La concurrence suffixale dans la construction des adjectifs dénominaux en Russe : analyse des suffixes -n-, -sk- et -ov-

Natalia Bobkova

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 2 - Jean Jaurès

Présentée et soutenue par

Natalia BOBKOVA

Le 20 septembre 2023

La concurrence suffixale dans la construction des adjectifs dénominaux en russe : analyse des suffixes -n-, -sk- et -ov-

---

Ecole doctorale : **CLESCO - Comportement, Langage, Education, Socialisation, Cognition**

Spécialité : **Sciences du langage**

Unité de recherche :
CLLE - Unité Cognition, Langues, Langage, Ergonomie

Thèse dirigée par
Vladimir BELIAKOV et Fabio MONTERMINI

Jury

M. Olivier BONAMI, Rapporteur
Mme Irina KOR CHAHINE, Rapporteure
M. Dunstan BROWN, Examinateur
M. Nabil HATHOUT, Examinateur
M. Sergueï SAKHNO, Examinateur
Mme Juliette THUILIER, Examinatrice
M. Vladimir BELIAKOV, Directeur de thèse
M. Fabio MONTERMINI, Co-directeur de thèse

Remerciements

Je souhaite exprimer ma gratitude envers mes deux directeurs de thèse, qui ont chacun contribué de manière significative à ce travail. Je tiens à remercier Vladimir Beliakov pour avoir accepté de diriger cette recherche, pour sa rigueur et sa réactivité, ainsi que pour nos échanges enrichissants autour de la thèse. Les suggestions de lecture qu'il m'a faites m'ont permis de progresser de manière autonome dans mon parcours académique. Je tiens également à exprimer ma reconnaissance à Fabio Montermini pour sa pédagogie et sa patience lors de nos discussions sur la thèse et sur la morphologie et la linguistique en général. Son accompagnement dans mes recherches, ses recommandations bibliographiques et méthodologiques, ainsi que son investissement dans le contenu et la structure de mes textes, ont été d'une grande aide.

Je remercie Irina Kor Chahine et Olivier Bonami pour avoir dédié leur temps en tant que rapporteurs de ma thèse, leur expertise et leur expérience sont d'une grande valeur. Mes remerciements vont également à Dunstan Brown, Nabil Hathout, Sergueï Sakhno et Juliette Thuilier pour leur participation au jury, leur implication est hautement appréciée.

Ce travail porte l'empreinte de plusieurs personnes dont la contribution a été déterminante tant pour sa forme que pour sa direction actuelles. En premier lieu, je tiens à remercier Frédéric Pont pour sa passion contagieuse pour la programmation et d'avoir joué un rôle crucial dans l'élaboration de la version bêta de mon corpus. Un remerciement particulier va également à Juliette Thuilier, qui m'a initiée à l'univers des statistiques et avec qui j'ai passé de moments intenses tentant de résoudre la problématique de la concurrence suffixale en russe au début de ce parcours quantitatif. Je tiens aussi à exprimer ma gratitude à Nabil Hathout, qui m'a aidée à dévoiler les mystères de la sémantique distributionnelle et aussi pour ses retours constructifs en tant que membre de mon comité de suivi de thèse. Enfin, mes remerciements vont à Sergueï Sakhno, le deuxième membre de mon comité, pour ses précieuses remarques sur ma recherche et pour son travail inspirant qui m'a largement motivée à entreprendre cette thèse.

Je remercie tous les membres du laboratoire CLLE pour leur accueil : Cécile Fabre, Josette Rebeyrolle, Franck Amadieu, Anne Przewozny, Anne Condamines, Fabio Del Prete, Basilio Calderone, Franck Sajous, Dejan Stosic, Anne Le Draoulec, et j'en passe.

Je souhaite exprimer ma reconnaissance à tous ceux qui ont rendu possible le financement de mes recherches. En particulier, je tiens à remercier Émilie Merdy pour m'avoir fait confiance et intégrée dans son équipe à Synapse Développement, ainsi que Patrick Séguéla pour m'avoir permis de découvrir le secteur privé. Je voudrais

exprimer mes sentiments chaleureux à toute l'équipe des synapsiens, en surtout aux anciens : à Anouk et Audrey pour leur esprit d'équipe ; à Émilie pour avoir cherché à me faire développer les compétences en TAL que j'aime tant ; à Anaïs, Guilhem, Nicolas, Thibault D., Loïc, Quentin, Corentin, Kevin et Vincent pour leur bienveillance et surtout pour leur patience et disponibilité pour tout debug ; à Sophie qui m'a beaucoup appris le relationnel. Une reconnaissance spéciale est adressée à l'équipe Altaïr dans sa version finale : Angéline, Charles, Clément et Thibault B. Merci pour l'ambiance, la solidarité et le partage d'expériences et de tâches passionnantes. Au-delà du monde de l'entreprise, j'ai également eu la chance de continuer à enseigner le russe au département de slavistique de l'Université de Toulouse 2 Jean Jaurès au début de ma thèse. Je suis reconnaissante envers le département des sciences du langage où j'ai acquis de nouvelles compétences en enseignement de linguistique et surtout à Jean-Michel Tarrier, Mai Ho-Dac, Ludovic Tanguy, Myriam Bras (et aussi à Fabio Montermini). Merci de m'avoir confié vos cours et pour tous les échanges que nous avons eus. Enfin, je remercie le département LPGA de la Sorbonne la Nouvelle, et tout particulièrement Marine Wauquier et Yoann Dupont, pour leur accueil et pour avoir facilité non seulement mon intégration, mais aussi la préparation des cours.

Je remercie mes collègues de bureau pour tous les moments que nous avons partagés, que ce soit autour d'un repas, d'un café ou d'une discussion cryptique en regex au tableau : Marine, Marc Philippe, Lison, Chiara, Alain et Lena, ces petits moments ensemble ont toujours été agréables. Mes pensées vont à Émilie et Karla, dont le soutien a été inestimable durant des moments particulièrement éprouvants. Je tiens également à remercier chaleureusement mes amis doctorants : Julie R., Giusi, Aleksandra, Nataly, Léa, Pavel, Bénédicte, Camilla, Julie H., Luce, Maxime, Daniele, Silvia, Filip, Yizhe, Océane. Votre amitié a rendu ce voyage universitaire très enrichissant.

Je tiens à remercier mes amis qui ont été mon soutien et une source de réconfort tout au long de ce parcours. Julie, avec qui nous avons fait tant de projets passionnants, cherchant une harmonie entre les équations de Lotka-Volterra et les suffixes russes. Marie, pour les moments partagés autour du thé et de la cuisine, et pour l'ambiance accueillante et inspirante de sa maison, qui a souvent été mon refuge pour la rédaction. Anastasia, pour m'avoir fait voyager à travers le monde entier tout en restant chez moi. Laurent pour sa passion pour russe, son assiduité et sa bonne humeur. Finalement, Sasha, Dima, Kristina, Zhenia, bien que séparés par la distance, vous avez su être là. Nos retrouvailles, même après de longs intervalles, ont toujours eu l'air d'un rendez-vous continu, comme si nous nous étions vus la veille.

Ce travail n'aurait pas été accompli sans l'appui constant de mes parents, Liudmila et Sergueï, ainsi que de ma tante Valentina. Leur encouragement et leur confiance ont été un soutien indispensable tout au long de ce parcours, me stimulant à progresser, même face aux défis les plus redoutables.

Enfin, un merci tout particulier à Léo. Ta constance face au chaos de ma thèse m'a offert l'espace de sérénité nécessaire pour mener à bien ce travail. Je n'en serais pas là où je suis aujourd'hui sans ton amour et ton soutien.

Sommaire

Liste des figures

Liste des tableaux

Conventions typographiques

Dans cette thèse, les exemples en russe seront systématiquement écrits en utilisant des caractères cyrilliques en petites majuscules pour les lexèmes (1a) ou en italique pour mots-formes (1b), accompagnés d'une translittération et d'une traduction entre guillemets. Si un suffixe flexionnel est présent dans un mot-forme, il sera séparé du thème par un tiret ('-'). Les gloses morphologiques seront renseignées pour chaque mot-forme.

(1) a. КНИГА 'livre'
KNIGA

b. МОСТ 'pont' : *мост-ов* 'pont-M.GEN.PL'
MOST *most-ov*

Les conventions d'annotation pour les adjectifs qualificatifs et relationnels sont différentes : les adjectifs qualificatifs seront accompagnés de leur traduction (2a), tandis que la traduction des adjectifs de relation correspondra à celle du nom de base (2b). Si l'adjectif apparaît seul dans l'exemple (2c), la traduction sera accompagnée d'une glose correspondante[1].

(2) a. УМ 'esprit' → УМНЫЙ 'intelligent'
UM UMNYJ

b. КНИГА 'livre' → КНИЖНЫЙ
KNIGA KNIŽNYJ

c. ВКУСОВОЙ 'goût$_{\text{REL}}$'
VKUSOVOJ

Les exemples marqués en cyrillique pourront être lus en utilisant le système de translittération du russe, tel que décrit par Aslanoff (1986). Ce système reproduit la forme écrite du mot en combinant la transcription phonologique et la translittération

[1]La discussion sur la sémantique de différentes classes adjectivales sera présentée dans la section 2.2.3.

graphique communément utilisées en linguistique des langues slaves (Corbett et Comrie, 2003, pp.xii-xiii,57-58 : Timberlake, 2004, p.25). Pour citer des affixes dans le texte, nous utiliserons la translittération (par exemple, pour traiter les suffixes dérivationnels, comme -*n*- ou -*sk*-, ou les suffixes flexionnel – -*ij* ou -*yj*).

La translittération ne fournit pas d'informations sur la prononciation des mots. Cependant, certains cas nécessitent une analyse phonologique plus approfondie. Il n'est généralement pas nécessaire d'utiliser la transcription phonétique ; la plupart des phénomènes linguistiques examinés dans cette étude peuvent être représentés par la transcription phonologique. Pour ce faire, nous utiliserons deux variantes de transcription phonologique en parallèle. La première correspond à ce qui est communément pratiqué en linguistique au niveau international (Corbett et Comrie, 2003, p.54) et utilise les caractères en API. Dans cette transcription, seuls les phénomènes majeurs de la phonologie sont marqués (il s'agit de la position de l'accent (ˈ) et de la palatalisation (^{j}) pour le russe)[2]. La deuxième transcription est propre aux slavistes (Garde, 1998, pp.30-33), et utilise les caractères latins avec des diacritiques à la place de l'API. Cette transcription reflète également les phénomènes majeurs, avec une autre représentation pour la palatalisation (') ; quant à l'accent, il est marqué directement sur les voyelles. L'avantage de ce système est qu'il est plus proche de la translittération, et donc plus lisible pour les non-linguistes. Pour citer les phonèmes dans le corps du texte, nous utiliserons par commodité la transcription des slavistes (par exemple, pour présenter les alternances entre /k/ et /č/).

Le tableau 1 présente une synthèse des trois conventions utilisées dans cette thèse pour chaque graphème (`Orth`) : la translittération linguistique (`Tr`), la transcription phonologique en API (`API`) et la transcription phonologique des slavistes (`Slav`).

Il convient de préciser certaines caractéristiques des transcriptions phonologiques utilisées dans cette thèse[3].

Premièrement, dans les deux transcriptions, il existe deux manières de transcrire les lettres <*e, ë, ю, я*> : /e, o, u, a/ et /je, jo, ju, ja/. Le premier ensemble de phonèmes est utilisé lorsque les lettres correspondantes sont précédées d'une consonne molle (3a) ; la deuxième – lorsqu'elles apparaissent en début de mot (3b), après une voyelle (3c) ou après les signes dur et mou (3d). Il est important de noter que les transcriptions phonologiques pour <ъ> 'signe dur' et <ь> 'signe mou' sont vides, car ces deux graphèmes ne correspondent pas à des phonèmes. Le signe mou indique que la consonne précédente est molle (3e), tandis que le signe dur indique qu'elle ne l'est pas (3f).

(3) a. мять /m^{j}at^j/ /m'at'/ 'froisser'
mjat'

b. ель /jel^j/ /jel'/ 'sapin'
el'

[2]Ces phénomènes serons discutés en détail dans la section 1.2.

[3]Les phénomènes cités brièvement ici seront traités plus en détail dans la section 1.2.

		Transcription				Transcription	
Orth	Tr	API	Slav	Orth	Tr	API	Slav
а	a	/a/	/a/	**п**	p	/p/, /pʲ/	/p/, /p'/
б	b	/b/, /bʲ/	/b/, /b'/	**р**	r	/r/, /rʲ/	/r/, /r'/
в	v	/v/, /vʲ/	/v/, /v'/	**с**	s	/s/, /sʲ/	/s/, /s'/
г	g	/g/, /gʲ/	/g/, /g'/	**т**	t	/t/, /tʲ/	/t/, /t'/
д	d	/d/, /dʲ/	/d/, /d'/	**у**	u	/u/	/u/
е	e	/je/, /e/	/je/, /jo/	**ф**	f	/f/, /fʲ/	/f/, /f'/
			/e/, /o/	**х**	x	/x/, /xʲ/	/x/, /x'/
ё	ë	/jo/, /o/	/jo/, /o/	**ц**	c	/t͡s/	/c/
ж	ž	/ʒ/	/ž/	**ч**	č	/t͡ʃʲ/	/č/
з	z	/z/, /zʲ/	/z/, /z'/	**ш**	š	/ʃ/	/š/
и	i	/i/	/i/	**щ**	šč	/ʃʲː/	/šč/
й	j	/j/	/j/	**ъ**	”	//	//
к	k	/k/, /kʲ/	/k/, /k'/	**ы**	y	/ɨ/	/i/
л	l	/l/, /lʲ/	/l/, /l'/	**ь**	’	//	//
м	m	/m/, /mʲ/	/m/, /m'/	**э**	è	/e/	/e/
н	n	/n/, /nʲ/	/n/, /n'/	**ю**	ju	/ju/, /u/	/ju/, /u/
о	o	/o/	/o/	**я**	ja	/ja/, /a/	/ja/, /a/

Tableau 1: Translittération linguistique et transcriptions phonologiques – en API et celle des slavistes – du cyrillique

c. НОЯБРЬ /noˈjabrʲ/ /nojábr'/ ‘novembre’
NOJABR’

d. РУЖЬЁ /ruʒˈjo/ /ružjó/ ‘fusil’
RUŽ’Ë

e. ПОЛКА /ˈpolka/ /pólka/ ‘étagère’
POLKA
ПОЛЬКА /ˈpolʲka/ /pól'ka/ ‘polka (danse)’
POL’KA

f. СЕСТЬ /sʲestʲ/ /s'est'/ ‘s’asseoir’
SEST’
СЪЕСТЬ /s͡jestʲ/ /sjest'/ ‘manger’
S"EST’

Deuxièmement, dans la tradition slaviste, la lettre <*e*> peut correspondre à /e/ ou /je/, mais aussi à /o/ ou /jo/. Le choix dépend de la position de l’accent dans le paradigme flexionnel d’un mot. Si le <*e*> est toujours présent en position accentuée, /e/ ou /je/ sont utilisés dans la transcription (4a). Sinon, lorsque <*ë*> apparaît au moins une fois sous l’accent (4b), les phonèmes /o/ et /jo/ seront utilisés. Selon l’école phonologique de Moscou, certains affixes présentent également des anomalies.

Par exemple, le suffixe flexionnel adjectival du masculin singulier /oj/ correspond graphiquement à <*oй*> en position accentuée (4c). Cependant, en position atone, il est orthographié comme <*ый*> ou <*ий*> (4d)[4].

(4) a. ЛЕС ‘forêt’ :
LES
лес-а /lʲeˈsa/ /l'esá/ ‘forêt-M.NOM.PL’
cf. *лес* / lʲes/ /l'es/ ‘forêt$_{\text{M.NOM.SG}}$’

b. НЕСТИ ‘porter’ :
NESTI
нес-л-а /nʲesl'a/ /n'oslá/ ‘porter-PST-F.SG’
cf. *нёс* / nʲos/ /n'os/ ‘porter$_{\text{PST.M.SG}}$’

c. ПРОСТОЙ /prosˈtoj/ /prostój/ ‘simple’
PROSTOJ
ГЛУХОЙ /gluˈxoj/ /gluxój/ ‘sourd’
GLUXOJ

d. НОВЫЙ /ˈnovɨj/ /nóvoj/ ‘nouveau’
NOVYJ
РЕДКИЙ /ˈrʲedkʲij/ /r'édkoj/ ‘rare’
REDKIJ

Enfin, il est important de noter que toutes les représentations phonologiques données ci-dessus dans le tableau 1 ne sont pas considérées comme des phonèmes. Ainsi, /ja, je, jo, ju/ représentent chacun un groupe de deux phonèmes. De même, /g', k', x'/ ainsi que /ɨ/ représentent les variantes des phonèmes /g, k, x/ et /i/ selon l’école phonologique de Moscou[5].

[4]Pour plus d’anomalies, cf. Garde (1998, p.44).
[5]La discussion plus détaillée sera présente dans la section 1.2.

Notations

$\models$	Implication
*	Construction morphologique ou syntaxique inacceptable
$>$	Relation étymologique
?	Construction difficilement acceptable ; acceptable sous certaines conditions
$\sim$	Relation phonologique, morphologique ou lexicale
$\nrightarrow$	Relation dérivationnelle orientée absente
$\rightarrow$	Relation dérivationnelle orientée
ACC	Accusatif
ADV	Adverbe
APPT	Adjectif d'appartenance
A	Adjectif
DAT	Datif
GEN	Génitif
INSTR	Instrumental
LOC	Locatif
M	Masculin
NOM	Nominatif
N	Nom

N	Neutre
PL	Pluriel
REL	Adjectif de relation
SG	Singulier

Introduction

Le but de la présente recherche sera d'étudier le phénomène de la concurrence suffixale en s'intéressant en particulier aux adjectifs dénominaux en russe.

L'utilisation d'adjectifs est considérable dans le système nominal du russe, et une grande proportion d'entre eux sont dérivés. Selon Townsend (1975), il existe seulement entre 200 et 300 adjectifs simples en russe contemporain. La dérivation des adjectifs est principalement effectuée par l'affixation et la composition ; la suffixation demeure une opération morphologique primordiale (Šanskij et Tixonov, 1987). Sorokina (1984, p.18) indique que près de 98% des adjectifs russes sont dérivés à partir des noms, tandis que les 2% restants sont dérivés des verbes, des autres adjectifs ou des adverbes. Les néologismes adjectivaux, formés à partir des substantifs, occupent la deuxième place après les néologismes nominaux. Généralement, l'apparition d'un nouveau nom dans la langue (par exemple, un emprunt) entraîne la formation d'un adjectif dénominal ayant un sens générique 'relatif à ce que le nom de base désigne' (5), selon les travaux de Zemskaja (2000).

(5) a. ВАУЧЕР 'coupon' → ВАУЧЕРСКИЙ
VAUČER VAUČERSKIJ

b. КОТИРОВКА 'cotation' → КОТИРОВОЧНЫЙ
KOTIROVKA KOTIROVOČNYJ

c. РЕЙТИНГ 'classement' → РЕЙТИНГОВЫЙ
REJTING REJTINGOVYJ

d. ИМИДЖ 'réputation' → ИМИДЖЕВЫЙ
IMIDŽ IMIDŽEVYJ

Actuellement, la langue russe possède un système de suffixation complexe pour dériver les adjectifs à partir des noms. Les règles de formation les plus productives pour cette catégorie d'adjectifs comprennent les suffixes -*n*-, -*sk*- et -*Ov*-[6]. D'après

[6]Le *O* majuscule dans ce cas représente une voyelle qui peut correspondre, graphiquement, à <*o*> ou <*e*>, cf. (5c) et (5d).

Zemskaja (1965), ces trois suffixes produisent la majorité des adjectifs russes et se combinent avec des noms de base de plus en plus variés en termes de structure et de classes sémantiques. Ces trois suffixes adjectivaux sont les plus productifs dans tous les types de discours (neutre, littéraire, technique, etc.) (Zemskaja, 2015, p.206,227). Vinogradov (1952, p.163) précise que *-n-*, *-sk-* et *-Ov-* permettent de construire des adjectifs à la fois qualificatifs et relationnels.

En raison de leur productivité élevée et de leur utilisation pour la formation d'adjectifs qualificatifs et relationnels qui sont présents dans des discours variés, les suffixes *-n-*, *-sk-* et *-Ov-* feront l'objet d'une étude approfondie dans la présente thèse.

