII–XXIII, pp. 39–55.

Zeboudj K., 2011, *Les dénominations monoréférentielles dans un guide touristique sur l'Algérie : approches linguistique et traductologique*, thèse de doctorat, Université de la Sorbonne Nouvelle – Paris III.

SITOGRAPHIE

http://nlp.actaforte.pl:8080/Nomina/Nazwiska

http://clarin-pl.eu/en/uslugi/

http://horajec.republika.pl/fakt28.html

http://nazwiska-polskie.pl/

http://stankiewicze.com/index.php?kat=44

http://szlachtarp.pl/lista-nazwisk.html

http://www.bip19.098.pl/index.php

http://www.futrega.org/etc/nazwiska.zip

http://www.herby.com.pl

http://www.jezykowedylematy.pl/2017/03/odmiana-nazwisk-dwuczlonowych-saryusz-wolski/

http://www.listaipn.pl/

http://www.forum-norwegia.pl/viewtopic.php?t=26370&start=108

https://pl.wikipedia.org/wiki/Kategoria:Alfabetyczna_lista_imion

https://sjp.pwn.pl/zasady/;629611

https://polandtimes.wordpress.com/2016/07/05/planowany-efekt-domina-finasowego-trzech-krajow

https://nk.pl/szkola/3541/forum/44

https://eurosport.interia.pl/justyna-kowalczyk/news-justyna-kowalczyk-ujawnia-przezylam-zalamanie-nerwowe,nId,1436878

https://forum.trojmiasto.pl/WYDAJE-MI-SIE-RYWALIZACJA-PONIEWAZ-t166780,1,170.html

LES PRÉNOMS ET LES PATRONYMES DANS LES RESSOURCES DICTIONNAIRIQUES POUR LE TRAITEMENT AUTOMATIQUE DU POLONAIS PAR NOOJ

Résumé

Cet article rend compte d'une recherche qui s' inscrit dans une perspective plus large de mettre à la disposition des chercheurs des ressources linguistiques – dictionnaires et grammaires locales – dédiées au polonais.

En premier lieu, nous présentons un dictionnaire électronique morphologique des prénoms et des patronymes au format NooJ. Le corpus pris en compte pour l'élaboration de cette ressource, constitué à partir de plusieurs sources publiées sur Internet, contient plus de 466.000 vedettes (7.586 prénoms et 458.244 patronymes). Cherchant à réduire les dimensions du dictionnaire, nous avons proposé une description modulaire qui a nécessité la création de plus de 40 grammaires locales pour les patronymes et presque le double pour les prénoms. En conséquence, le dictionnaire reconnaît plus de 33 Mo de formes. La solution ci-dessus – dictionnairique – présentant l'inconvénient d'être peu économique en ce qui concerne le temps et la taille des fichiers finals, nous avons proposé une approche grammaticale. Dans la dernière partie de l'article, nous expliquons cette démarche aussi bien que les avantages et les inconvénients des deux méthodes et des ambiguïtés sémantiques et grammaticales générées par elles.

Ensuite, nous justifions notre choix d'élaborer cette partie du lexique et, après un bref survol des propriétés qui distinguent les noms propres des noms communs, nous présentons celles qui en polonais ont un impact direct sur la forme des mots retenus et constituent les principaux axes d'opposition entre eux. Outre les catégories grammaticales ayant un impact direct sur la forme (cas, genre et nombre), nous mentionnons, pour les prénoms, leur origine (slave, latine, grecque, biblique, etc.). Face aux hésitations quant à l'observation des règles d'usage restrictives, mais souvent non observées, nous avons décidé d'adopter une position libérale qui n'exclut pas certaines formes même si elles sont considérées comme erronées par les puristes.

Mots-clés : NooJ, traitement automatique des langues naturelles, patronymes, prénoms

DICTIONARY OF FIRST NAMES AND SURNAMES FOR THE AUTOMATIC TREATMENT OF POLISH BY NOOJ

Summary

This paper reports on a study whose purpose was to provide researchers specializing in the automatic treatment of natural languages with linguistic resources dedicated to Polish, namely dictionaries and local grammars.

Firstly, a morphological dictionary of first names and surnames in NooJ format is presented. The corpus for the dictionary, made up of texts collected from several sources published on the Internet, contains more than 466,000 headwords (7 586 first names and 458 244 surnames). Seeking to reduce the size of the dictionary, we propose a modular approach for the construction of local grammars. It requires, however, the creation of more than 40 local grammars for surnames and almost double for first names. The dictionary recognizes altogether about 33 MB of forms. As the solution based on a list of first names and surnames is time- and disc space-consuming, we introduce another approach – based on local grammars only. In the final part of the paper, we discuss the advantages and disadvantages of both solutions, as well as semantic and grammatical ambiguities that cannot be overcome in both approaches.

Secondly, we discuss the reasons for the choice of this part of the lexicon, and next, having given a brief overview of the properties that distinguish proper nouns from the common names, we describe these properties that have a direct impact on the forms of surnames in Polish and constitute the main sources of opposition among them. In addition to the grammatical categories (case, gender and number) affecting surnames' forms, we also point out their origin (Slavic, Latin, Greek, biblical etc.). As for the observance of the usage rules of Polish surnames, very strict or more flexible, we have adopted a liberal approach that does not exclude certain forms, although they can be considered erroneous by purists.

Key words: NooJ, automatic treatment of natural languages, surnames, first names