Lorsque plusieurs suffixes productifs sont utilisés pour dériver les adjectifs en appliquant une opération sémantique similaire, leurs schémas de construction peuvent alors être en concurrence dans le système dérivationnel. Cette similarité offre aux locuteurs un éventail de possibilités au niveau formel. Les développements récents en morphologie dérivationnelle considèrent que différents types de contraintes (phonologiques, morphologiques, sémantiques, pragmatiques, etc.) présentent une interaction complexe qui résulte en choix d'un des suffixes concurrents (Baayen *et al.*, 2013).

En dehors des cas où la concurrence est résolue grâce aux préférences des suffixes à des informations linguistiques spécifiques (6a), il existe des situations où deux, voire même trois, formes concurrentes coexistent (6b).

(6) a. Париж 'Paris' →
Pariž
парижский / *парижный / *парижевый
parižskij parižnyj pariževyj

b. слесарь 'menuisier' →
slesar'
слесарный / слесарский / слесаревый
slesarnyj slesarskij slesarevyj

La littérature sur la morphologie du russe décrit le phénomène de la concurrence entre les différents procédés morphologiques, mais ces études restent principalement descriptives et se concentrent sur des cas spécifiques (Townsend, 1975 ; Švedova, 1980 ; Hénault et Sakhno, 2015 ; Zemskaja, 2015). Cependant, comme le soulignent Trubeckoj et Jakobson (2004), une analyse statistique de la morphologie est nécessaire pour évaluer les phénomènes observés dans la langue, tels que la productivité, la concurrence et la coexistence de doublets.

La linguistique contemporaine a fait un pas important en passant de l'approche systémique à l'approche basée sur l'usage. L'approche systémique étudie le langage idéal en mettant de côté les méthodes expérimentales et objectives. L'intuition précède l'analyse pour éliminer la variation et obtenir un système langagier parfait et statique. L'approche basée sur l'usage, quant à elle, étudie les textes (base empirique) dont les particularités ne sont pas connues avant l'analyse et découvertes au cours de celle-ci,

en utilisant un corpus représentatif de la langue. Elle se focalise sur l'étude des faits et non pas des normes, et sur l'admission du fait qu'il existe plusieurs stratégies pour exprimer le sens (Plungjan, 2008). L'attitude envers ce qu'on appelle une norme linguistique est plus flexible, la distinction entre erreur et variante marginale est plus fluide. L'approche basée sur l'usage considère qu'il n'y a pas de langue en tant que telle, mais des structures dominantes qui peuvent être étudiées. Même dans un domaine relativement conservateur comme la morphologie, il existe de nombreuses différences par rapport à l'image présentée par les descriptions normatives. L'analyse de corpus implique l'étude de toutes les formes rencontrées dans les textes, qui sont plus nombreuses que celles prescrites par la norme.

Bien que l'analyse quantitative soit un champ de recherche en croissance en linguistique russe, il y a encore de nombreux sujets de grammaire et de lexique qui n'ont pas encore été couverts. C'est pourquoi cette thèse se basera sur des méthodes numériques, quantitatives et statistiques pour étudier le phénomène de concurrence morphologique.

Pour effectuer des analyses statistiques, il est nécessaire d'utiliser des corpus de grande taille. Nous allons fonder notre étude sur les données du Corpus National de la Langue Russe (RusCorpora).

Les hypothèses que nous allons vérifier dans la présente recherche sont les suivantes :

H_1 : Les propriétés des noms de base imposent des contraintes sur le choix d'un des suffixes adjectivaux ; ce choix est déterminé par l'interaction complexe de ces contraintes ;

H_2 : Les adjectifs de basse fréquence (et notamment des hapax) ne suivent pas les mêmes tendances constructionnelles que les adjectifs de haute fréquence, l'inventaire de propriétés de leurs noms de base étant plus large ;

H_3 : Malgré le fait que les trois suffixes en question soient considérés comme les plus productifs en russe moderne, leur productivité peut être exprimée en termes mathématiques précis permettant de définir le suffixe le plus productif et le moins productif ;

H_4 : L'émergence des doublets ne doit pas être interprétée comme une preuve de synonymie absolue, mais doit être examinée en fonction de la distribution des propriétés des noms de base et des fréquences relatives des adjectifs, ainsi que selon des contextes partagés par chaque adjectif dans un couple de doublets.

Du point de vue théorique, cette thèse adopte une approche lexématique, selon laquelle un lexème est défini comme une unité manipulée par la morphologie dérivationnelle, qui possède trois niveaux d'informations linguistiques : la forme (phonologique ou graphique), la catégorie lexicale et le sens lexical référentiel. Le lexème englobe un ensemble de mots-formes, selon les définitions proposées par Matthews, 1974 ; Mel'čuk, 1993 ; Fradin, 2003 ; Haspelmath et Sims, 2013. Il est doté d'un ou plusieurs thèmes sur lesquels les formes fléchies sont formées, un de ces thèmes peut également être utilisé pour la dérivation. Il est possible qu'un de ces thèmes soit également utilisé dans la dérivation. Toutefois, le thème peut subir certaines

modifications pour s'adapter, ce qui conduit à la construction d'un radical distinct pour la dérivation (Roché, 2010). Conformément à la notation proposée par Matthews (1974), les lexèmes seront indiqués avec des petites majuscules et accompagnés d'une traduction entre guillemets. Les mots-formes seront transcrits en italique et accompagnés de gloses morphologiques, comme mentionné dans les Conventions typographiques.

La structure de la recherche se présente comme suit.

Dans la partie I, nous présenterons des considérations théoriques relatives à la phonologie et à la morphologie du système nominal et adjectival en russe. Nous aborderons les allomorphies thématiques dans le système nominal, qui sont nécessaires pour comprendre le fonctionnement du russe et pour les modélisations détaillées dans la partie III. Dans le chapitre 1, nous décrirons les particularités phonologiques et morphologiques du russe. Dans le chapitre 2, nous délimiterons l'objet d'étude des adjectifs dénominaux et présenterons l'état actuel des connaissances sur les différentes classifications des adjectifs, leurs propriétés syntaxiques et sémantiques, ainsi que les différents moyens de dérivation des adjectifs en russe. Enfin, dans le chapitre 3, nous examinerons le phénomène de la concurrence en morphologie, en mettant l'accent sur la concurrence entre les suffixes -*n*-, -*sk*- et -*Ov*-. Cette analyse nécessitera une discussion sur la productivité des règles morphologiques, les différentes méthodes de blocage, ainsi que les régimes de coexistence et de spécialisation des doublets.

La partie II traitera de la constitution du corpus général des adjectifs et de leurs noms de base, ainsi que des annotations associées. Dans le chapitre 4, nous présenterons le Corpus National de la Langue Russe, les méthodes utilisées pour extraire les adjectifs avec leurs fréquences, ainsi que les trois sous-corpus de données finaux : les adjectifs de haute fréquence, les adjectifs de basse fréquence (notamment les hapax) et les données de doublets. Le chapitre 5 décrira les méthodes utilisées pour reconstituer les noms de base des adjectifs, et traitera des questions de double motivation pour les adjectifs dénominaux. Nous présenterons également les particularités des annotations phonologiques, morphologiques, sémantiques et étymologiques des noms, ainsi que les statistiques descriptives associées. Dans le chapitre 6, nous décrirons les méthodes utilisées pour réduire la subjectivité des annotations manuelles, en particulier en ce qui concerne l'étymologie et la sémantique des noms.

La partie III sera consacrée aux études de cas sur la concurrence entre les suffixes -*n*-, -*sk*-, -*Ov*-. Le chapitre 7 présentera une modélisation statistique de la concurrence entre les suffixes à l'aide de méthodes multifactorielles. Les propriétés phonologiques, morphologiques, sémantiques et étymologiques seront examinées pour évaluer leur pertinence dans le choix d'un des suffixes concurrents. L'interaction de ces contraintes sera également analysée. Dans le chapitre 8, nous étudierons les fréquences absolues et relatives des adjectifs et de leurs noms de base. Les fréquences relatives permettront de mesurer à quel point un lexème construit est considéré comme une unité à part entière ou comme une unité composée d'un radical et d'un suffixe. Nous étudierons également la productivité des suffixes en question. Selon la littérature, les trois suffixes sont productifs ; nous établirons le degré de productivité de chaque suffixe. Enfin,

le chapitre 9 présentera une étude exploratoire des doublets. Nous analyserons les régimes de coexistence de deux adjectifs formés à partir de la même base nominale en examinant la distribution des propriétés de leurs noms de base, de leurs fréquences, ainsi que l'inventaire des noms recteurs qu'ils modifient.

Partie I

État de l'art

Chapitre 1

Phonologie et morphologie du russe

Sommaire

Introduction

Dans ce chapitre, nous exposerons les caractéristiques fondamentales de la phonologie et de la morphologie du russe. La section 1.1 aura pour objectif de faire un panorama général sur la langue russe, en présentant sa place parmi les langues slaves ainsi que ses caractéristiques principales. La section 1.2 sera dédiée aux spécificités des systèmes consonantique et vocalique. Les catégories morpho-syntaxiques majeures des noms et des adjectifs seront examinées dans la section 1.3. Enfin, la section 1.4 présentera le sujet des allomorphies thématiques qui se produisent au cours de la flexion et de la dérivation russes, notamment des adjectifs dénominaux.

1.1 Présentation du russe

Les langues slaves sont principalement parlées dans une grande partie de l'Europe orientale et centrale. Le russe, la langue avec le plus grand nombre de locuteurs, s'est répandue depuis son berceau en Europe de l'est vers l'Asie du nord jusqu'à l'océan Pacifique (Corbett et Comrie, 2003, p.1). Il est utilisé comme langue maternelle ou seconde par environ 150 millions de personnes. Il est la langue officielle de la Fédération de Russie, ainsi que l'une des langues officielles du Bélarus, du Kazakhstan, du Kirghizistan et du Tadjikistan, dominante dans certaines régions de l'Ukraine et utilisée comme langue de communication dans la Communauté des États indépendants (CEI). Le russe est également parlé dans d'autres pays de l'ex-URSS, ainsi que dans les communautés russes à travers le monde.

Le russe standard moderne est une langue slave appartenant au groupe slave oriental (avec l'ukrainien et le biélorusse), il présente également des caractéristiques de langue synthétique. Le terme 'russe standard' implique que son usage est reconnu comme correct par l'ensemble de la communauté : il est admis dans la communication écrite et orale, prend en compte toute la diversité des styles (style littéraire, journalistique, poétique, familier, vulgaire) mais exclut les dialectes et les régionalismes. Le terme 'russe moderne' exclut toute considération historique et, selon certaines études (Garde, 1998, pp.9-11), ne remonte pas plus haut que le XIXème siècle.

Le russe est écrit avec les caractères cyrilliques. Le cyrillique, avec quelques variations, est également utilisé par d'autres langues telles que l'ukrainien, le biélorusse, le bulgare, le macédonien et le serbe. L'alphabet cyrillique a été utilisé dès le Moyen Âge pour l'écriture des langues des peuples slaves orthodoxes (vieux slave) et s'est répandu en Bulgarie à partir du IXe siècle, puis en Russie, Ukraine, Bélarus, Serbie, Bosnie, Macédonie et Monténégro (Corbett et Comrie, 2003, p.20). Après la formation de l'URSS, cet alphabet a également été utilisé pour les langues turques, finno-ougriennes, caucasiennes et mongoles. Le cyrillique russe a pris sa forme actuelle après la réforme orthographique de 1918 et les principes établis lors de cette réforme ont été formalisés en 1956 (Timberlake, 2004, pp.10-23).

Depuis le XVIIe siècle, le russe a été enrichi par des emprunts aux langues européennes qui se distinguent du fond russe et slave. Le russe a également subi quelques modifications à la fin des années 1980 et au début des années 1990, après la chute de l'URSS, principalement des changements dans le lexique avec l'enrichissement par de nombreux emprunts, ainsi que leur intégration dans le système morphologique du russe et l'apparition de nombreux dérivés (Timberlake, 2004, p.3 ; Garde, 1998, p.11).

1.2 Le système phonologique

1.2.1 Les phonèmes russes

Avant de décrire les particularités phonologiques du russe, il est pertinent de présenter son système graphique. L'alphabet russe, également connu sous le nom de cyrillique, comprend 33 caractères : 21 consonnes, 10 voyelles, un signe mou <ь> et un signe dur <ъ>. Le principe de cet alphabet est basé sur la phonologie : il est possible de passer de la succession des lettres d'un mot écrit à celle des phonèmes et inversement en utilisant un nombre limité de règles simples (cf. les Conventions typographiques) (Garde, 1998, p.31).

La description phonologique du russe est complexe en raison de l'absence de consensus quant au nombre de phonèmes qui composent son système phonologique.

Il est généralement admis que le russe possède 5 phonèmes pour les voyelles (/a/, /e/, /i/, /o/ et /u/). Cependant, les linguistes de l'école de Saint-Pétersbourg attribuent un statut phonémique à */ɨ/ (<ы>). Selon l'école phonologique de Moscou, /ɨ/ se trouve en distribution complémentaire par rapport à /i/ et est considéré comme un allophone de /i/ (<*u*>). En effet, /i/ et /ɨ/ ne partagent pas les mêmes contextes phonologiques : /i/ se trouve après les consonnes molles (7a), tandis que /ɨ/ se trouve après les consonnes dures (7b). En position initiale c'est le /i/ qui est utilisé, /ɨ/ en début des mot fait partie des exceptions (7c), généralement des noms propres ou des toponymes d'origine étrangère.

(7) a. СИЛА /ˈsʲila/ /s'íla/ 'force'
SILA

b. СЫР /sɨr/ /sɪr/ 'fromage'
SYR

c. Ыйзу /ˈɨjzu/ /íjzu/ 'Õisu (Estonie)'
Yjzu

Dans cette recherche, nous adopterons la perspective de l'école phonologique de Moscou, selon laquelle le russe possède cinq phonèmes vocaliques. Ces phonèmes vocaliques correspondent aux deux séries de graphèmes vocaliques systématisées dans le tableau 1.1.

Première série	а	э	о	у	ы
Deuxième série	я	е	ё	ю	и
Phonème	/a/	/e/	/o/	/u/	/i/

Tableau 1.1: Séries des graphèmes vocaliques et leurs phonèmes correspondants

Ainsi, chaque phonème vocalique peut être transcrit en utilisant deux graphèmes différents. Par exemple, le phonème vocalique /a/ peut être représenté par les graphèmes

<*а*> ou <*я*> (8a), le phonème vocalique /e/ peut être représenté par les graphèmes <*э*> ou <*е*> (8b), etc.

(8) a. БА**Г**РЯНЕЦ /bagˈrʲanʲets͡/ /bagr'án'ec/ ‘pourpre$_{N}$’
BAGRJANEC

b. **Э**КЗАМ**Е**Н /ekˈzamʲen/ /ekzám'en/ ‘examen’
ÈKZAMEN

Le nombre de phonèmes consonantiques en russe est aussi sujet à débat parmi les linguistes. Selon l'école linguistique de Moscou, il existe 32 phonèmes consonantiques dans la langue russe alors que l'école linguistique de Saint-Pétersbourg en liste 36 (Garde, 1998, p.17). Il existe des différences de classification phonologique entre les deux écoles, notamment lorsqu'il s'agit de phonèmes vélaires /k, g, x/ lorsqu'ils sont dotés du trait de mouillure, et quant à l'existence du phonème /šč/ correspondant au graphème <*щ*>[1].

Dans la suite de cette thèse, le point de vue de l'école phonologique de Moscou, qui considère qu'il existe 32 phonèmes consonantiques en russe, sera adopté.

Il est à noter que chaque voyelle en russe forme toujours une syllabe. Les phonèmes consonantiques, à leur tour, ne peuvent jamais former une syllabe à eux seuls.

Le nombre total de phonèmes en russe est de 37. Ainsi, il existe un déséquilibre entre le nombre de graphèmes de consonnes et de voyelles dans l'alphabet russe. Il y a 21 graphèmes de consonnes pour 32 phonèmes consonantiques, et 10 graphèmes de voyelles pour 5 phonèmes vocaliques. Ce phénomène est expliqué par l'évolution de la langue russe au cours des siècles, où la moitié des phonèmes vocaliques a été perdue, tandis que le nombre de phonèmes consonantiques a doublé (Breuillard et Viellard, 2015, p.19). Ainsi, dans la langue russe contemporaine, le même graphème peut être utilisé pour représenter deux phonèmes consonantiques différents (9).

(9) ВЕ**С** /vʲes/ /v'es/ ‘poid’
VES
ВЕ**С**Ь /vʲesʲ/ /v'es'/ ‘entier’
VES’

Il existe une corrélation entre l'abondance des graphèmes vocaliques et le manque de graphèmes consonantiques en russe. Le fait que certains phonèmes consonantiques soient représentés par les mêmes graphèmes est lié au fait que les graphèmes vocaliques changent pour les mêmes phonèmes correspondants (10).

(10) МАЛЫЙ ‘petit’ : ***ма****л* /mal/ /mal/ ‘petit$_{\text{M.NOM.SG}}$ (court)’
MALYJ *mal*
МЯТЬ ‘froisser’ : ***мя****-л* /mʲal/ /m'al/ ‘froisser-PST.M.SG’
MJAT’ *mja-l*

[1]Pour une discussion plus détaillée sur les divergences entre les deux écoles phonologiques russes, cf. Comtet (1995).

La phonologie du russe est marquée principalement par deux phénomènes (Timberlake, 2004, pp.28-29) : l'altération du trait de mouillure pour les consonnes et la variation de la position de l'accent tonique pour les voyelles.

1.2.2 Variantes des consonnes

Le trait de mouillure est considéré comme un trait distinctif pour les langues slaves selon Corbett et Comrie (2003, p.6) : de nombreuses consonnes forment des paires minimales distinguées par le trait de mouillure, le cas le plus extrême étant celui du russe. L'altération de la propriété de mouillure est considérée comme fondamentale et positionnelle[2] pour les consonnes russes. Dans ce qui suit, nous traiterons le trait de mouillure comme un trait fondamental ; les propriétés positionnelles seront discutées dans la section 1.4.2[3].

En russe, la plupart des consonnes peuvent être prononcées de deux façons distinctes : ces consonnes possèdent un point d'articulation palatal secondaire lorsque la partie médiane de la langue se rapproche du palais dur et adopte ainsi une position similaire à celle de la voyelle /i/ ou la semi-voyelle /j/. Dans la tradition phonologique russe, les consonnes articulées de cette manière sont appelées *consonnes molles*. Les consonnes désignées comme *consonnes dures* ne présentent pas cette articulation[4]. Les consonnes dures ressemblent généralement aux consonnes françaises.

Comme les couples de mouillure sont formés de phonèmes distincts, ces derniers peuvent être utilisés pour différencier des lexèmes (11a), ainsi que des mots-formes (11b).

(11) a. ТОМНЫЙ /ˈtomnɨj/ /tómnoj/ 'langoureux'
TOMNYJ
ТЁМНЫЙ /ˈtʲomnɨj/ /tˈómnoj/ 'sombre'
TËMNYJ

b. ГОТОВЫЙ 'prêt': *готов* /goˈtov/ /gotóv/ 'prêt$_{\text{M.NOM.SG}}$ (court)'
GOTOVYJ *gotov*

[2]Nous utilisons la terminologie proposée par Garde (1998, p.52) selon laquelle une variante fondamentale est celle dont la position forte ne peut être définie que par la formule négative 'dans tous les autres cas'. Quant à la variante positionnelle, elle est caractérisée par sa présence dans des positions qui peuvent être définies de manière positive, ainsi qu'en position faible.

[3]Dans la présente recherche, nous nous limitons à l'altération du trait de mouillure. Cf. Garde (1998, pp.59-74) ou Timberlake (2004, pp.68-74) pour d'autres types d'altération, comme altération du trait de sonorité.

[4]En slavisitique, à côté du terme mouillure, on retrouve aussi *mollesse* et *palatalité* (Breuillard et Viellard, 2015, p.99). Dans la tradition anglo-saxonne, ce phénomène est généralement désigné sous le nom de *palatalization*, cf. par exemple, Timberlake (2004, pp.56-63) ou (Corbett et Comrie, 2003, pp.828-831). Cependant, le terme *palatalisation* est réservé dans la slavistique au domaine de la phonétique historique et désigne les déplacements du point d'articulation vers le palais dur, ce qui entraîne un changement d'ordre pour les consonnes : une vélaire devient chuintante ou sifflante (Breuillard et Viellard, 2015, p.65). Nous utiliserons le terme de palatalisation pour parler des allomorphies thématiques dans la section 1.4.

ГОТОВИТЬ ‘préparer’ :	*готовь* /go'tov^{j}/	/gotóv'/ ‘préparer$_{\text{imp.sg}}$’
GOTOVIT’	*gotov’*	

En général, les graphèmes consonantiques en russe correspondent aux deux phonèmes qui forment des couples composés de deux consonnes qui ne diffèrent que par la présence ou l’absence de mouillure. Ces consonnes sont les dentales et les labiales. Les autres consonnes ne possèdent pas de correspondant qui ne diffère d’elles que par la mouillure, telles que les consones affriquées, fricatives, vélaires et la semi-voyelle /j/.

Il est important de noter que les consonnes /k, g, x/ ont des variantes molles positionnelles qui ne se manifestent pas dans tous les contextes phonologiques. Par exemple, les vélaires molles se retrouvent devant les voyelles /e, i/ (12a-12c), leur apparition devant les /o, a, u/ est très rare et se produit dans les noms propres et les toponymes d’origine étrangère (12d-12e)[5] ou dans les abréviations. Les vélaires molles n’apparaissent jamais en position finale des mots ou devant une voyelle. Les variantes molles ne servent pas à différencier les mots en opposition aux variantes dures. Ces variantes sont en distribution complémentaire et représentent pour chacune des trois vélaires deux réalisations d’un seul et même phonème (Garde, 1998, p.60).

(12) a. КИНУТЬ /'k^{j}inut^{j}/ /k'ínut'/ ‘jeter’
KINUT’

b. ХИТРЫЙ /'x^{j}itrɨj/ /x'ítroj/ ‘rusé’
XITRYJ

c. ГЕЛЬ /g^{j}el^{j}/ /g'el'/ ‘gel’
GEL’

d. ГЁТЕ /'g^{j}ote/ /g'óte/ ‘Goethe’
GËTE

e. КЯХТА /'k^{j}axta/ /k'áxta/ ‘Kiakhta (Mongolie)’
KJAXTA

Le trait de mouillure peut être systématisé de la manière suivante :

- Couples de mouillure :
  - Labiales : /p-p'/, /b-b'/, /f-f'/, /v-v'/, /m-m'/
  - Dentales : /t-t'/, /d-d'/, /s-s'/, /z-z'/, /n-n'/, /l-l'/, /r-r'/
- Toujours dures : /c, š, ž/
- Toujours molles : /č, j/
- Consonnes à trait de mouillure conditionné : /k, g, x/

[5]Ces rares apparitions sont toutefois suffisantes pour considérer les vélaires molles comme phonèmes à part entière, notamment par l’école phonologique de Saint-Pétersbourg.

Les consonnes qui ne font pas partie de couples (qui sont toujours dures ou toujours molles) ne dépendent pas du contexte phonologique. Du point de vue algorithmique, la répartition des consonnes en couples en fonction du trait de mouillure révèle la nature régressive de cette altération : elle est conditionnée par ce qui se trouve à droite de la consonne en question dans le contexte phonologique :

- Les consonnes dures sont suivies par les voyelles de la première série <*а, э, ы, о, у*> ;
- Les consonnes molles sont suivies par les voyelles de la deuxième série <*я, е, и, ё, ю*>.

Par exemple, les consonnes molles se retrouvent dans (13a), et les consonnes dures – dans (13b). Si aucune voyelle ne suit la consonne (par exemple, en position finale), le trait de mouillure est marqué par un signe mou <*ь*> (13c). En absence de signe mou la consonne reste dure (13d).

(13) a. земля /z^{j}em'l^{j}a/ /z'eml'á/ 'terre'
ZEMLJA

b. зола /zo'la/ /zolá/ >'cendres'
ZOLA

c. боль /bol^{j}/ /bol'/ 'douleur'
BOL'

d. пол /pol/ /pol/ 'sol'
POL

1.2.3 Position de l'accent

En russe, chaque mot-forme comporte une syllabe accentuée. L'accentuation en russe est caractérisée par deux propriétés principales : son impact sur la prononciation des mots, notamment des voyelles, et sa position variable en ce qui concerne les formes fléchies d'un même lexème.

L'alphabet cyrillique constitue un bon guide de prononciation car il est généralement clair comment une séquence de lettres doit être prononcée. Cependant, le cas des voyelles pose des problèmes : les mots russes contiennent généralement une syllabe accentuée, dont la voyelle est prononcée plus nettement et plus clairement que les autres voyelles non accentuées, qui peuvent être plus ou moins réduites. Ainsi, une fois que le locuteur sait où se trouve l'accent dans un mot donné, il peut le prononcer correctement.

Du point de vue phonétique, une syllabe accentuée est caractérisée par une plus grande tension des organes d'articulation, un timbre plus élevé, une durée plus longue et, par conséquent, une prononciation plus nette (Švedova, 1980, pp.90-92 ; Garde, 1998, p.27). Les voyelles de la syllabe préaccentuée ont une durée intermédiaire, elles

sont plus courtes que la voyelle accentuée mais plus longues que les autres voyelles. Les voyelles qui sont plus éloignées de la syllabe accentuée sont très courtes (Timberlake, 2004, pp.29-30,132), voire même réduites. Cela concerne surtout <*o*> (14a) et <*e*> 14b).

(14) a. *волос* [ˈvoɬəs] ‘cheveu’
volos

b. *ветер* [ˈvʲetʲɪr] ‘vent’
veter[6]

Dans les exemples cités précédemment, les voyelles des syllabes accentuées sont plus longues, plus intenses et plus hautes que les autres ; <*o*> et <*e*> non accentuées sont réduites à /ə/ et /ɪ/ respectivement.

De manière générale, un mot-forme de lexème (nom, adjectif, verbe, adverbe) comporte une seule syllabe accentuée, cependant, deux cas de figure particuliers peuvent être cités. Premièrement, il existe de nombreux phénomènes de variabilité de l’accent pour un seul lexème. Dans ce cas, une des options de prononciation peut être normative, tandis que l’autre peut violer la norme et être typique du langage parlé ou familier (15a). Deuxièmement, quelques composés peuvent avoir un accent secondaire. Dans ce cas, l’accent principale est placé vers la fin du mot, tandis que l’accent secondaire se trouve généralement plus au début (15b) (Švedova, 1980 ; Roon, 2006). Tout comme l’accentuation principale, l’accentuation secondaire se caractérise par une absence de réduction des voyelles (Karpacheva, 2000).

(15) a. ТВОРОГ ‘fromage blanc’
TVOROG
/tvoˈrog/ /tvoróg/ (standard)
/ˈtvorog/ /tvórog/ (familier)

b. ДАЛЬНЕВОСТОЧНЫЙ ‘extrême-orient$_{\text{REL}}$’
DAL’NEVOSTOČNYJ
/ˌdalʲnʲevosˈtot͡ʃnɨj/ /dal'n'evostóčnoj/
САМОЛЁТОСТРОЕНИЕ ‘construction d’avions’
SAMOLËTOSTROENIE
/samoˌlʲotostroˈjenʲije/ /samol'otostrojén'ije/

Le système phonologique du russe se caractérise également par la position libre de l’accent, qui n’est pas automatiquement assigné à une syllabe précise d’un mot-forme (première, dernière, avant-dernière, etc.). Il n’y a généralement pas de règle phonologique pour prévoir la place de l’accent dans un mot, sauf dans les cas où la

[6]La transcription phonétique est fournie ici par soucis d’illustration. Cependant, comme cela a été mentionné dans les Conventions typographiques, la transcription phonologique suffit pour décrire la majorité des phénomènes en russe, y compris les phénomènes auxquels nous nous intéressons dans la présente recherche. Ainsi, la transcription phonétique ne sera plus utilisée.

structure morphologique peut la déterminer (Garde, 1998, p.119). Dans l'exemple (16a), l'accent est sur /*i*/ comme dans la plupart des mots avec le suffixe suffixe *-itel'*. Dans la majorité des cas, cependant, la position de l'accent est arbitraire pour la même distribution des syllabes et des voyelles (16b et 16e). De manière similaire, l'accent n'est pas attaché à un élément caractérisable du point de vue morphologique : il peut marquer le thème ou les affixes. Par exemple, l'accent peut être placé sur le thème (16b), le suffixe dérivationnel (16c), le préfixe (16d) ou sur le suffixe flexionnel (16e).

(16) a. УЧИТЕЛЬ /uˈt͡ʃitʲelʲ/ /učít'el'/ 'professeur'
UČITEL'

b. ПИВО /ˈpʲivo/ /p'ívo/ 'bierre'
PIVO

c. ФУТБОЛИСТ /futboˈlʲist/ /futbol'íst/ 'footballeur'
FUTBOLIST

d. ВЫХОД /ˈvɨxod/ /víxod/ 'sortie'
VYXOD

e. ВИНО 'vin' : *вин-о* /vʲiˈno/ /v'inó/ 'vin-N.NOM.SG'
VINO *vin-o*

L'accentuation en russe est variable, de ce fait c'est une propriété individuelle de chaque mot, et cette propriété peut être distinctive au niveau lexical (17a-17b). Cependant, l'accent ne marque pas nécessairement la même syllabe pour tous les mots-formes des lexèmes. Lors de la réalisation de formes fléchies, l'accent peut se déplacer d'une syllabe à l'autre. Ainsi, l'accent en russe est mobile et peut servir à distinguer les formes d'un même lexème (17c), dans ce cas, il devient un marqueur morphologique (Timberlake, 2004, p.132). Dans d'autres cas, ce sont les formes fléchies de lexèmes différents qui peuvent être distinguées grâce à la position de l'accent (17d).

(17) a. ЗАМОК$_1$ /ˈzamok/ /zámok/ 'château'
ZAMOK
ЗАМОК$_2$ /zaˈmok/ /zamók/ 'cadenas'
ZAMOK

b. МУКА$_1$ /ˈmuka/ /múka/ 'supplice'
MUKA
МУКА$_2$ /muˈka/ /muká/ 'farine'
MUKA

c. ГОРОД 'ville' :
GOROD
город-а /ˈgoroda/ /góroda/ 'ville-M.GEN.SG'
gorod-a
город-а /goroˈda/ /gorodá/ 'ville-M.NOM.PL'
gorod-a

d. БЕЛКА ‘écureuil’ :
BELKA
белк-а /ˈbʲelka/ /bˈélka/ ‘écureuil-F.NOM.SG’
belk-a
БЕЛОК ‘blanc d’œuf’ :
BELOK
белк-а /bʲelˈka/ /bˈelká/ ‘blanc d’œuf-M.GEN.SG’
belk-a

La classification des lexèmes qui ont des formes fléchies en russe est effectuée en fonction des types accentuels. Le type accentuel correspond à un schéma d’accentuation pour les mots-formes donnés d’une classe lexicale. Comme mentionné précédemment, les formes fléchies peuvent avoir un accent sur le thème (accent non final) ou sur la flexion (accent final). Pour décrire le système accentuel russe, Švedova (1980, pp.94-95,509-528), par exemple, propose quatre types accentuels qu’elle désigne par les lettres de A à D (avec quelques variations et sous-types). Cette classification prend en compte la position de l’accent dans l’ensemble du paradigme des formes fléchies. On y retrouve les types accentuels avec des accents finaux, les types avec des accents non-finaux, ainsi que des combinaisons différentes d’accents.

Une classification alternative complémentaire est proposée par Zaliznjak (2003, pp.25-76) qui distingue 6 classes accentuelles, numérotées de *a* à *f*[7]. Cette classification est généralement similaire à celle des types et sous-types accentuels proposée par Švedova (1980).

Dans le présent travail, nous utiliserons deux approches à la position de l’accent : une basée sur la forme de citation des lexèmes (NOM.SG.) et la position de l’accent en fonction des syllabes (positon finale, initiale, etc.) ; une autre approche est celle de Zaliznjak (2003) qui analyse les propriétés accentuelles de tout le paradigme des noms de base. Cette approche distingue six classes accentuelles pour les noms en russe[8], qui sont les suivantes :

a. Noms dont l’accent est toujours sur le thème :
КАРТА ‘carte’ : $к\boldsymbol{а}рт\text{-}а_{\text{nom.sg.}}$, $к\boldsymbol{а}рт\text{-}ъ\acute{\imath}_{\text{gen.sg.}}$, $к\boldsymbol{а}рт\text{-}ы_{\text{nom.pl.}}$, $карт_{\text{gen.pl.}}$,
KARTA
СПОР ‘dispute’ : $спор_{\text{nom.sg.}}$, $сп\boldsymbol{о}р\text{-}а_{\text{gen.sg.}}$, $сп\boldsymbol{о}р\text{-}ы_{\text{nom.pl.}}$, $сп\boldsymbol{о}р\text{-}ов_{\text{gen.pl.}}$;
SPOR

b. Noms dont l’accent est toujours sur le suffixe flexionnel, sauf quand ce dernier est absent :

[7]La classification proposée par Zaliznjak (2003) est valable pour les noms et pour les adjectifs. Cependant, dans le présent travail, nous nous intéressons uniquement aux classes accentuelles nominales. Dans ce qui suit, nous allons faire référence aux noms pour la présentation de cette classification.

[8]Une classe supplémentaire 0 peut y être rajoutée, pour faire référence à des noms indéclinables.

ОЧКО ‘point’ : *очк-**о***$_{\text{nom.sg.}}$, *очк-**а***$_{\text{gen.sg.}}$, *очк-**и***$_{\text{nom.pl.}}$, *очк-**ов***$_{\text{gen.pl.}}$,
OČKO
СТОЛ ‘table’ : *стол*$_{\text{nom.sg.}}$, *стол-**а***$_{\text{gen.sg.}}$, *стол-**ы***$_{\text{nom.pl.}}$, *стол-**ов***$_{\text{gen.pl.}}$;
STOL

c. Noms dont l’accent est sur le thème au singulier et sur le suffixe flexionnel au pluriel :
МОРЕ ‘mer’ : *мо**р**-е*$_{\text{nom.sg.}}$, *мо**р**-я*$_{\text{gen.sg.}}$, *мор-**я***$_{\text{nom.pl.}}$, *мор-**ей***$_{\text{gen.pl.}}$,
MORE
САД ‘jardin’ : *сад*$_{\text{nom.sg.}}$, *с**а**д-а*$_{\text{gen.sg.}}$, *сад-**ы***$_{\text{nom.pl.}}$, *сад-**ов***$_{\text{gen.pl.}}$;
SAD

d. Noms dont l’accent est sur le suffixe flexionnel au singulier et sur le thème au pluriel :
ВИНО ‘vin’ : *вин-**о***$_{\text{nom.sg.}}$, *вин-**а***$_{\text{gen.sg.}}$, *в**и**н-а*$_{\text{nom.pl.}}$, *вин*$_{\text{gen.pl.}}$,
VINO
ЛИСТ ‘feuille’ : *лист*$_{\text{nom.sg.}}$, *лист-**а***$_{\text{gen.sg.}}$, *л**и**стъ-я*$_{\text{nom.pl.}}$, *л**и**стъ-ев*$_{\text{gen.pl.}}$;
LIST

e. Noms dont l’accent est sur le thème au singulier et au nominatif pluriel, mais sur le suffixe flexionnel pour les autres cas du pluriel :
ЗУБ ‘dent’ : *зуб*$_{\text{nom.sg.}}$, *з**у**б-а*$_{\text{gen.sg.}}$, *з**у**б-ы*$_{\text{nom.pl.}}$, *зуб-**ов***$_{\text{gen.pl.}}$,
ZUB
ВЕЩ-Ь ‘chose’ : *вещь*$_{\text{nom.sg.}}$, *в**е**щ-и*$_{\text{gen.sg.}}$, *в**е**щ-и*$_{\text{nom.pl.}}$, *вещ-**ей***$_{\text{gen.pl.}}$;
VEŠČ

f. Noms dont l’accent est sur le thème au nominatif pluriel et sur le suffixe flexionnel à tous les autres cas :
ГУБА ‘lèvre’ : *губ-**а***$_{\text{nom.sg.}}$, *губ-**ы***$_{\text{gen.sg.}}$, *г**у**б-ы*$_{\text{nom.pl.}}$, *губ*$_{\text{gen.pl.}}$,
GUBA
КОНЬ ‘cheval’ : *конь*$_{\text{nom.sg.}}$, *кон-**я***$_{\text{gen.sg.}}$, *к**о**н-и*$_{\text{nom.pl.}}$, *кон-**ей***$_{\text{gen.pl.}}$.
KON’

1.3 Le système morphologique

Les langues slaves sont caractérisées par un système morphologique complexe, notamment en ce qui concerne la morphologie flexionnelle. Ainsi, elles sont considérées comme des langues conservatrices de la branche des langues indo-européennes (Corbett et Comrie, 2003, pp.6-7).

Ces langues présentent également un système développé d’accords, notamment entre le nom et l’adjectif, ainsi qu’entre le nom et le verbe. Le russe ne fait pas exception à cette règle, et est caractérisé par un système morphologique complexe et l’utilisation extensive de la flexion. Chaque forme de substantif et d’adjectif est marquée par les

traits morphosyntaxiques de genre, nombre et cas. Pour marquer le nombre et le cas, le russe utilise des systèmes de suffixation et d'accentuation. Le genre, le nombre et le cas sont contextuels pour les adjectifs, car ces derniers s'accordent avec les noms qu'ils modifient. En plus des catégories d'accord, il y a des catégories propres de l'adjectif : l'opposition forme longue/courte et les degrés de comparaison. Pour les substantifs, le genre est intrinsèque, le nombre est inhérent et le cas est contextuel.

1.3.1 Le paradigme morphologique des noms

Les noms russes peuvent être fléchis selon six cas différents : nominatif, génitif, datif, accusatif, instrumental et locatif, ainsi que selon deux nombres : singulier et pluriel. Les noms peuvent également être marqués par l'un des trois genres : masculin, féminin ou neutre. Le paradigme morphologique des noms est donc composé d'au moins douze formes qui résultent de la combinaison des traits de nombre et de cas.

1.3.1.1 Nombre

Il est important de noter que tous les noms en russe ne varient pas seulement entre singulier et pluriel (18a). Certains noms n'ont qu'une forme au singulier, *singularia tantum* (18b), tandis que d'autres n'ont qu'une forme au pluriel, *pluralia tantum*, (18c).

(18) a. ХВОСТ 'queue' : *хвост* 'queue$_{\text{M.NOM.SG}}$' ; *хвост-ы* 'queue-M.NOM.PL'
XVOST *xvost* *xvost-y*

b. ЛИСТВА 'feuillage$_{\text{SG}}$'
LISTVA

c. НОЖНИЦЫ 'ciseaux$_{\text{PL}}$'
NOŽNICY

Le pluriel en russe est généralement marqué par les suffixes flexionnels -*i* ou -*y*.

Les noms concrets, c'est-à-dire les noms désignant des êtres animés ou des objets inanimés, constituent un sous-ensemble majoritaire des noms auxquels l'opposition singulier/pluriel est applicable de manière régulière. Les noms qui n'ont qu'une forme au singulier ou qu'une forme au pluriel font partie de groupes sémantiques distincts. Ainsi, les noms qui n'ont qu'une forme au singulier sont des noms massifs, qui peuvent être mesurés mais non comptés (19a) ; des noms collectifs dérivés d'adjectifs ou de noms (19b) ; et des noms abstraits dérivés de verbes ou d'adjectifs (19c). Les noms qui n'ont qu'une forme au pluriel désignent des objets qui se composent de deux parties (19d) ; des substances et des matières (19e) ; des aliments (19f) ; des actions et des processus (19g) ; et des périodes (19h) (Švedova, 1980, pp.469-473).

(19) a. ЗОЛОТО 'or$_{\text{SG}}$'
ZOLOTO

b. БЕДНОТА 'les pauvres$_{\text{SG}}$'
BEDNOTA

c. СИНЕВА ‘bleu$_{\text{N.SG}}$’
SINEVA

d. ОЧКИ ‘lunettes$_{\text{PL}}$’
OČKI

e. ДУХИ ‘parfum$_{\text{PL}}$’
DUXI

f. СЛИВКИ ‘crème$_{\text{PL}}$’
SLIVKI

g. РОДЫ ‘accouchement$_{\text{PL}}$’
RODY

h. СУМЕРКИ ‘crépuscule$_{\text{PL}}$’
SUMERKI

1.3.1.2 Genre

La distinction entre les genres masculin, féminin et neutre n’est pas le seul moyen d’aborder le trait de genre en russe (Švedova, 1980, pp.463-469). Certaines études incluent aussi dans la discussion le trait d’animé/inanimé, également appelé sous-genre (Garde, 1998, pp.141-143 ; Corbett, 1991, pp.165-167), ce qui entraîne l’identification de six genres en russe au lieu de trois. Cependant, comme le montrent Fraser et Corbett (1995, pp.130-132), la notion d’animé/inanimé est plus spécifique que celle de genre : elle ne concerne que certaines cases du paradigme nominal (l’accusatif des noms masculins) et elle peut être facilement déduite à partir de la sémantique des noms (les noms animés désignant les humains, les animaux et les insectes). Le trait d’animé/inanimé peut donc être considéré comme une catégorie sémantique des noms[9].

Les trois genres – masculin, féminin et neutre (20a-20c) – sont distingués pour les formes nominales au singulier. Ainsi, Timberlake (2004, p.92) évoque quatre classes situées à l’intersection des traits de nombre et de genre : trois genres au singulier et le pluriel. Le masculin et le féminin sont sémantiquement motivés pour les êtres animés, ce qui n’est pas le cas pour les noms inanimés pour lesquels le masculin et le féminin sont arbitraires. Le neutre caractérise majoritairement les noms inanimés. Il existe un sous-ensemble de noms de genre commun (20d) qui désignent un groupe de personnes selon leurs activités. Ces noms de genre commun nécessitent un accord masculin ou féminin, en fonction de leur référence.

(20) a. РОМАН ‘roman$_{\text{M}}$’
ROMAN

b. ПОГОДА ‘temps$_{\text{F}}$’
POGODA

c. ЯБЛОКО ‘pomme$_{\text{N}}$’
JABLOKO

[9] L’opposition animé/inanimé sera abordé plus en détail dans la section 5.2.4.

d. ЗУБРИЛА ‘bûcheur$_{M/F}$’
ZUBRILA

Dans la grande majorité des cas, il est possible de déduire le genre des noms à partir de leur forme au nominatif singulier. Le féminin est marqué par les suffixes flexionnels *-a* ou *-ja* ; le masculin est marqué par l'absence de suffixes flexionnels. Le neutre, à son tour, est généralement marqué par les suffixes flexionnels *-o* ou *-e*. La confusion apparaît lorsque le thème d'un nom comporte une consonne molle en position finale : dans ce cas, il n'y a pas de moyen formel pour déterminer s'il s'agit du masculin ou du féminin en observant juste la forme de citation (nominatif singulier) du lexème (21)[10].

(21) ТЕНЬ ‘ombre$_F$’
TEN’
ДЕНЬ ‘jour$_M$’
DEN’

1.3.1.3 Cas

Les six cas mentionnés précédemment sont généralement considérés comme les cas principaux pour les noms en russe. Cependant, il existe des classifications plus fines qui incluent des cas secondaires. Parmi eux, ceux qui sont le plus souvent mentionnés sont le deuxième cas génitif (également appelé génitif partitif), le deuxième cas locatif, le vocatif et l'adnumératif (Zaliznjak, 2002, pp.36-52).

La variation morphologique en cas correspond à la notion de classe de flexion : une classe de flexion regroupe les noms dont les cases du paradigme sont remplies selon les mêmes règles morphologiques. En russe, les classes de flexion se distinguent formellement davantage au singulier qu'au pluriel.

Les classes flexionnelles des noms en russe sont liées au trait de genre, cependant, le genre ne les définit pas de manière systématique. Comme pour le trait de genre, l'attribution des noms russes aux classes flexionnelles n'est pas une tâche triviale, car ceux-ci présentent de nombreuses irrégularités. Il n'y a pas de consensus dans la slavistique quant au nombre de déclinaisons qui existent en russe. Traditionnellement, on distingue trois classes de flexion (Švedova, 1980, pp.481-482) : la première, la deuxième et la troisième. La première classe flexionnelle englobe les noms masculins dépourvus de flexion au nominatif singulier et les noms neutres avec les suffixes flexionnels qui correspondent aux /*o*/ et /*e*/ au nominatif singulier (22a). La deuxième classe flexionnelle contient les noms masculins et féminins avec le suffixe flexionnel /*a*/, ainsi que les noms du genre commun (22b). La troisième classe flexionnelle concerne les noms féminins sans suffixe flexionnel au nominatif singulier et dont le thème se termine par une consonne molle ou par une chuintante (22c). Graphiquement, les noms de la troisième classe de flexion contiennent un signe mou final même après les chuintantes.

[10]Cependant, les noms masculins et féminins dont le nominatif singulier se termine par un signe mou ont des paradigmes distincts, par exemple *тен-и* (*ten-i*) ‘ombre-F.GEN.SG’ mais *дн-я* (*dn-ja*) ‘jour-M.GEN.SG’.

(22) a. СТОЛ ‘table$_{\text{I.M}}$’
STOL
ОКНО ‘fenêtre$_{\text{I.N}}$’
OKNO
РУЖЬЁ ‘fusil$_{\text{I.N}}$’
RUŽ’Ë
ПОЛЕ ‘champs$_{\text{I.N}}$’
POLE

b. ТРУБА ‘tuyau$_{\text{II.F}}$’
TRUBA
ЗЕМЛЯ ‘terre$_{\text{II.F}}$’
ZEMLJA
МУЖЧИНА ‘homme$_{\text{II.M}}$’
MUŽČINA
СУДЬЯ ‘juge$_{\text{II.M}}$’
SUD’JA
СИРОТА ‘orphelin$_{\text{II.M/F}}$’
SIROTA

c. КРОВАТЬ ‘lit$_{\text{III.F}}$’
KROVAT’
МЫШЬ ‘souris$_{\text{III.F}}$’
MYŠ’

Les classifications plus fines (Timberlake, 2004, pp.103-131 ; Fraser et Corbett, 1995, p.128) subdivisent les classes flexionnelles I et II en fonction des genres (par exemple, la classe Ia contient uniquement les noms masculins et la classe Ib, uniquement les noms neutres).

Les noms déclinables ont des formes distinctes qui correspondent aux différents cas, tandis que les noms indéclinables ont des formes homonymes dans tous les cases de leur paradigme. En général, il s’agit de mots d’origine étrangère, animés et inanimés, ainsi que de toponymes et de noms de famille, tant russes qu’étrangers (23) (Švedova, 1980, pp.504-505).

(23) ДЕПО ‘dépôt’
DEPO
Баку ‘Bakou’
Baku
Гёте ‘Goethe’
Gёte
Живаго ‘Jivago’
Živago

Du point de vue graphique, les classes flexionnelles I et II présentent deux variations : les noms dont les thèmes se terminent par une consonne dure, une chuintante ou un

/c/ (des thèmes durs) et les noms dont les thèmes ont une consonne molle ou un /j/ en position finale (thèmes mous) (Švedova, 1980, pp.481-482). Cette variation, ainsi que la partition des noms en déclinables et indéclinables et les particularités du graphème finale du thème, servent de base à la typologie des classes flexionnelles proposée par Zaliznjak (2003, pp.25-76).

Zaliznjak distingue neuf classes de flexion (numérotées de 0 à 8) des noms en fonction de leur graphème finale[11]. Cette classification est basée sur la forme graphique : si un nom se termine par une voyelle ou un signe mou, ces lettres sont enlevées pour accéder au thème graphique (*акул-а* (*akul-a*) 'requin', *тополь* (*topol-'*) 'peuplier') ; si le mot se termine par une consonne, la forme graphique du thème correspond au nominatif singulier (*поезд* (*poezd*) 'train'). Les classes flexionnelles sont ainsi les suivantes :

0. Noms indéclinables ;

1. Noms dont le thème se termine par une consonne dure :
   ТОПОР 'hache' : *топор*$_{\text{nom.sg.}}$,
   TOPOR
   КОБРА 'cobra' : *кобр-а*$_{\text{nom.sg.}}$,
   KOBRA
   ОЛОВО 'étain' : *олов-о*$_{\text{nom.sg.}}$;
   OLOVO

2. Noms dont le thème se termine par une consonne molle :
   ТЮЛЕНЬ 'phoque' : *тюлен-ь*$_{\text{nom.sg.}}$,
   TJULEN'
   ЦАПЛЯ 'héron' : *цапл-я*$_{\text{nom.sg.}}$,
   CAPLJA
   ПОЛЕ 'champs' : *пол-е*$_{\text{nom.sg.}}$;
   POLE

3. Noms se terminant par une gutturale <*г*>, <*к*> ou <*х*> :
   ПЕТУХ 'coq' *петух*$_{\text{nom.sg.}}$,
   PETUX
   САПОГ 'botte' : *сапог*$_{\text{nom.sg.}}$,
   SAPOG
   ЧАЙНИК 'théière' : *чайник*$_{\text{nom.sg.}}$;
   ČAJNIK

4. Noms se terminant par une chuintante <*ж*> <*ш*> <*щ*> <*ч*> :
   МЯЧ 'ballon' : *мяч*$_{\text{nom.sg.}}$,
   MJAČ

[11] Tout comme les types accentuels, les classes flexionnelles de Zaliznjak (2003) s'appliquent également aux adjectifs. Nous nous concentrerons exclusivement sur les noms dans la présentation de cette classification.

ГАЛОША ‘chaussure en caoutchouc’ : *галош-а*$_{\text{nom.sg.}}$,
GALOŠA
ЖИЛИЩЕ ‘domicile’ : *жилищ-е*$_{\text{nom.sg.}}$,
ŽILIŠČE
ЛУЖА ‘flaque’ : *луж-а*$_{\text{nom.sg.}}$;
LUŽA

5. Noms se terminant par la sifflante <ц> :
   ДЕВИЦА ‘jeune fille’ : *девиц-а*$_{\text{nom.sg.}}$;
   DEVICA

6. Noms se terminant par une voyelle autre que <и> ou par la semi-voyelle <й> :
   БОЙ ‘combat’ : *бой*$_{\text{nom.sg.}}$;
   BOJ

7. Noms dont la base se termine par <и> :
   СЛОЖЕНИЕ ‘addition’ : *сложени-е*$_{\text{nom.sg.}}$,
   SLOŽENIE
   МАНИЯ ‘obsession’ : *мани-я*$_{\text{nom.sg.}}$;
   MANIJA

8. Noms qui font traditionnellement partie de la troisième déclinaison :
   ЛЮБОВЬ ‘amour’ : *любов-ь*$_{\text{nom.sg.}}$,
   LJUBOV’
   ТЕТРАДЬ ‘cahier’ : *тетрад-ь*$_{\text{nom.sg.}}$.
   TETRAD’

Dans cette recherche, nous nous baserons à la fois sur la distinctions entre les trois classes flexionnelles canoniques et sur la classification proposée par Zaliznjak (2003). Cette dernière permet de prendre en compte les variations principales des paradigmes flexionnels des noms russes sans entrer dans des détails trop fins[12].

1.3.2 Le paradigme morphologique des adjectifs

Au sein d’un syntagme nominal, l’accord entre le nom et l’adjectif est obligatoire en genre, nombre et cas. Les adjectifs sont accordés avec les noms en fonction de ces trois traits morpho-syntaxiques, ce qui les amène à avoir au moins 36 formes fléchies (Švedova, 1980, pp.545-554). Le taux de syncrétisme formel des adjectifs est très élevé (24) :

[12]Il est cependant possible de combiner les huit classes flexionnelles avec les types accentuels proposés également par Zaliznjak et discutés dans la section 1.2.3 pour arriver à près de 50 classes de flexion. Nous traitons séparément l’accentuation et la déclinaison des noms. Le nombre de classes de flexion peut augmenter davantage si les différentes allomorphies thématiques sont incluses dans l’analyse (voir Parker et Sims (2020) qui identifient 89 classes flexionnelles ou Guzmán Naranjo (2020) qui en identifie 108).

(24) НОВЫЙ ‘nouveau’ :
NOVYJ

нов-ой ‘nouveau-F.GEN.SG’ ;	*нов-ой* ‘nouveau-F.DAT.SG’ ;
nov-oj	*nov-oj*
нов-ой ‘nouveau-F.INS.SG’ ;	*нов-ой* ‘nouveau-F.LOC.SG’
nov-oj	*nov-oj*

Il existe également un groupe d’adjectifs qui peuvent avoir des formes longues et courtes dans leur paradigme. Un autre trait pertinent qui n’est pas partagé avec les noms est celui des degrés de comparaison (certains adjectifs russes ont une forme de comparatif et de superlatif). Les deux traits – degrés de comparaison et la distinction entre formes longues et courtes – sont liés à la classification des adjectifs en qualificatifs et relationnels ; ces sujets seront traités dans les sections 2.1.1 et 2.3.1. Dans cette section, nous nous concentrerons sur les catégories morpho-syntaxiques des adjectifs.

Comme mentionné précédemment, l’adjectif russe s’accorde avec le nom en genre. Cependant, il y a quelques cas où l’adjectif constitue le seul marqueur de genre : lorsqu’il modifie les noms de la classe flexionnelle 0 (25a) et les noms du genre commun (25b).

(25) a.

известн-ый	*конферансье*	‘un conférencier célèbre’
izvestn-yj	*konferanc’e*	
célèbre-M.NOM.SG	‘conférencier$_{\text{M.NOM.SG}}$’	
американск-ая	*визави*	‘une vis-à-vis américaine’
amerikansk-aja	*vizavi*	
américain-F.NOM.SG	vis-à-vis$_{\text{F.NOM.SG}}$	

b.

больш-ой	*невеж-а*	‘un grand ignorant’
bol’š-oj	*nevež-a*	
grand-M.NOM.SG	ignorant-M.NOM.SG	
нов-ая	*коллег-а*	‘une nouvelle collègue’
nov-aja	*kolleg-a*	
nouveau-F.NOM.SG	collègue-F.NOM.SG	

Dans le cas des noms appartenant à la classe flexionnelle 0, les adjectifs représentent également le seul marqueur de nombre (26a) et de cas (26b).

(26) a.

розов-ый	*фламинго*	‘un flamand rose’
rozov-yj	*flamingo*	
rose-M.NOM.SG	‘flamand$_{\text{M.NOM.SG}}$’	
розов-ые	*фламинго*	‘des flamands roses’
rozov-ye	*flamingo*	
rose-M.NOM.PL	flamand$_{\text{M.NOM.PL}}$’	

b.

[я	*купи-л]*	*нов-ое*	*пальто*
[ja	*kupi-l]*	*nov-oe*	*pal’to*
1.NOM.SG	acheter-PST.M.SG	nouveau-N.ACC.SG	manteau$_{\text{N.ACC.SG}}$

‘[j’ai acheté un] nouveau manteau’
[u *menja* *net]* *nov-ogo* *pal’to*
PREP 1.GEN.SG avoir$_{\text{PRS.NEG}}$ nouveau-N.GEN.SG manteau$_{\text{N.GEN.SG}}$
‘[je n’ai pas de] nouveau manteau’
[ja *rasskaž-u* *o]* *nov-om* *pal’to*
1.NOM.SG parler-FUT.1.SG PREP nouveau-N.LOC.SG manteau$_{\text{N.LOC.SG}}$
‘[je parlerai du] nouveau manteau’

Les degrés de comparaison sont pertinents pour le système adjectival russe (27). Les formes de comparatif sont dépourvues des traits de genre, nombre et cas, et sont ainsi indéclinables.

(27) ГРУБЫЙ ‘rude’ : *груб-ее* ‘rude-COMP’
GRUBYJ *grub-ee*
ЛЕВЫЙ ‘gauche’ : *лев-ее* ‘gauche-COMP’
LEVYJ *lev-ee*

Cependant, tous les adjectifs ne sont pas concernés par la formation des degrés de comparaison, c’est le cas notamment des adjectifs de relation. Parmi les adjectifs qualificatifs, il existe un groupe d’adjectifs qui ne forment pas non plus de comparatif, en fonction de leurs sens (Švedova, 1980, pp.560-563) : ce sont, par exemple, les adjectifs qui désignent les propriétés non mesurables (28a), les adjectifs de couleurs construits (28b).

(28) a. ГЛУХОЙ ‘sourd’ : **глуш-е* ‘sourd-COMP’
GLUXOJ *gluš-e*
b. БОРДОВЫЙ ‘bordeaux’ : **бордов-ее* ‘bordeaux-COMP’
BORDOVYJ *bordov-ee*

Les adjectifs qualificatifs sont caractérisés également par la présence de formes longues et de formes courtes dans leur paradigme. Ces formes peuvent être fléchies uniquement en genre et nombre, elles ne sont pas concernées par le cas. Du point de vue syntaxique, les formes courtes remplissent une fonction d’attribut du sujet (29b), tandis que les formes longues sont employées comme épithètes (29a).

(29) a. *весёл-ый* *ребёнок* ‘un enfant joyeux’
vesël-yj *rebënok*
joyeux-M.NOM.SG enfant$_{\text{M.NOM.SG}}$
b. *ребёнок* *весел* ‘l’enfant est joyeux’
rebënok *vesel*
enfant$_{\text{M.NOM.SG}}$ joyeux$_{\text{M.NOM.SG}}$ (court)

1.4 Allomorphies thématiques

Lors de la formation des mots-formes ou des dérivés, on peut observer des alternances au niveau des thèmes ainsi qu'au niveau des affixes. Nous nous concentrerons sur les allomorphies thématiques dans cette section et étudierons les modifications qui peuvent se produire au niveau formel des thèmes pour construire un radicale adapté à la dérivation (Roché, 2010). Les allomorphies suffixales seront traitées dans la section 2.2.2.

En russe, toutes les combinaisons de sons ne sont pas autorisées entre le thème et les affixes – flexionnels ou dérivationnels. Plusieurs types de phénomènes peuvent se produire. Les transformations non-linéaires se manifestent par des alternances de phonèmes (finaux – concernant la dernière consonne du thème, et non-finaux – alternances vocaliques au milieu du thème). Ce sont les transformations non-linéaires qui reflètent les changements phonologiques en diachronie. Les transformations linéaires consistent en un ajout de matériel phonologique supplémentaire entre le thème et l'affixe, suppression de matériel phonologique superflu, superposition ou fusion des segments (si le dernier phonème ou séquence de phonèmes du thème est identique à celui/ceux au début du suffixe) (Zemskaja, 2011, pp.80-81 ; Lopatin, 1977, pp.114-124 ; Švedova, 1980, pp.410-442).

Tous ces phénomènes impliquent la présence dans l'espace thématique des lexèmes de base de plus d'un seul thème.

1.4.1 Espaces thématiques

Le thème représente une séquence commune entre tous les mots-formes d'un même lexème. Dans le cas le plus simple, les adjectifs russes sont formés par une simple adjonction d'un suffixe (*-n-*, *-sk-* ou *-Ov-*, par exemple, dans le cadre de cette étude) au thème du nom de base. Cependant, la correspondance formelle entre le thème de base et le radical de l'adjectif (30a) n'est pas toujours respectée dans certains cas particuliers (30b-30f).

(30) a. СТОЛ 'table' → СТОЛОВЫЙ
STOL STOLOVYJ

b. РУ**К**А 'main' → РУ**Ч**НОЙ
RUKA RUČNOJ

c. Д**Е**НЬ 'jour' → ДНЕВНОЙ
DEN' DNEVNOJ

d. КОНЬ /konʲ/ /kon'/ 'cheval' → КОНСКИЙ /'konskʲij/ /kónskoj/
KON' KONSKIJ

e. БУРЖУА 'bourgeois' → БУРЖУА**З**НЫЙ
BURŽUA BURŽUAZNYJ

f. Чикаго‘Chicago’ → чикагский
ČIKAGO ČIKAGSKIJ

Il n’y a pas de consensus entre les slavistes sur le statut de ces thèmes. D’après Arutjunova (1961, p.73), il est nécessaire de distinguer les thèmes de base pertinents pour la flexion de ceux utilisés lors de la dérivation. Certains chercheurs considèrent qu’il s’agit de variants d’un seul et même thème (Vinokur, 1959, pp.419-442), ou des allomorphes (Zemskaja, 2011, pp.21,22). D’autres estiment qu’il s’agit de thèmes différents, un thème flexionnel et un thème dérivationnel[13] (Švedova, 1980). Lopatin (1977, p.106-111) distingue un thème au sens restreint (utilisé dans les mots-formes) et un thème au sens large (utilisés à la fois en fléxion et en dérivation) et établit une hiérarchie des thèmes. Selon lui, lors de la dérivation, un thème flexionnel est utilisé comme thème primaire, et des modifications formelles sont effectuées pour adapter ce thème à l’affixe et produire un thème dérivationnel. Dans le cas de la dérivation des adjectifs à partir des noms, c’est le thème correspondant au nominatif singulier qui est considéré comme le thème primaire par Lopatin.

En linguistique française, il est généralement admis qu’un lexème peut contenir plusieurs thèmes et que le radical des lexèmes construits ne leur est souvent pas identique. Le sujet des allomorphies thématiques a été largement étudié, tant en morphologie flexionnelle (Bonami et Boyé, 2003 ; Bonami et Boyé, 2005) qu’en morphologie dérivationnelle (Boyé et Plénat, 2015 ; Bonami *et al.*, 2009 ; Roché, 2010). La distinction entre le thème et le radical est effectuée notamment par Roché (2010), le thème faisant partie de l’information lexicale, le radical étant construit lors de la dérivation. C’est cette distinction que nous adoptons dans notre travail. Ainsi, dans l’exemple (30a), le radical de l’adjectif est identique au thème, l’espace thématique ne contient alors qu’un seul thème. Dans tous les autres cas (30b-30f), différentes opérations formelles ont été mises en place pour obtenir la forme du radical, ce qui sous-entend que l’espace thématique des lexèmes en question contient plus d’un seul thème.

Il est à noter que toutes les modifications observées dans la dérivation en (30b-30f) ne sont pas présentes dans la flexion des noms de base (31a-31e), à l’exception de день ‘jour’ en (30c), où la suppression de la voyelle se produit également dans les formes fléchies de ce lexème (31b).

(31) a. рука ‘main’ :
RUKA

рук-а ‘main-F.NOM.SG’ ;	*рук-и* ‘main-F.NOM.PL’ ;	*рук* ‘main$_{\text{F.GEN.PL}}$’
ruk-a	*ruk-i*	*ruk*

b. день ‘jour’ :
DEN’

[13] *Основа мотивирующего слова* ‘thème de base motivante’ et *мотивирующая основа* ‘thème motivant’ dans la terminologie de Švedova (1980).

день 'jour$_{\text{M.NOM.SG}}$' ; *дн-и* 'jour-M.NOM.PL' ; *дн-ей* 'jour-M.GEN.PL'
den' *dn-i* *dn-ej*

c. КОНЬ 'cheval' :
KON'
конь 'cheval$_{\text{M.NOM.SG}}$' *кон-и* 'cheval-M.NOM.PL' *кон-ей* 'cheval-M.GEN.PL'
kon' *kon-i* *kon-ej*

d. БУРЖУА 'bourgeois' :
BURŽUA
буржуа 'bourgeois$_{\text{M.NOM.SG}}$' ; *буржуа* 'bourgeois$_{\text{M.GEN.PL}}$'
buržua *buržua*

e. ЧИКАГО 'Chicago' :
ČIKAGO
Чикаго 'Chicago$_{\text{M.NOM.SG}}$' ; *Чикаго* 'Chicago$_{\text{M.GEN.SG}}$'
Čikago *Čikago*

Les thèmes et les radicaux des exemples de lexèmes en (30) et (31) sont présentés dans le tableau 1.2.

Lexème	Thème(s)	Radical(aux)	Opération formelle
СТОЛ	стол	стол	-
РУКА	рук	руч	Palatalisation
ДЕНЬ	день, дн	дн	Voyelle mobile
КОНЬ	конь	кон	Mouillure
БУРЖУА	буржуа	буржуаз	Ajout
ЧИКАГО	чикаго	чикаг	Suppression

Tableau 1.2: Lexèmes, thèmes, radicaux

1.4.2 Alternances non linéaires

Les alternances non-linéaires concernent deux types de modifications pour les thèmes : les modifications non-finales des phonèmes vocaliques (alternance entre une voyelle pleine et une voyelle ∅) et les modifications finales des phonèmes consonantiques (mouillure, palatalisation) selon Lopatin (1977, pp.209-259).

1.4.2.1 Voyelle mobile

Le premier type d'allomorphie thématique que nous allons étudier ici est l'alternance entre une voyelle pleine et voyelle ∅. Ainsi, en russe, il peut y avoir, dans l'espace thématique des lexèmes, un thème avec une voyelle (forme pleine) et un thème sans

voyelle (forme réduite)[14]. Les voyelles qui peuvent disparaître ainsi sont /o/ (35a) et /e/ (35b).

(32) a. МОХ 'mousse' : *мо**о**х* 'mousse$_{\text{M.NOM.SG}}$' ; *мх-а* 'mousse-M.GEN.SG'
MOX *mox* *mx-a*

b. БОБЁР 'castor' : *бобёр* 'castor$_{\text{M.NOM.SG}}$' ; *бобр-а* 'castor-M.GEN.SG'
BOBËR *bobër* *bobr-a*

Cette alternance entre une voyelle pleine et ∅ intervient à l'intérieur du thème, d'où l'appellation alternance non-finale. Ce type d'alternance est le résultat de processus historiques en phonologie. Comme discuté dans la section 1.2.3, les phonèmes vocaliques qui ne sont pas accentués sont plus courts que les phonèmes vocaliques accentués, voire même réduits. C'est une caractéristique importante de la langue russe moderne. Le vieux slave, à son tour, était marqué par la présence de phonèmes vocaliques ultrabrefs. Entre le XIe et le XIIIe siècle, ces phonèmes ont subi deux modifications majeures : la disparition du timbre (ce qui a entraîné l'amuïssement des ultrabrèves) et l'éclaircissement du timbre en valeur vocalique (Breuillard et Viellard, 2015, pp.85-86 ; Timberlake, 2004, p.88). Ces deux phénomènes sont à l'origine de l'alternance voyelle pleine / voyelle ∅. L'alternance voyelle/∅ est présente à la fois dans la flexion (33a) et dans la dérivation (33b).

(33) a. ПОСОЛ 'ambassadeur' :
POSOL
*посо**о**л* 'ambassadeur$_{\text{M.NOM.SG}}$' ; *посл-а* 'ambassadeur-M.GEN.SG'
posol *posl-a*
СОН 'rêve' :
SON
*со**о**н* 'rêve$_{\text{M.NOM.SG}}$' ; *сн-а* 'rêve-M.GEN.SG'
son *sn-a*

b. Д**Е**НЬ 'jour' → ДНЕВНОЙ 'journalier'
DEN' DNEVNOJ

La présence de la voyelle mobile dans les thèmes n'est pas prévisible à partir de la structure phonologique des lexèmes concernés (34).

(34) ЛЕНЬ 'paresse' : *лень* 'paresse$_{\text{F.NOM.SG}}$' ; *лен-и* 'paresse-F.GEN.SG'
LEN' *len'* *len-i*
ПЕНЬ 'souche' : *пень* 'souche$_{\text{M.NOM.SG}}$' ; *пн-я* 'souche-M.GEN.SG'
PEN' *pen'* *pn-ja*

Toutefois, la structure morphologique des noms construits peut fournir des indices sur la présence potentielle d'une voyelle mobile. Par exemple, le suffixe dérivationnel

[14] *Mobile, fugitive, fleeting vowel* (Corbett et Comrie, 2003, p.8 ; Timberlake, 2004, p.88) ou encore *vowel-zero stem alternation* (Sims, 2017, p.491) dans la tradition anglo-saxonne.

-ec (35a) qui forme les gentilés ainsi que le suffixe *-ok* (35b) utilisé dans les diminutifs contiennent les phonèmes /o/ et /e/ qui disparaissent dans certaines cases du paradigme flexionnel.

(35) a. САРАТОВЕЦ 'habitant de Saratov' :
SARATOVEC
*саратов**е**ц* 'habitant de Saratov$_{\text{M.NOM.SG}}$' ;
saratovec
саратовц-а 'habitant de Saratov-M.GEN.SG'
saratovc-a

b. ФЛАЖОК 'petit drapeau' :
FLAŽOK
*флаж**о**к* 'petit drapeau$_{\text{M.NOM.SG}}$' ;
flažok
флажк-а 'petit drapeau-M.GEN.SG'
flažk-a

En ce qui concerne la dérivation, lorsqu'un thème de base présente une voyelle pleine au nominatif singulier, le radical sera marqué par une voyelle ∅ (36a) et vice versa (36b) (Zemskaja, 2011, p.22,112).

(36) a. ДЁГ**О**ТЬ 'goudron' → ДЁГТЕВЫЙ
DËGOT' DËGTEVYJ

b. КАСКА 'casque' → КАС**О**ЧНЫЙ [15]
KASKA KASOČNYJ

Contrairement aux alternances non-finales, les alternances qui surviennent à la fin du thème sont beaucoup plus répandues dans la dérivation et sont étroitement liées au suffixe dérivationnel qui se combine avec le thème en question (Lopatin, 1977, pp.120,225). Il s'agit des alternances de trait de mouillure morphologiquement conditionnées et de la palatalisation.

1.4.2.2 Mouillure

Le trait de mouillure est une propriété fondamentale de certaines consonnes russes, comme discuté dans la section 1.2.2. Cependant, certaines consonnes (notamment les labiales et les dentales) peuvent être réalisées avec une altération du trait de mouillure dans certains contextes morphologiques (37) ; dans ce cas, la mouillure devient une propriété positionnelle.

[15] Il est à noter que dans cet exemple l'alternance voyelle/∅ est accompagnée par la palatalisation /k/ ∼ /č/ ; nous allons aborder ce type d'alternance plus bas dans cette section.

(37) КОНЬ /konʲ/ /kon'/ 'cheval' → КОНСКИЙ /'konskʲij/ /kónskoj/
KON' KONSKIJ
ВОКЗАЛ /vok'zal/ /vokzál/ 'gare' →
VOKZAL
ВОКЗАЛЬНЫЙ /vok'zalʲnɨj/ /vokzál'noj/
VOKZAL'NYJ

L'alternance du trait de mouillure est aussi un phénomène phonologique qui a émergé initialement en raison des changements dans le système vocalique en vieux slave. Les consonnes ont été marquées par le trait de mouillure devant les voyelles ultrabrèves, et après l'amuïssement de ces dernières, les consonnes ont conservé ce trait. Cependant, cette alternance n'est plus conditionnée phonologiquement en russe moderne, cette propriété est déterminée par la position de la consonne dans le mot (Breuillard et Viellard, 2015, p.100 ; Corbett et Comrie, 2003, p.67).

1.4.2.3 Palatalisation

Enfin, la palatalisation est un autre type d'alternance qui se produit à la fin du thème[16]. Il s'agit d'un changement qui transforme un phonème vélaire en une palatale ou dentale affriquée. Ce phénomène est également lié à des processus historiques, et concerne principalement trois phonèmes consonantiques, avec quatre changements possibles :

(38) a. /k/ ∼ /č/
b. /g/ ∼ /ž/
c. /x/ ∼ /š/
d. /k/ ∼ /c/

Comme les alternances abordées plus haut, la palatalisation est un phénomène qui a émergé en vieux slave en raison des changements consonantiques qui se sont produits à l'époque. Ces changements, connus sous le nom de palatalisations slaves, notamment de la première palatalisation slave, se sont terminés vers le Ve siècle. Ils ont impliqué la transformation des consonnes vélaires en consonnes palatales devant les voyelles rétractées /e/ et /i/ (Breuillard et Viellard, 2015, pp.65-67 ; Timberlake, 2004, pp.82-84). Ces transformations avaient un caractère phonologique ; en russe contemporain, cependant, la palatalisation n'est plus conditionnée par le contexte phonologique ; elle apparaît dans certains mots sous certaines conditions morphologiques, c'est-à-dire en présence de certains suffixes flexionnels (39a-39c) du paradigme verbal ou devant les suffixes dérivationnels (39d-39f).

[16]Nous utilisons le terme de palatalisation suivant les slavistes français (Garde, 1998). Dans la tradition anglo-saxonne, ce phénomène est appelé *palatalization*. Cependant, pour le différencier de la mouillure – également appelée *palatalization* dans la littérature anglo-saxonne (cf. la section 1.2.2) – des termes tels que *velar palatalizations* (Corbett et Comrie, 2003, p.8 ; Kapatsinski, 2010, p.361) ou *velar mutations* (Timberlake, 2004, pp.82-84 ; Sims, 2017, p.497) sont utilisés.

(39) a. ПЛАКАТЬ ‘pleurer’ :
PLAKAT’
плак-ать ‘pleurer-INF’ ; *плач-у* ‘pleurer-PRS.1.SG’
plak-at’ *plač-u*

b. ЖЕЧЬ ‘brûler’ :
ŽEČ’
жг-у ‘brûler-PRS.1.SG’ ; *жж-ёшь* ‘brûler-PRS.2.SG’
žg-u *žž-ëš’*

c. ПАХАТЬ ‘labourer’ :
PAXAT’
пах-ать ‘labourer-INF’ ; *паш-у* ‘labourer-PRS.1.SG’
pax-at’ *paš-u*

d. КРИ**К** ‘cri’ → КРИ**Ч**АТЬ ‘crier’
KRIK KRIČAT’

e. НО**Г**А ‘pied’ → НО**Ж**КА ‘petit pied’
NOGA NOŽKA

f. ПЕ**Ш**ИЙ ‘piéton’ → ПЕ**Х**ОТА ‘infanterie’
PEŠIJ PEXOTA

Il est à noter que, en plus de la palatalisation (39a-39e), le phénomène de dépalatalisation (39f) peut avoir lieu.

Les quatre alternances consonantiques identifiées représentent les principales palatalisations ayant eu lieu en vieux slave. Cependant, il existe d’autres cas spécifiques, tels que ceux répertoriés par Švedova (1980, p.435,505-506). Une liste non exhaustive de ces cas spécifiques est présentée en (40).

(40) a. /t/ ∼ /č/
СВЕ**Т**ИТЬ ‘éclairer’ → СВЕ**Ч**ЕНИЕ ‘lueur’
SVETIT’ SVEČENIE

b. /t/ ∼ /šč/
ОСВЕ**Т**ИТЬ ‘éclairer’ → ОСВЕ**Щ**ЕНИЕ ‘éclairage’
OSVETIT’ OSVEŠČENIE

c. /d/ ∼ /ž/
МЕДВЕ**Д**Ь ‘ours’ → МЕДВЕ**Ж**ОНОК ‘ourson’
MEDVED’ MEDVEŽONOK

d. /s/ ∼ /š/
ЛЕ**С** ‘forêt’ → ЛЕ**Ш**ИЙ ‘esprit de forêt’
LES LEŠIJ

e. /z/ ∼ /ž/
ФРАНЦУ**З** ‘ français’ → ФРАНЦУ**Ж**ЕНКА ‘française’
FRANCUZ FRANCUŽENKA

f. /st/ ∼ /šč/
толСТЫЙ ‘gros’ → толЩИНА ‘épaisseur’
TOLSTYJ TOLŠČINA

Les différents types de palatalisation représentent un ensemble d'alternances qui ont chacune leur propre distribution et leurs propres conditions. En outre, la productivité de ces palatalisations n'est pas homogène. Ainsi, Kapatsinski (2010, p.361) fait la distinction entre les noms de base d'origine slave et les noms de base empruntés. Les premiers présentent généralement des allomophies thématiques, tandis que dans les seconds, la palatalisation souvent ne se produit pas. De manière générale, Zemskaja (2011, p.114) met en évidence le fait que ce type d'alternance s'affaiblit dans la dérivation en russe moderne : par exemple, les formes adjectivales avec alternance sont progressivement remplacées par des formes sans alternance (41).

(41) a. казаХ ‘Kazakh’
KAZAX
казаШСКИЙ[17] ∼ казаХСКИЙ
KAZAŠSKIJ KAZAXSKIJ

b. калмыК ‘Kalmouk’
KALMYK
калмыЦКИЙ ∼ калмыКСКИЙ
KALMYCKIJ KALMYKSKIJ

c. устюГ ‘Oustioug (Russie)’
USTJUG
устюЖСКИЙ ∼ устюГСКИЙ
USTJUŽSKIJ USTJUGSKIJ

Pour systématiser les alternances non-linéaires qui surviennent lors de la dérivation des adjectifs, nous allons conclure cette section avec la classification proposée par Lopatin (1977, p.225-226), adaptée pour les besoins de cette recherche. Cette classification ne concerne que les alternances finales de consonnes, car, comme nous l'avons constaté en (36b), un seul processus dérivationnel peut impliquer à la fois une alternance de voyelle pleine avec une voyelle ∅ et une palatalisation. Ces phénomènes devraient donc être traités séparément. La classification que nous utiliserons est ainsi la suivante :

1. Absence d'alternance
байкер ‘motard’ → байкерский
BAJKER BAJKERSKIJ
Анадырь ‘Anadyr (Russie)’ → анадырьский
ANADYR' ANADYR'SKIJ

2. Alternance consonne dure / consonne molle
ампула ‘ampoule’ → ампульный
AMPULA AMPUL'NYJ

Alternance consonne molle / consonne dure
ель 'sapin' → еловый
el' elovyj

3. Alternance /k/ ∼ /č/, /g/ ∼ /ž/, /x/ ∼ /š/
каблук 'talon' → каблучный
kabluk kabločnyj
Волга 'Volga (Russie)' → волжский
Volga volžskij
чех 'Tchèque$_{N.}$' → чешский
čex češskij

L'adaptation que nous avons faite se manifeste notamment dans la simplification. La classification originale proposée par Lopatin prend en compte les alternances non linéaires qui concernent le système verbal, en termes de flexion et de dérivation. Lopatin distingue également la dépalatalisation, ainsi que les alternances suivantes : /b/ ∼ /bl/, /p/ ∼ /pl/, /t/ ∼ /č/, /t/ ∼ /šč/, /d/ ∼ /ž/, etc, les alternances inverses correspondantes et l'alternance /k/ ∼ /c/. Cependant, ces types d'alternances ne sont pas pertinents pour la dérivation des adjectifs dénominaux.

1.4.3 Transformations linéaires

Les allomorphies thématiques peuvent également résulter de transformations linéaires des thèmes. Quatre types de cas sont à considérer : l'ajout ou la suppression de matériel phonologique (Zemskaja, 2011, pp.117-155 ; Lopatin, 1977, pp.201-259, Švedova, 1980, pp.138-140), le remplacement et l'interférence des segments (Švedova, 1980, pp.420-424,449). Généralement, ces types de modifications ont lieu dans des cas où des séquences phonologiques ne sont pas admissibles en russe.

Par exemple, les noms d'origine étrangère ayant une voyelle finale n'ont pas de formes distinctes dans le paradigme flexionnel, cette voyelle faisant partie du thème des noms. Cependant, les voyelles finales des noms russes appartiennent à la flexion. Un moyen pour résoudre ce problème est d'ajouter une séquence supplémentaire, à savoir une consonne épenthétique, entre la voyelle finale du thème et le suffixe adjectival (42).

(42) Рено 'Renault' → рено**ш**ный
Reno renošnyj
кабаре 'cabaret' → кабаре**т**ный
kabare kabaretnyj
купе 'compartiment' → купе**й**ный
kupe kupejnyj

Cependant, tous les noms indéclinables se terminant par une voyelle ne sont pas concernés par l'ajout d'une consonne. Un autre moyen de résoudre le problème phonologique est la suppression d'un ou plusieurs phonèmes finaux du thème. Les

emprunts (43a) et notamment les noms géographiques d'origine étrangère (43b) subissent souvent la suppression des voyelles finales lors de la dérivation pour adapter l'adjectif construit à la langue.

(43) a. жалюз**и** 'stores' → жалюзевый
žaljuzi žaljuzevyj
пралин**е** 'praliné' → пралиновый
praline pralinovyj

b. Тбилис**и** 'Tbilissi' → тбилисский
Tbilisi tbilisskij
Чикаг**о** 'Chicago' → чикагский
Čikago čikagskij

Il existe toutefois des cas où aucune modification ne touche le thème d'un nom étranger se terminant par une voyelle (44) (Švedova, 1980, pp.420-424).

(44) Борд**о** 'Bordeaux' → бордоский
Bordo bordoskij
Грюна**у** 'Grünau' → грюнауский
Grjunau grjunauskij

Les séquences suivantes peuvent être également supprimées dans les noms d'origine slave (45a-45d) et étrangère (45e-45g) :

(45) a. *-k(a)*
Камчат**к**а 'Kamtchatka' → камчатский
Kamčatka kamčatskij

b. *-ec*
тунеяд**ец** 'fainéant' → тунеядный
tunejadec tunejadnyj

c. *-ok*
ублюд**ок** 'bâtard' → ублюдский
ubljudok ubljudskij

d. *-/ij/*
мгновен**ие** 'instant' → мгновенный
mgnovenie mgnovennyj

e. *-ik(a)*
диакрит**ик**а 'Diacritique' → диакритный
diakritika diakritnyj

f. *-/ij/*
лаборатори**я** 'laboratoire' → лабораторный
laboratorija laboratornyj

g. *-as/-us/-jaj/-os*

Капсук**ас** 'Marijampolė (Lituanie)'	→	капсукский
Kapsukas		Kapsukskij
Паневеж**ис** 'Panevėžys (Lituanie)'	→	паневежский
Panevežis		panevežskij
Райсен**яй** 'Raseiniai (Lituanie)'	→	райсенский
Rajsenjaj		rajsenskij

Comme dans le cas des voyelles finales, les séquences listées ci-dessus ne sont pas systématiquement supprimées, par exemple, la séquence -/*ij*/ (46).

(46)

бытие 'génèse'	→	бытийный
bytie		bytijnyj
Латвия 'Lettonie'	→	латвийский
Latvija		latvijskij

Dans le but d'adapter la forme des adjectifs dérivés à la phonologie, le russe propose également une option de remplacement de segments, ce qui peut être considéré comme un cas particulier d'ajout et de suppression simultanés. Par exemple, dans (47a), une des voyelles finales est supprimée, et une consonne est insérée après la voyelle restante. Dans (47b), le suffixoïde *-um* est enlevé et une semi-voyelle /j/ est placée entre la voyelle du thème et le suffixe adjectival.

(47)

a.

алоэ 'aloès'	→	алойный
aloè		alojnyj

b.

опиум 'opium'	→	опийный
opium		opijnyj

Le dernier type d'allomorphies concerne le phénomène d'interférence qui se produit lorsque la séquence finale du thème est totalement ou partiellement identique au suffixe (Zemskaja, 2011, pp.155-162). Dans ce cas, une fusion totale (48a) ou partielle (48b) est observée (Švedova, 1980, p.449).

(48)

a.

жернов 'meule'	→	жерновой
žernov		žernovoj
Гданьск 'Gdańsk'	→	гданьский
Gdan'sk		gdan'skij

b.

Кузбасс 'Kouzbass (Russie)'	→	кузбасский
Kuzbass		kuzbasskij
Дамаск 'Damas'	→	дамасский
Damask		damasskij

Conclusion

Le russe présente des propriétés phonologiques et morphologiques très complexes. Le système phonologique est caractérisé par la présence de l'accent primaire dont la position n'est pas fixe, ainsi que par la distinction entre les consonnes dures et les consonnes molles. Le système nominal se distingue par plusieurs classes de flexion pour les noms et l'utilisation de cas pour exprimer la fonction syntaxique. Le système adjectival, quant à lui, est tout aussi sophistiqué, avec des formes courtes et longues pour les adjectifs ainsi que la présence de comparatifs synthétiques et analytiques. En outre, le russe est caractérisé par des allomorphies variées incluant la mouillure, la palatalisation et l'alternance entre les voyelles pleines et les voyelles zéro, ainsi que par la suppression des séquences finales de thèmes et l'ajout de phonèmes épenthétiques.

Les propriétés phonologiques, morphologiques et morphophonologiques des noms peuvent imposer des contraintes sur le choix d'un des suffixes adjectivaux concurrents. L'ensemble des propriétés morphologiques adjectivales, à son tour, n'est pas distribué de manière uniforme entre les adjectifs qualificatifs et relationnels.

Chapitre 2

Les adjectifs comme classe lexicale

Sommaire

Introduction

La classification canonique différencie les adjectifs de qualité et les adjectifs de relation au sein de la catégorie lexicale des adjectifs. Cependant, cette classification est sujette à discussion en raison des frontières peu claires entre les deux classes du point de vue morphologique, syntaxique et sémantique. Ces questions seront abordées dans la section 2.1, où nous présenterons également l'importance de se concentrer sur les adjectifs dénominaux. La section 2.2 sera entièrement consacrée aux adjectifs dénominaux russes, incluant les différentes façons de les construire et une revue de leurs

propriétés sémantiques. L'analyse des adjectifs formés avec les suffixes *-n-*, *-sk-* et *-Ov-* sera alors justifiée. La dernière section 2.3 examinera les adjectifs construits avec les trois suffixes retenus, en mettant en évidence leurs différentes propriétés linguistiques.

2.1 Sous-classes des adjectifs

2.1.1 Propriétés des adjectifs

La caractéristique principale des adjectifs est qu'ils s'ajoutent à un nom dans le processus de désignation (Roché, 2006, p.375). En russe, les substantifs et les adjectifs sont regroupés sous la même catégorie de *имя* 'nom', avec des distinctions entre *имя существительное* 'nom substantif' et *имя прилагательное* 'nom adjectif' (Archaimbault, 1992, p.213). Ainsi, du point de vue terminologique, l'adjectif n'est pas distinct du nom. En ce qui concerne la distribution des adjectifs et leur combinaison avec les noms, tous les adjectifs ne partagent pas les mêmes propriétés.

La classification canonique (Švedova, 1980, p.538) subdivise les adjectifs en deux catégories en fonction de la propriété qu'ils désignent : les adjectifs qualificatifs, qui font référence aux propriétés qui ne peuvent pas être déduites à partir d'autres entités, et les adjectifs relationnels, qui se réfèrent aux relations qui peuvent être déduites à partir d'autres classes lexicales. Les adjectifs qualificatifs désignent des propriétés qui peuvent être perçues par les cinq sens (49a), les qualités spatiales et temporelles (49b), les couleurs (49c), les caractéristiques physiques (49d) ou mentales (49e) des individus, etc. (Švedova, 1980, p.541).

(49) a. ГОРЯЧИЙ 'chaud'
GORJAČIJ
ГРОМКИЙ 'fort (son)'
GROMKIJ
ДУШИСТЫЙ 'parfumé'
DUŠISTYJ

b. ДАЛЁКИЙ 'lointain'
DALËKIJ
КОРОТКИЙ 'court'
KOROTKIJ

c. КРАСНЫЙ 'rouge'
KRASNYJ
СИНИЙ 'bleu'
SINIJ

d. ГЛУХОЙ 'sourd'
GLUXOJ
СТАРЫЙ 'vieux'
STARYJ

e. ДОБРЫЙ ‘gentil’
DOBRYJ
СКУПОЙ ‘radin’
SKUPOJ

Les adjectifs qualificatifs impliquent une expérience subjective du locuteur, tandis que les adjectifs de relation reflètent la connaissance objective sur les entités et leurs propriétés (Arutjunova, 1999, p.39). Les adjectifs qualificatifs peuvent être divisés en deux sous-catégories : ceux qui désignent une propriété absolue, indépendante de la perception du locuteur (50a), et ceux qui transmettent un jugement subjectif, appelés adjectifs évaluatifs (50b).

(50) a. *немой человек* ‘un homme muet’
nemoj čelovek
полосатая рубашка ‘une chemise à carreaux’
polosataja rubaška

b. *длинная дорога* ‘une longue route’
dlinnaja doroga
сложный экзамен ‘un examen difficile’
složnyj èkzamen

Les adjectifs qualificatifs peuvent être simples (51a) ou construits (51b-51c).

(51) a. ПЛОХОЙ ‘mauvais’
PLOXOJ
ТЁПЛЫЙ ‘tiède’
TËPLYJ

b. СИНИЙ ‘bleu’ → СИНЕВАТЫЙ ‘bleuâtre’
SINIJ SINEVATYJ
ГЛАЗ ‘oeil’ → ГЛАЗАСТЫЙ ‘aux grands yeux’
GLAZ GLAZASTYJ

c. УМ ‘intelligence’ → УМНЫЙ ‘intelligent’
UM UMNYJ

Dans le cas des adjectifs construits exemplifié en (51c), comme le souligne Mezhevich (2002), même s’ils sont dérivés d’un nom, ils expriment une qualité plutôt qu’une relation entre un individu et un concept. Une tendance à la désémantisation pour les adjectifs qualificatifs construits est également observée.

Les adjectifs relationnels, quant à eux, désignent une relation entre le nom qu’ils modifient et une entité, une action ou une autre propriété désignée par le lexème à partir duquel ils sont dérivés. La nature de cette relation peut être très diverse, comme une relation matérielle (52a), de possession (52b), destination (52c), etc.

(52) a. ДЕРЕВО ‘bois’ → ДЕРЕВЯННЫЙ
DEREVO DEREVJANNYJ
СТАЛЬ ‘acier’ → СТАЛЬНОЙ
STAL’ STAL’NOJ

b. РЫБА ‘poisson’ → РЫБИЙ ‘poisson$_{\text{APPT}}$’
RYBA RYBIJ
МУЖ ‘mari’ → МУЖНИН ‘mari$_{\text{APPT}}$’
MUŽ MUŽNIN

c. РЕБЁНОК ‘enfant’ → ДЕТСКИЙ [1]
REBËNOK DETSKIJ

Les adjectifs de relation sont généralement construits par le biais de suffixation, contrairement aux adjectifs qualificatifs qui peuvent être simples ou construits (McNally et Boleda, 2004, pp.181-183). Les catégories lexicales qui peuvent servir de base pour les adjectifs de relation incluent les noms (53a), les verbes (53b), les numéraux (53c)[2] ou les adverbes (53d) (Švedova, 1980, p.541).

(53) a. ЖЕЛЕЗО ‘fer’ → ЖЕЛЕЗНЫЙ
ŽELEZO ŽELEZNYJ
ЛАМПА ‘lampe’ → ЛАМПОВЫЙ
LAMPA LAMPOVYJ

b. ПЛАВАТЬ ‘nager’ → ПЛАВАТЕЛЬНЫЙ
PLAVAT’ PLAVATEL’NYJ
ЛЕЧИТЬ ‘soigner’ → ЛЕЧЕБНЫЙ
LEČIT’ LEČEBNYJ

c. СОРОК ‘quarante’ → СОРОКОВОЙ
SOROK SOROKOVOJ
ДВЕСТИ ‘deux cent’ → ДВУХСОТЫЙ
DVESTI DVUXSOTYJ

d. БЛИЗКО ‘près’ → БЛИЖНИЙ
BLIZKO BLIŽNIJ
ВЧЕРА ‘hier’ → ВЧЕРАШНИЙ
VČERA VČERAŠNIJ

Les adjectifs de relation, qui représentent une grande proportion des adjectifs russes et continuent d’augmenter, sont différenciés des adjectifs qualificatifs : ces deux classes ne partagent pas l’ensemble de propriétés adjectivales. Selon Goes (1999), certains

[1] L’adjectif ДЕТСКИЙ (DETSKIJ) est formé sur le thème 2 de РЕБЁНОК (REBËNOK) ‘enfant’ qui est également utilisé dans formes fléchies au pluriel : *ребёнок* (*rebënok*) ‘enfant$_{\text{M.NOM.SG}}$’ ; *дет-и* (*det-i*) ‘enfant-M.NOM.PL’.

[2] Il s’agit ici des adjectifs numéraux ordinaux ; Švedova (1980, p.541) les considère comme faisant partie des adjectifs de relation car ils expriment une relation de quantité.

adjectifs peuvent être plus prototypiques que d'autres ; ce sont les adjectifs qualificatifs qui peuvent être considérés comme étant les adjectifs prototypiques, car ils englobent l'ensemble le plus complet des propriétés adjectivales. Selon la terminologie utilisée par Arutjunova (1999, p.39), ils sont considérés comme des adjectifs 'purs' car ils permettent de décrire notre perception du monde.

Les propriétés de distribution des adjectifs qualificatifs et relationnels pour le français sont systématisées dans le tableau 2.1 (Fradin, 2008, 2017)[3]. Du point de vue syntaxique, les adjectifs de relation se distinguent par leur usage non prédicatif et leur incapacité à être gradables. En général, ils ne se placent pas en position antéposée par rapport au nom qu'ils modifient.

Propriété	Adjectifs qualificatifs	Adjectifs relationnels
Emploi attribut	*le service est réglementaire*	**la visite est ducale*
Épithète antéposé	*(de) broussailleux sourcils*	**ducale visite*
Épithète postposé	*tronc moussu*	*visite ducale*
Gradabilité	*voix très caverneuse*	**la visite très ducale*

Tableau 2.1: Propriétés des adjectifs qualificatifs et relationnels en français

Les adjectifs qualificatifs, surtout ceux qui ne sont pas dérivés et appartiennent au vieux fond de la langue, présentent un ensemble des quatre propriétés listées par Fradin (2008, 2017). Corbett (2004, pp.200-219), à son tour, cite les propriétés des adjectifs prototypiques russes :

- Emploi attribut ;
- Emploi épithète ;
- Présence des formes longue et courte ;
- Accord avec le nom recteur en genre, nombre, cas (forme longue) et en genre, nombre (court) ;
- Présence du comparatif synthétique.

Comme dans le cas du français, les adjectifs russes prototypiques (qualificatifs) peuvent apparaître dans des constructions syntaxiques en tant qu'attribut (54a) et en tant qu'épithète (54b).

(54) a. *Катя очень красивая* 'Katja est très jolie'
Katja očen' krasivaja

[3]Tiré et simplifié de Fradin (2008, 2017). Dans l'original, il y a plus de cases : Fradin distingue notamment deux propriétés supplémentaires : la répétition (A, A N ; N A, A) et l'apposition (N, A, ...). Toutefois, Fradin note que seules les trois propriétés sont essentielles : la gradabilité, la possibilité d'être attribut et la possibilité d'être épithète.

b. *Какая красивая девушка!* ‘Quelle jolie jeune fille !’
Kakaja krasivaja devuška

Les adjectifs russes peuvent avoir deux formes : une forme longue utilisée en tant qu’épithète et une forme courte utilisée uniquement en tant qu’attribut, qui ne s’accorde qu’en genre et nombre avec le nom qu’elle modifie (cf. la section 1.3.2). Les adjectifs de qualité sont dotés de deux formes (55a) tandis que les adjectifs de relation ne possèdent qu’une forme longue (55b).

(55) a. *горьк-ая* *реальность* ‘une réalité amère’
gor’ka-ja *real’nost’*
amer-F.NOM.SG réalité$_{\text{F.NOM.SG}}$
∼ *реальность* *горьк-а* ‘la réalité est amère’
real’nost’ *gor’k-a*
réalité$_{\text{F.NOM.SG}}$ amer-F.NOM.SG (court)
бел-ое *лиц-о* ‘visage blanc’
bel-oe *lic-o*
blanc-N.NOM.SG visage-N.NOM.SG
∼ *лиц-о* *бел-о* ‘le visage est blanc’
lic-o *bel-o*
visage-N.NOM.SG blanc-N.NOM.SG (court)

b. *президентск-ий* *дворец* ‘palais présidentiel’
prezidentsk-ij *dvorec*
présidentiel-M.NOM.SG palais$_{\text{M.NOM.SG}}$
**дворец* *президентск* ‘le palais est présidentiel’
dvorec *prezidentsk*
palais$_{\text{M.NOM.SG}}$ présidentiel$_{\text{M.NOM.SG}}$ (court)
книжн-ый *магазин* ‘magasin de livres’
knižn-yj *magazin*
livre$_{\text{REL}}$-M.NOM.SG magasin$_{\text{M.NOM.SG}}$
**магазин* *книжен* ‘le magasin est de livres’
magazin *knižen*
magasin$_{\text{M.NOM.SG}}$ livre$_{\text{REL.M.NOM.SG}}$ (court)

La distinction entre les formes courtes et longues des adjectifs russes permet de les classer selon leur fonction syntaxique : les formes courtes sont utilisées de manière prédicative, rapprochant ainsi ces adjectifs des verbes, tandis que les formes longues fléchies sont plus proches des noms (Corbett, 2004, pp.200-219)[4]. Les adjectifs ayant une distinction entre les formes longue et courte peuvent également être gradables et avoir un comparatif synthétique (56a). Par contre, les adjectifs de relation ne peuvent pas désigner des propriétés mesurables (56b).

[4]Selon cette perspective, Corbett (2004) décrit un continuum entre les noms et les verbes, sur lequel les adjectifs sont placés en fonction de leurs propriétés morphologiques et syntaxiques.

(56) a. *важн-ая* *истори-я*
važna-ja *istori-ja*
important-F.NOM.SG histoire-F.NOM.SG
'histoire importante'
∼ *эта* *истори-я* *важн-ее*
èta *istori-ja* *važn-ee*
DEM histoire-F.NOM.SG important-COMP
'cette histoire est plus importante'
добр-ый *человек*
dobr-yj *čelovek*
gentil-M.NOM.SG homme$_{\text{M.NOM.SG}}$
'gentil homme'
∼ *этот* *человек* *добр-ее*
ètot *čelovek* *dobr-ee*
DEM homme$_{\text{M.NOM.SG}}$ gentil-COMP
'cet homme est plus gentil'

b. *диакритичн-ое* *письм-о*
diakritičn-oe *pis'm-o*
diacritique-N.NOM.SG écriture-N.NOM.SG
'écriture diacritique'
**это* *письм-о* *диакричн-ее*
èto *pis'm-o* *diakritičn-ee*
DEM écriture-N.NOM.SG diacritique-COMP
'cette écriture est plus diacritique'
французск-ий *гимн*
francuzsk-ij *gimn*
français-M.NOM.SG hymne$_{\text{M.NOM.SG}}$
'hymne français'
**этот* *гимн* *француж-е*
ètot *gimn* *francuž-e*
DEM hymne$_{\text{M.NOM.SG}}$ français-COMP
'cet hymne est plus français'

Étant donnée la nature gradable de l'adjectif de qualité, il peut être modifié par l'adverbe ОЧЕНЬ (OČEN') 'très' (57a) ce qui n'est pas le cas des adjectifs de relation (57b), selon les observations de Mezhevich (2002, p.100).

(57) a. *красивая девушка* 'jolie fille'
krasivaja devuška
∼*очень красивая девушка* 'très jolie fille'
očen' krasivaja devuška

b. *президентский дворец* 'palais présidentiel'
prezidentskij dvorec

**очень президентский дворец* ‘palais très présidentiel’
očen’ prezidentskij dvorec

Du point de vue syntaxique, lorsqu’un syntagme nominal contient un adjectif de relation et un adjectif de qualité, c’est l’adjectif de relation (ИГРУШЕЧНЫЙ (IGRUŠEČNYJ) ‘jouet$_{\text{REL}}$’) qui précède généralement le nom (58a). Ces deux sous-classes d’adjectifs ne peuvent pas être reliées par une conjonction (58b).

(58) a. *большой игрушечный магазин* ‘grand magasin de jouets’
bol’šoj igrušečnyj magazin
**игрушечный большой магазин*
igrušečnyj bol’šoj magazin

b. **красивый и деревянный дом* ‘une maison belle et en bois’
krasivyj i derevjannyj dom

Jusqu’à présent nous nous sommes concentrés sur les propriétés syntaxiques des adjectifs qualificatifs et relationnels. Avant de conclure cette section, il convient de préciser que ces deux sous-classes se distinguent dans leurs propriétés dérivationnelles (Švedova, 1980, p.541). Ainsi, les adjectifs de qualité peuvent être utilisés comme base pour la formation d’adverbes en utilisant les suffixes *-o* et *-e* (59a). Ils peuvent également servir de base pour la dérivation d’autres adjectifs qualificatifs exprimant des degrés différents de la même qualité (59b), ainsi que pour la formation de noms abstraits (59c). Ce n’est pas le cas pour les adjectifs de relation.

(59) a. ГОРЯЧИЙ ‘chaud$_{\text{A}}$’ → ГОРЯЧО ‘chaud$_{\text{Adv}}$’
GORJAČIJ GORJAČO
МУДРЫЙ ‘sage$_{\text{A}}$’ → МУДРО ‘sage$_{\text{Adv}}$’
MUDRYJ MUDRO

b. БЕЛЫЙ ‘blanc’ → БЕЛОВАТЫЙ ‘blanchâtre’
BELYJ BELOVATYJ
БОЛЬШОЙ ‘grand’ → БОЛЬШУЩИЙ ‘le plus grand’
BOL’ŠOJ BOL’ŠUŠČIJ

c. СМЕЛЫЙ ‘courageux’ → СМЕЛОСТЬ ‘courage’
SMELYJ SMELOST’
ПУСТОЙ ‘vide$_{\text{A}}$’ → ПУСТОТА ‘vide$_{\text{N}}$’
PUSTOJ PUSTOTA

Il est à noter que tous les adjectifs qualificatifs ne possèdent pas les propriétés listées dans le tableau 2.1 : Fradin (2008) cite des exemples d’adjectifs non prédicables (CADET), non antéposables (PANSU), prédicables mais non gradables et non antéposables (BORGNE)[5]. Il en va de même pour les adjectifs de relation : ainsi, McNally et Boleda,

[5]Ces particularités poussent Fradin à distinguer jusqu’à six sous-groupes d’adjectifs, au lieu de parler des adjectifs qualificatifs et relationnels seuls.

2004 ; Roché, 2006 ; Fradin, 2008 et Rainer, 2013 évoquent la possibilité pour ces adjectifs d'être prédicables (par exemple, MENSUEL, dans *Notre revenue est mensuelle*).

2.1.2 Frontières entre les classes

La distinction entre les adjectifs de qualité et les adjectifs de relation repose principalement sur leur sémantique. Cependant, les frontières entre les deux sous-classes d'adjectifs ne sont pas claires, ni du point de vue sémantique, ni morphologique, ni syntaxique.

En ce qui concerne le niveau sémantique, les adjectifs de relations en russe peuvent acquérir une signification qualificative. Le sens qualificatif de ces adjectifs peut être déduit à partir du contexte (60a-60b), cependant, le nom recteur ne suffit pas à déduire la nature de l'adjectif, comme dans (60c) où *игрушечный магазин* (*igrušečnyj magazin*) peut désigner un magasin qui vend des jouets, un magasin qui est lui-même un jouet (Mezhevich, 2002, p.99) ou un vrai magasin qui est très petit en taille.

(60) a. *железная руда* 'minerai de fer'
železnaja ruda
железное здоровье 'santé de fer'
železnoe zdorov'e

b. *детские игрушки* 'jouets d'enfants'
detskie igruški
детское поведение 'comportement enfantin'
detskoe povedenie

c. *игрушечный магазин* 'magasin de jouets, magasin-jouet, petit magasin'
igrušečnyj magazin

Selon Vinogradov (1952, pp.160-162), la frontière entre le sens qualificatif et relationnel varie selon les cas. Par exemple, le mot КУСТАРНЫЙ (KUSTARNYJ) peut avoir un sens qualificatif positif ('artisanal, non industriel') en faisant partie des adjectifs de relation (61a), mais il a également développé un nouveau sens qualificatif péjoratif (61b).

(61) a. *кустарная промышленность* 'industrie de l'artisanat'
kustarnaja promyšlennost'
кустарные изделия 'marchandises artisanales'
kustarnye izdelija

b. *кустарный способ производства* 'méthode de production rudimentaire'
kustarnyj sposob proizvodstva

Le caractère qualificatif des adjectifs de relation est marqué soit par la gradabilité (62a) soit par une métaphore (62b)[6], selon Zemskaja (2015, pp.233-239).

[6]Zemskaja (2015, pp.233-239) utilise le terme *имплицитное сравнение* 'comparaison implicite' pour analyser ce cas.

(62) a. *Она выбрала самый конкурсный фильм*
Ona vybrala samyj konkursnyj fil'm
'elle a choisi le film le plus concurrentiel'

b. *страусиная политика*
strausinaja politika
'politique de l'autruche (comme chez l'autruche qui cache sa tête)'

Švedova (1980, pp.539-543) souligne que les différents degrés de qualification peuvent être propres à tous les adjectifs de relation (à l'exception des adjectifs d'appartenance). Cependant, Zemskaja (2015, pp.233-239) précise que la plupart des adjectifs qui développent un sens qualificatif sont formés à partir de noms de base qui localisent un objet dans l'espace. Ainsi, l'adjectif peut acquérir le sens 'propre à [l'endroit]' (63).

(63) *ярмарочная публика* 'public forain'
jarmaročnaja publika

La frontière au niveau morphologique entre les deux sous-classes d'adjectifs n'est pas stable non plus. Švedova (1980, pp.539-543) mentionne que certains adjectifs qualificatifs qui désignent les races des animaux (64a) et des degrés élevés de qualité (64b) n'ont pas de formes courtes.

(64) a. БУЛАНЫЙ 'isabelle (cheval)' : **булан* 'isabelle$_{\text{M.NOM.SG}}$ (court)'
BULANYJ *bulan*
ГНЕДОЙ 'champagne (cheval)' : **гнед* 'champagne$_{\text{M.NOM.SG}}$ (court)'
GNEDOJ *gned*

b. ПРЕБОГАТЫЙ 'le plus riche' :
PREBOGATYJ
**пребогат* 'le plus riche$_{\text{M.NOM.SG}}$ (court)'
prebogat
НАИПРЕКРАСНЕЙШИЙ 'le plus beau' :
NAIPREKRASNEJŠIJ
**наипрекраснейш* 'le plus beau$_{\text{M.NOM.SG}}$ (court)'
naiprekrasnejš

Certains adjectifs ne possèdent pas de forme courte dans tous les contextes, cela dépend de l'activation d'un de leurs sens. Par exemple, l'adjectif ГРАМОТНЫЙ (GRAMOTNYJ) a une forme courte dans le sens 'qui sait lire, instruit' (65a), mais pas dans le sens 'fait sans fautes' (65b). De même, certains adjectifs font partie d'expressions figées et n'ont qu'une forme longue (65c).

(65) a. *грамотн-ый* *мальчик* 'garçon instruit'
gramotn-yj *mal'čik*
instruit-M.NOM.SG garçon$_{\text{M.NOM.SG}}$

	∼*мальчик*	*грамотен*	‘le garçon est instruit’
	mal’čik	*gramoten*	
	garçon$_{\text{M.NOM.SG}}$	instruit$_{\text{M.NOM.SG}}$ (court)	
b.	*грамотн-ый*	*чертёж*	‘bon [précis] croquis’
	gramotn-yj	*čertëž*	
	précis-M.NOM.SG	croquis$_{\text{M.NOM.SG}}$	
	**чертёж*	*грамотен*	‘le croquis est bon [précis]’
	čertëž	*gramoten*	
	croquis$_{\text{M.NOM.SG}}$	précis$_{\text{M.NOM.SG}}$ (court)	
c.	*бел-ый*	*свет*	‘le monde entier [blanc]’
	bel-yj	*svet*	
	blanc-M.NOM.SG	monde$_{\text{M.NOM.SG}}$	
	**свет*	*бел*	‘le monde est entier [blanc]’
	svet	*bel*	
	monde$_{\text{M.NOM.SG}}$	blanc$_{\text{M.NOM.SG}}$ (court)	
	кругл-ый	*дурак*	‘pauvre [rond] fou’
	krugl-yj	*durak*	
	rond-M.NOM.SG	fou$_{\text{N.M.NOM.SG}}$	
	**дурак*	*кругл*	‘le fou est pauvre [rond]’
	durak	*krugl*	
	fou$_{\text{N.M.NOM.SG}}$	rond$_{\text{M.NOM.SG}}$ (court)	

Inversement, les adjectifs de relation qui développent un sens qualificatif possèdent des formes courtes (66a) et des degrés de comparaison (66b).

(66) a. *рек-а* *черн-а* *и* *бархатн-а*
rek-a *čern-a* *i* *barxatn-a*
fleuve-F.NOM.SG noir-F.NOM.SG (court) CONJ velours-F.NOM.SG (court)
‘le fleuve est noir et en velours’

b. *слов-а* *железн-ей* *нет*
slov-a *železn-ej* *net*
mot-N.GEN.SG fer$_{\text{REL}}$-COMP avoir$_{\text{PRS.NEG}}$
‘il n’y a pas de mot qui soit plus sûr [en fer]’

L’incertitude des frontières entre les classes d’adjectifs pousse Vinogradov (1952, pp.160-162) à affirmer que tous les adjectifs de relation peuvent potentiellement développer un sens qualificatif. Selon Roché (2006, p.385), il s’agit plutôt de la recréation de ce sens, car les adjectifs qualificatifs précèdent historiquement les adjectifs de relation (Bartning et Noailly, 1993). Le sens qualificatif est dû aux différentes connotations sémantiques du nom de base, la métaphorisation faisant apparaître ce sens, souvent figuré. Les différences morphologiques et syntaxiques entre les adjectifs qualificatifs et relationnels ne sont ni stables, ni obligatoires, car les adjectifs de relation

en développant un sens qualificatif peuvent élargir leurs paradigmes avec des formes courtes (cf. la section 1.3.2), ainsi que construire des adverbes.

Ainsi, les adjectifs de relation et les adjectifs qualificatifs ne forment pas deux sous-classes d'adjectifs fermées. Les frontières entre les deux sont floues, avec un grand nombre d'adjectifs pouvant se trouver à leur intersection ou varier selon les contextes entre une interprétation relationnelle et une interprétation qualificative. De plus, Fradin (2008) et Strnadová (2014) soutiennent l'idée qu'il n'est pas nécessaire ni important de classer les adjectifs en qualificatifs et relationnels. La tâche la plus importante est plutôt de repérer les facteurs de changement de leur comportement pour pouvoir prédire leur interprétation qualificative ou relationnelle dans un contexte donné.

2.1.3 Vers d'autres classifications

La sémantique formelle distingue trois sous-classes d'adjectifs : les adjectifs intersectifs, les adjectifs subsectifs et les adjectifs intensionnels (McNally et Boleda, 2004 ; Partee, 2009). Les adjectifs prototypiques, c'est-à-dire les adjectifs qualificatifs gradables employés en tant qu'épithètes et en tant qu'attributs, sont intersectifs et dénotent les propriétés des noms. Lorsqu'ils sont utilisés comme épithètes dans des constructions telles que (67a) *Jean est un chanteur parisien*, il est possible d'en faire deux inférences (Jean est chanteur ; Jean est parisien). À l'inverse, les adjectifs subsectifs désignent les propriétés des propriétés et ne permettent qu'une seule inférence (67b). Les adjectifs intensionnels, similaires aux adjectifs subsectifs en ce qu'ils désignent également les propriétés des propriétés, ne permettent pas de faire d'inférences (67c).

(67) a. *Jean est un chanteur parisien*
   ⊨ Jean est chanteur
   ⊨ Jean est parisien

b. *Jean est un grand chanteur*
   ⊨ Jean est un chanteur
   ⊭ Jean est grand

c. *Jean est un ancien chanteur*
   ⊭ Jean est un chanteur
   ⊭ Jean est ancien

La sémantique des adjectifs de relation est similaire à celle de leur nom de base, et le prédicat qui instancie la relation dépend du contexte. Le nom de base peut alors devenir un argument potentiel d'une relation sémantique. Ces adjectifs ont été étudiés dans le cadre de la sémantique formelle. Par exemple, McNally et Boleda, 2004 ; Arsenijevic *et al.*, 2010 ; Gehrke et McNally, 2015 et Fradin, 2017 proposent une lecture intersective pour les adjectifs de relation, qu'ils soient dénominaux ou non. Cette analyse intersective est basée notamment sur la capacité de ces adjectifs à

remplir un rôle d'attribut. De plus, Fradin (2017) remarque qu'il n'a pas été observé de cas où les adjectifs dénominaux sont analysés comme subsectifs ou intensionnels.

La dichotomie stricte entre les adjectifs de relation et les adjectifs de qualité n'est plus considérée comme nécessaire par les chercheurs en sémantique formelle. Roché (2006) propose de s'intéresser aux adjectifs construits à la place, tandis que Fradin (2017) se concentre sur les adjectifs dénominaux, qui peuvent être considérés comme des adjectifs intersectifs. Dans la suite de cette étude, nous discuterons de l'intérêt d'utiliser les adjectifs dénominaux comme objet d'étude, y compris pour le russe.

Il est fréquent d'associer les adjectifs relationnels aux adjectifs dénominaux (Bally, 1944, p.87), et certains auteurs affirment même que tous les adjectifs relationnels sont dérivés sur une base nominale (Noailly, 1999, p.22). Cependant, comme le montrent les exemples (53b-53d), les adjectifs de relation en russe peuvent également être dérivés de verbes, de numéraux et d'adverbes. Néanmoins, les adjectifs dénominaux sont majoritaires dans les constructions adjectivales.

En outre, les adjectifs dénominaux ne forment pas une classe homogène : certains d'entre eux ne sont pas dérivés, mais se comportent comme des adjectifs de relation, comme, par exemple, en français : TERRESTRE ($<$ lat. *terrestris*) ou SOLAIRE ($<$ lat. *solaris*). Ces adjectifs sont sémantiquement associés aux noms TERRE et SOLEIL (Fradin, 2008, 2017). D'autres adjectifs ont un suffixe dérivationnel et sont lexicalement liés aux noms de base, sans pour autant être lié morphologiquement (VILLE $\rightarrow$ URBAIN). Dans ces cas s'agit alors d'une famille dérivationnelle lexicale (Hathout, 2011), et non d'une famille dérivationnelle morphologique.

Il est donc important de noter que le fait d'être dénominal ne constitue pas une propriété nécessaire pour les adjectifs de relation (McNally et Boleda, 2004, pp.181-183). De plus, il existe des adjectifs dénominaux qui peuvent exprimer une relation, tandis que d'autres désignent une qualité. Il apparaît donc que la majorité des adjectifs relationnels est incluse dans le groupe des adjectifs dénominaux, mais que tous les adjectifs de relation ne sont pas nécessairement dénominaux, et vice versa.

2.2 Les adjectifs dénominaux

Jusqu'à présent, nous avons examiné les propriétés des adjectifs et analysé les principales classifications existantes dans cette catégorie lexicale. Nous avons également fait le point sur les adjectifs russes dans leur ensemble. Dans cette section, nous allons nous concentrer sur la formation des adjectifs dénominaux russes en examinant les suffixes *-n-*, *-sk-*, *-Ov-* et en expliquant les raisons de notre choix de limiter notre étude à ces suffixes.

2.2.1 La place des adjectifs d'appartenance

Les adjectifs d'appartenance (68) ne font pas partie de cette étude, cependant une discussion est nécessaire pour justifier l'exclusion de cette classe des adjectifs.

(68) a. ОТЕЦ ‘père’ → ОТЦОВ ‘père$_{\text{APPT}}$’
OTEC OTCOV

b. МАМА ‘mère’ → МАМИН ‘mère$_{\text{APPT}}$’
MAMA MAMIN

c. МУЖ ‘mari’ → МУЖНИН ‘mari$_{\text{APPT}}$’
MUŽ MUŽNIN

d. РЫБА ‘poisson’ → РЫБИЙ ‘poisson$_{\text{APPT}}$’
RYBA RYBIJ

En slavistique, les adjectifs d’appartenance sont souvent considérés comme un cas particulier d’adjectifs relationnels en raison de la nature des relations qu’ils expriment, qui est celle de la possession. Cependant, il est important de noter que les adjectifs d’appartenance et les adjectifs de relation sont souvent regroupés car ils sont tous les deux construits à partir de noms, majoritairement pour les adjectifs de relation et toujours pour les adjectifs d’appartenance. Les adjectifs de relation expriment une variété de relations entre les entités, tandis que les adjectifs d’appartenance expriment uniquement des relations de possession.

Toutefois, la sémantique plus restreinte des adjectifs d’appartenance par rapport à celle des adjectifs de relation soulève quelques problèmes. Selon Vinogradov (1952, pp.151-158), leur statut adjectival est très subjectif car ils ont plutôt une fonction individualisante. Cet auteur souligne également que l’idée d’appartenance, ce qui représente le noyau sémantique de ces adjectifs, est difficilement compatible avec la sémantique de l’adjectif en tant que classe lexicale qui est celle de qualité. Depuis la discussion de Vinogradov, les grammairiens russes n’ont pas consacré beaucoup d’attention aux adjectifs d’appartenance.

Les adjectifs d’appartenance sont dérivés en russe avec les suffixes -*Ov* (68a), -*in* (68b), -*nin* (68c) ou par le biais de la conversion (68d).

Du point de vue sémantique, il existe des restrictions quant à la dérivation des adjectifs d’appartenance à partir des noms. Selon Corbett (1995), ces restrictions sont nombreuses. Les adjectifs d’appartenance sont principalement formés à partir de noms désignant des êtres humains (69a), mais ils peuvent également être dérivés de noms désignant des animaux, généralement des animaux supérieurs (69b). Dans le style littéraire, il est possible de former des adjectifs d’appartenance à partir de noms communs désignant des personnes (69a) et de noms propres (69c).

(69) a. ГОСУДАРЬ ‘gouverneur’ → ГОСУДАРЕВ ‘gouverneur$_{\text{APPT}}$’
GOSUDAR’ GOSUDAREV
ЧЕЛОВЕК ‘homme’ → ЧЕЛОВЕЧИЙ ‘homme$_{\text{APPT}}$’
ČELOVEK ČELOVEČIJ
ЛЕСНИК ‘forestier’ → ЛЕСНИКОВ ‘forestier$_{\text{APPT}}$’
LESNIK LESNIKOV

b. ВОРОН ‘corbeau’ → ВОРОНОВ ‘corbeau$_{\text{APPT}}$’
VORON VORONOV
КУКУШКА ‘coucou’ → КУКУШКИН ‘coucou$_{\text{APPT}}$’
KUKUŠKA KUKUŠKIN

c. Игорь ‘Igor’ → ИГОРЕВ ‘Igor$_{\text{APPT}}$’
Igor’ IGOREV
Саша ‘Sacha’ → САШИН ‘Sacha$_{\text{APPT}}$’
Saša SAŠIN

En raison de ces restrictions sémantiques qui limitent la formation des adjectifs d’appartenance à certains types de noms, leur utilisation est généralement réservée à un contexte familier (Corbett, 1987). Il est toutefois possible d’utiliser les adjectifs d’appartenance dans différents types de discours. Par exemple, lorsque les adjectifs d’appartenance sont dérivés à partir de noms propres, ils peuvent être employés dans des expressions figées (70a) et dans la terminologie (70b).

(70) a. *марсово поле* ‘Champs de Mars’
marsovo pole
сизифов труд ‘travail de Sisyphe’
sizifov trud

b. *эвклидова геометрия* ‘géométrie euclidienne’
èvklidova geometrija

Du point de vue morphologique, les adjectifs d’appartenance ne sont pas fléchis de la même manière que les adjectifs qualificatifs et relationnels (voir la section 1.3.2). Ils ont un paradigme flexionnel mixte qui combine des flexions adjectivales avec des flexions nominales (voir la section 1.3.1). De plus, ils n’ont pas de forme courte.

Du point de vue syntaxique, selon Vinogradov (1952, pp.151-158), les adjectifs d’appartenance sont souvent remplacés par des formes plus universelles telles que les adjectifs de relation ou des constructions avec le génitif.

Dans la section 2.1.2, nous avons examiné l’instabilité des frontières entre les adjectifs de qualité et les adjectifs de relation. Les adjectifs d’appartenance, de leur côté, n’ont jamais de sens qualificatif, car ils dénotent une propriété individuelle. En raison de toutes ces particularités, Vinogradov (1952, pp.151-158) remarque que la formation et l’utilisation des adjectifs d’appartenance n’ont pas de perspectives en russe, étant donné leur formation limitée et leur utilisation de moins en moins fréquente[7].

En raison des particularités des adjectifs d’appartenance, telles que la sémantique limitée de leurs noms de base, la déclinaison mixte entre les noms et les adjectifs, ainsi que leur usage limité, nous avons choisi de ne pas les inclure dans notre analyse. Ainsi, dans la suite de notre étude, nous nous concentrons sur les adjectifs dénominaux – relationnels ou qualificatifs – formés à l’aide des suffixes *-n-*, *-sk-* et *-Ov-*.

[7]Voir Šmelëva (2008) pour une discussion sur le destin des adjectifs d’appartenance en russe.

2.2.2 Variantes suffixales en russe

La dérivation des adjectifs à partir de noms est un sujet complexe en morphologie russe, car l'inventaire des suffixes utilisés dans ce type de dérivation est très riche. Les grammaires telles que Townsend (1975) ou Švedova (1980), par exemple, énumèrent plus de 25 suffixes différents, dont les degrés de productivité varient.

Selon les études en synchronie (Zemskaja, 2015 ; Kustova, 2018), trois suffixes sont identifiés comme étant les plus productifs : *-n-* (71a), *-sk-* (71b) et *-Ov-* (71c)[8].

(71) a. РЮКЗАК 'sac à dos' → РЮКЗАЧНЫЙ
RJUKZAK RJUKZAČNYJ

b. ШВЕЙЦАРИЯ 'Suisse' → ШВЕЙЦАРСКИЙ
ŠVEJCARIJA ŠVEJCARSKIJ

c. ГАЗ 'gaz' → ГАЗОВЫЙ
GAZ GAZOVYJ
ГРЯЗЬ 'boue' → ГРЯЗЕВОЙ
GRJAZ' GRJAZEVOJ

A la différence du suffixe *-Ov-*, les suffixes *-n-* et *-sk-* possèdent de nombreusese variantes étendues[9] : ces suffixes se terminent par la séquence *-n-* ou *-sk-* et contiennent un matériel phonologique supplémentaire au début.

Les suffixes *-n-*, *-sk-* et *-Ov-* peuvent être considérés comme les trois principaux suffixes adjectivaux (les entités abstraites, marquées en majuscule), tandis que les représentations en minuscules marquent la réalisation concrète (Bobkova et Montermini, 2019).

- *-N-*:
  - *-n-* : ЛИТЕРАТУРА 'littérature' → ЛИТЕРАТУРНЫЙ,
    LITERATURA LITERATURNYJ
  - *-Ovn-* : ВИНА 'faute' → ВИНОВНЫЙ,
    VINA VINOVNYJ
  - *-ičn-* : ЦИКЛ 'cycle' → ЦИКЛИЧНЫЙ,
    CIKL CIKLIČNYJ
  - *-ivn-* : ЭКСПРЕССИЯ 'expression' → ЭКСПРЕССИВНЫЙ,
    ÈKSPRESSIJA ÈKSPRESSIVNYJ
  - *-On(n)-* : ДИВИЗИЯ 'division' → ДИВИЗИОННЫЙ,
    DIVIZIJA DIVIZIONNYJ

[8]Pour rappel, le *O* majuscule correspond graphiquement à <*o*> ou <*e*> ce qui résulte en deux suffixes : *-ov-* et *-ev-* ; cf. 71c.

[9]Appelés *expansions* dans la terminologie de Corbett et Comrie (2003, p.856).

- *-(e)stven(n)-* : НАСИЛИЕ ‘violence’ → НАСИЛЬСТВЕННЫЙ,
  NASILIE NASIL’STVENNYJ
- *-ozn-* : РЕЛИГИЯ ‘religion’ → РЕЛИГИОЗНЫЙ,
  RELIGIJA RELIGIOZNYJ
- *-al’n-* : ГЕНИЙ ‘génie’ → ГЕНИАЛЬНЫЙ,
  GENIJ GENIAL’NYJ
- *-onaln-* : НАЦИЯ ‘nation’ → НАЦИОНАЛЬНЫЙ,
  NACIJA NACIONAL’NYJ
- *-arn-* : МОЛЕКУЛА ‘molécule’ → МОЛЕКУЛЯРНЫЙ ;
  MOLEKULA MOLEKULJARNYJ

• *-SK-*:

- *-sk-* : ГРУЗИН ‘Géorgien$_N$’ → ГРУЗИНСКИЙ,
  GRUZIN GRUZINSKIJ
- *-esk-* : ТОВАРИЩ ‘camarade’ → ТОВАРИЩЕСКИЙ,
  TOVARIŠČ TOVARIŠČESKIJ
- *-česk-* : ПОВЕДЕНИЕ ‘comportement’ → ПОВЕДЕНЧЕСКИЙ,
  POVEDENIE POVEDENČESKIJ
- *-ičesk-* : ПРОЗА ‘prose’ → ПРОЗАИЧЕСКИЙ,
  PROZA PROZAIČESKIJ
- *-ijsk-* : Альпы ‘Alpes’ → АЛЬПИЙСКИЙ,
  Alpy AL’PIJSKIJ
- *-(j)ansk-* : Конфуций ‘Confucius’ → КОНФУЦИАНСКИЙ,
  Konfucij KONFUCIANSKIJ
- *-insk-* : Куба ‘Cuba’ → КУБИНСКИЙ,
  Kuba KUBINSKIJ
- *-Ovsk-* : ОТЕЦ ‘père’ → ОТЦОВСКИЙ ;
  OTEC OTCOVSKIJ

• *-OV-*:

- *-Ov-* : БЕРЕГ ‘rive’ → БЕРЕГОВОЙ.
  BEREG BEREGOVOJ

Le statut des variantes de *-n-* et *-sk-* est discutable et dépend de chaque variant.

En ce qui concerne les suffixes *-Ovsk-* (72a) et *-insk-* (72b), par exemple, certains linguistes, comme Zemskaja (2011, pp.123-133,141), les considèrent comme des interfixes, car ces séquences ne font pas partie ni du thème de base, ni du suffixe *-sk-*.

(72) a. Ялта ‘Yalta (Ukraine)’ → ялтинский
Jalta jaltinskij

b. Орёл ‘Orel (Russie)’ → орловский
Orël orlovskij

Dans ces exemples, le suffixe principal -*sk*- est suffisant pour que l'adjectif dérivé exprime un sens ‘relatif à X’. De plus, Zemskaja affirme que les notions de suffixe simple et suffixe composé sont absentes en synchronie, l'ajout de ces séquences ne change pas le sens des adjectifs construits uniquement avec le suffixe principal. Par conséquent, les variantes des suffixes créent souvent des doublons : entre le mot dérivé avec le suffixe principal et une variante (73a) ou entre les mots dérivés avec les deux variantes (73b).

(73) a. завком ‘comité d'usine’ → завкомский / завкомовский
zavkom zavkomskij zavkomovskij
практика ‘pratique’ → практичный / практический
praktika praktičnyj praktičeskij

b. Брно ‘Brno (Tchéquie)’ → брненский / брновский
Brno brnenskij brnovskij
Воркута ‘Vorkouta (Russie)’ →
Vorkuta
воркутинский / воркутовский
vorkutinskij vorkutovskij

Selon d'autres études, les parties -*Ov*- dans -*Ovsk*- et -*in*- dans -*insk*- sont considérés comme des unités morphologiques à part entière. Ces travaux (Dement'ev, 1974, p.119) les rapprochent des suffixes d'appartenance -*in* et -*Ov*, discutés dans la section 2.2.1. Toutefois, cette corrélation n'est pas observée dans la langue russe contemporaine (Lopatin, 1977, p.43). D'autres études considèrent ces suffixes soit comme des *submorphes* (Lopatin, 1977, pp.57-63), c'est-à-dire des séquences qui n'ont pas de signification en elles-mêmes ou indépendantes de la signification ; soit comme des *morphes* (Švedova, 1980, p.123-125)[10].

Selon l'étude de Lopatin (1977, pp.41-57), il n'y a pas de restrictions méthodologiques pour considérer que les segments en question existent en dehors du suffixe principal et pour les séparer. Il serait donc approprié de les traiter comme faisant partie des thèmes (avoir la section 1.4.3) ou des affixes.

Pour distinguer si la séquence fait partie du thème ou du suffixe, il est nécessaire de déterminer si l'extension en question est régulière (Lopatin, 1977, pp.41-57). Si c'est le cas, elle fait partie de l'affixation, sinon, si elle reflète un caractère plus ou

[10]Dans cette terminologie, les *morphes* sont des unités linéaires trouvées à l'intérieur des mots, alors que les *morphèmes* sont des unités abstraites représentées par les *morphes*. Ainsi -*n*-, -*sk*- sont les morphèmes, -*enn*-, *estvenn*-, -*Ovsk*-, -*ičesk*- sont des morphes.

moins individuel des changements, cette séquence appartient au thème. Ainsi, dans (74) les extensions feraient partie du thème[11].

(74) ГЕРАНЬ ‘géranium’ → ГЕРАНИЕВЫЙ
GERAN’ GERANIEVYJ
БУРЖУА ‘bourgeois$_N$’ → БУРЖУАЗНЫЙ
BURŽUA BURŽUAZNYJ

En cas d’extension suffixale, l’apparition d’un segment supplémentaire est justifiée par des propriétés phonologiques, morphologiques ou sémantiques.

Certaines extensions peuvent modifier le sens des adjectifs par rapport aux formations avec un suffixe de base. Dans ces situations, le suffixe de base combiné à l’extension peut être considéré comme un suffixe à part entière. C’est le cas pour les suffixes *-tel’n-*, *-ičn-*, *-abel’n-* (75a) ou *-iansk-*/*-jansk-* (75b) (Vinogradov, 1952, p.167). Dans les exemples (75a), les adjectifs développent un sens qualificatif ; dans (75b), les adjectifs renvoient non pas à une personne, mais à un courant social, politique ou philosophique. En revanche, les suffixes *-sk-* et *-Ovsk-* (75c) renvoient tous les deux à la fois à une personne et à un courant : de ce point de vue, *-Ovsk-* ne peut pas être considéré comme une variante de *-sk-*.

(75) a. МУЧИТЬ ‘torturer’ → МУЧИТЕЛЬНЫЙ ‘pénible’
MUČIT’ MUČITEL’NYJ
ЭНЕРГИЯ ‘énergie’ → ЭНЕРГИЧНЫЙ ‘énergique’
ÈNERGIJA ÈNERGIČNYJ
РЕНТА ‘location’ → РЕНТАБЕЛЬНЫЙ ‘rentable’
RENTA RENTABEL’NYJ

b. КАНТ ‘Kant’ → КАНТИАНСКИЙ
KANT KANTIANSKIJ
ГЕГЕЛЬ ‘Hegel’ → ГЕГЕЛЬЯНСКИЙ
GEGEL’ GEGEL’JANSKIJ

c. ЛЕНИН ‘Lénine’ → ЛЕНИНСКИЙ
LENIN LENINSKIJ
ТОЛСТОЙ ‘Tolstoï’ → ТОЛСТОВСКИЙ
TOLSTOJ TOLSTOVSKIJ

D’autres extensions ne modifient pas la sémantique adjectivale mais se combinent avec les thèmes avec certaines particularités phonologiques et morphologiques. Dans certaines conditions, la dérivation avec un suffixe simple ne peut pas avoir lieu en raison des contraintes phonologiques ou morphologiques. D’après Arutjunova (1961, pp.54-64), les extensions sont utilisées pour éviter de violer ces contraintes. Il n’est

[11] Il existe cependant un point de vue selon lequel les discussions sur l’appartenance du segment au thème ou au suffixe sont superflues (Vinokur, 1959, pp.419-442).

pas approprié de considérer ces extensions comme des suffixes distincts, comme -*sk*-, -*esk*-, -*ičesk*-, etc. (76).

(76)

ИЮЛЬ 'juillet'	→	ИЮЛЬСКИЙ
IJUL'		IJUL'SKIJ
КУПЕЦ 'marchand'	→	КУПЕЧЕСКИЙ
KUPEC		KUPEČESKIJ
СЦЕНА 'scène'	→	СЦЕНИЧЕСКИЙ
SCENA		SCENIČESKIJ

Le choix entre le suffixe de base et le suffixe avec extension peut également être expliqué par les propriétés lexicales. La productivité de certains affixes est limitée par les bases ayant une certaine sémantique (Arutjunova, 1961, pp.54-64). Corbett et Comrie (2003, p.856) illustrent ce point avec le suffixe -*sk*- qui est utilisé pour dériver des adjectifs à partir des noms désignant des individus et des groupes d'individus. Les extensions de ce suffixe sont productives dans la dérivation d'adjectifs à partir des emprunts et dans le vocabulaire technique.

Contrairement aux suffixes -*n*-, -*sk*-, et -*Ov*- qui dérivent des adjectifs avec un sens générique 'relatif à ce que désigne le nom de base', les variantes étendues construisent des adjectifs à sémantique hétérogène. De plus, les adjectifs formés avec les trois suffixes principaux sont les plus fréquents, et le nombre limité d'adjectifs formés avec les variantes ne permet pas une analyse statistique robuste. Enfin, l'analyse des variantes impliquerait un grand nombre de comparaisons entre les différents suffixes disponibles. Par conséquent, notre étude se concentrera uniquement sur les trois suffixes principaux, -*n*-, -*sk*-, et -*Ov*- ; les variantes pourraient faire l'objet d'une recherche à part.

2.2.3 Sémantique

Nous avons brièvement abordé la question de la sémantique des variantes suffixales dans la section précédente (2.2.2) et avons mentionné que les suffixes -*n*-, -*sk*-, et -*Ov*- dérivent des adjectifs avec le sens générique 'relatif à ce que désigne le nom de base', le point de vue que nous adoptons dans la présente recherche. Cette section se consacrera aux différentes interprétations de la sémantique des adjectifs dénominaux, qualificatifs et relationnels.

La morphologie dérivationnelle implique des opérations sur trois niveaux : la forme, la catégorie, et le sens. Les questions relatives à la forme ont été discutées dans la section 1.4 sur les allomorphies thématiques. Nous allons nous concentrer sur le sens ici.

Il existe deux approches pour traiter le sens des mots construits : une approche morphémique qui sous-entend que la sémantique du mot dérivé se compose de la sémantique des morphèmes qui le constituent ; l'approche lexicale, à son tour, met au centre de l'étude le mot (lexème) en tant que tel, c'est le lexème qui représente

une unité minimale de sens ; les règles morphologiques sont applicables aux mots et non aux morphèmes (Halle, 1973 ; Matthews, 1974 ; Aronoff, 1976 ; Anderson, 1992 ; Scalise, 2011). Dans l'approche lexicale, les affixes n'ont pas de sens fixe, ce sens est défini et concrétisé pour chaque construction individuelle (Aronoff, 1976). En suivant Aronoff (2007, pp.803-806), nous adopterons le point de vue selon lequel les lexèmes morphologiquement construits ont un sens idiosyncratique qui ne peut être accédé qu'au travers d'un contexte ; leur sémantique ne peut pas être analysée *a priori* par le biais de la décomposition.

Comme le souligne Aronoff (2019), les affixes dérivationnels présentent des particularités sémantiques et pragmatiques. Les mots dérivés à l'aide de ces affixes peuvent avoir une grande variété de sens, non seulement en raison des glissements sémantiques, mais également parce que les règles morphologiques en question ne sont pas restrictives ni sémantiquement, ni pragmatiquement. Cependant, il existe des cas où un suffixe ne sert qu'à construire un lexème d'une catégorie lexicale différente de celle du lexème de base, sans apporter de changements au niveau sémantique.

Dans la section 2.1.1, nous avons mis en évidence qu'à la différence des adjectifs de qualité en russe, qui peuvent être simples ou construits, les adjectifs de relation sont généralement construits. En ce qui concerne les adjectifs de qualité dérivés, dans la grande majorité des cas, le processus de dérivation ajoute une complexité sémantique (77)[12]. Ainsi, les adjectifs qualificatifs dérivés n'établissent pas de relations entre le nom recteur et le nom de base, ils attribuent au nom recteur une propriété saillante incluse dans la sémantique du nom de base.

(77)

ГЛАЗ 'oeil'	→	ГЛАЗАСТЫЙ 'aux grands yeux'
GLAZ		GLAZASTYJ
БОРОДА 'barbe'	→	БОРОДАТЫЙ 'barbu'
BORODA		BORODATYJ

La dérivation des adjectifs qualificatifs se base sur un transfert métonymique (Roché, 2006, pp.383-384), où le nom sert à désigner une caractéristique d'une catégorie, et sur un rapport analogique. Ainsi, l'adjectif qualitatif utilise le nom de base pour exprimer une qualité, mais il ne renvoie pas directement à ce nom, contrairement aux adjectifs de relation.

La dérivation des adjectifs de relation présente des particularités supplémentaires, car l'adjectif dérivé peut avoir des valeurs ajoutées, telles que l'appartenance[13] (78a), traitée dans la section 2.2.1, ou une valeur sémantique qui ne peut être concrétisée

[12]Les adjectifs formés avec le suffixe *-ast-* expriment une intensification d'un trait externe présent dans le nom de base et sont souvent associés à une connotation familière. Les adjectifs formés avec le suffixe *-at-* ont une signification 'caractérisé par un trait externe que le nom de base désigne'. Ce type d'adjectifs se distingue du précédent par l'absence d'intensité ainsi que de connotation familière dans la plupart des cas (Švedova, 1980, p.284).

[13]Pour d'autres valeurs pouvant être exprimées par les adjectifs de relation, voir notamment Roché (2006) pour le français, Rainer (2013) pour l'allemand.

qu'en contexte (78b-78c). Il s'agit de relations sémantiquement marquées dans le premier cas et de relations non marquées dans le deuxième.

(78) a. *иваново детство* 'l'enfance d'Ivan'
ivanovo detstvo

b. *книжный магазин* 'magasin où l'on vend des livres'
knižnyj magazin

c. *книжный клуб* 'club où l'on lit des livres [club de lecture]'
knižnyj klub

Dans les cas (77) et (78a), l'ajout d'un suffixe entraîne la formation d'un dérivé ayant un sens précis : dans (77), caractérisé par une intensité forte d'une propriété désignée par le nom de base ou caractérisé par une propriété désignée par le nom de base ; dans (78a), appartenant à celui désigné par le nom de base. Cependant, l'ajout du suffixe -*n*- comme dans (78b-78c) ne génère pas toujours un sens de lieu de vente ou de lieu de lecture, et n'entraîne pas toujours une sémantique spatiale.

En ce qui concerne les adjectifs de relation, le type de relation désigné dépend fortement des deux substantifs utilisés: le nom de base et le nom recteur.

En général, les adjectifs dénominaux formés à l'aide des suffixes -*n*-, -*sk*-, -*Ov*- ou leurs variantes ont une sémantique générale de '{relatif à, lié à, appartenant à, destiné à, se trouvant dans, fait de, fonctionnant à l'aide de} *X*'. Cependant, l'inventaire de sens que ces adjectifs peuvent exprimer, ainsi que les adjectifs de relation généralement, est très important, comme le notent Apresjan (1974), Marchand (1960) et Aronoff (2019). Ces auteurs observent que la polysémie des adjectifs de relation peut être illimitée. Les adjectifs relationnels peuvent prendre toutes les valeurs sémantiques possibles, selon Roché (2006, p. 380). Le sens de l'adjectif sera alors déterminé par celui du nom recteur, mais ne sera pas déductible *a priori*. Ainsi, dans l'exemple (79)[14], ni l'adjectif relationnel *замочная* ni le nom *скважина* n'expriment la relation de but.

(79) *замочная скважина* 'trou de serrure'
zamočnaja skvažina

La polysémie exhaustive ne permet pas de dresser *a priori* une liste détaillée et précise de tous les sens possibles de ces adjectifs. En fin de compte, ils sont plus proches de la monosémie car ils n'ont qu'un seul sens générique, qui est : 'relatif à *X*' (Zemskaja, 2015, pp.230-233 ; Uluxanov, 1977, p.92). Par conséquent, le sens de ces adjectifs se concrétise uniquement dans le contexte, en fonction des noms recteurs et de l'expérience subjective de chaque locuteur.

Les suffixes qui forment les adjectifs de relation sont dépourvus de sens propre, leur unique fonction est de transformer un nom en un adjectif. Ces adjectifs relationnels

[14]Exemple tiré de Mezhevich (2002).

remplacent les noms sans altérer leur valeur sémantique (Bally, 1944, p.97)[15]. Ainsi, on peut les appeler *noms convertis* (McNally et Boleda, 2004, pp.181-183) ou *pseudo-adjectifs* (Postal, 1969), ou dérivés syntaxiques (Zemskaja, 2011, pp.192-200 ; Zemskaja, 2015, p.207)[16]. Comme l'opération sémantique n'est pas activée, la formation de ce type d'adjectifs est également appelée *adjectivation non marquée* par Maurel (1993, p.24). Les suffixes *-n-*, *-sk-*, *-Ov-* peuvent donc être à usage multiple.

Comme mentionné précédemment, les dérivés qualificatifs possèdent un sens concret prédéfini, indépendant du contexte et du nom qu'ils modifient.

En guise de conclusion, nous pouvons affirmer que les adjectifs de relation ont la même sémantique que le nom de base, mais cette sémantique nominale est exprimée par une autre catégorie grammaticale. Ces adjectifs se rapprochent des constructions syntaxiques (80).

(80) *леч-ить* *грязь-ю* 'traiter avec de la boue'
leč-it' *grjaz'-ju*
traiter-INF boue-F.INS.SG
∼*грязев-ое* *лечени-е* 'traitement par la boue'
grjazev-oe *lečeni-e*
boue$_{\text{REL}}$-N.NOM.SG traitement-N.NOM.SG
колес-о *автомобил-я* 'roue de voiture'
koles-o *avtomobil-ja*
roue-N.NOM.SG voiture-M.GEN.SG
∼*автомобильн-ое* *колес-о* 'roue de voiture'
avtomobil'n-oe *koles-o*
voiture$_{\text{REL}}$-N.NOM.SG roue-N.NOM.SG
ручк-а *двер-и* 'poignée de porte '
ručk-a *dver-i*
poignée-F.NOM.SG porte-F.GEN.SG
∼*двер-ая* *ручк-а* 'poignée de porte '
dver-aja *ručk-a*
porte$_{\text{REL}}$-F.NOM.SG poignée-F.NOM.SG

2.3 Les suffixes *-n-*, *-sk-*, *-ov-*

Nous avons exposé les propriétés morphologiques, syntaxiques et sémantiques des adjectifs en général, ainsi que celles des adjectifs qualificatifs et relationnels. Dans cette section, nous nous concentrons spécifiquement sur les propriétés des adjectifs construits avec les suffixes *-n-*, *-sk-*, et *-Ov-*.

[15] L'idée que l'adjectif de relation est dépourvue de sens a été notamment popularisée par Bally (1944, p.116) : l'adjectif change de valeur grammaticale sans changer la valeur sémantique.

[16] Les adjectifs de qualité sont caractérisés par une dérivation dite lexicale, lorsque le niveau sémantique est modifié. Dans la dérivation dite syntaxique, la catégorie grammaticale est modifiée mais pas le niveau sémantique.

2.3.1 Propriétés morphologiques et syntaxiques

Dans la section 2.1.1, nous avons énuméré les propriétés morphologiques et dérivationnelles des adjectifs prototypiques en russe, telles que l'emploi en attribut et épithète, la présence de formes longues et courtes, du comparatif synthétique, et la dérivation d'adverbes et de noms abstraits. Dans cette section, nous examinerons les propriétés des suffixes *-n-*, *-sk-*, et *-Ov-* en relation à ces critères.

Selon Graščenkov (2022, pp.49-71), les suffixes *-sk-* et *-Ov-* forment des adjectifs qui peuvent être utilisés uniquement en tant qu'épithètes, avec des propriétés morphologiques et dérivationnelles considérablement plus restreintes que celles des adjectifs formés à l'aide du suffixe *-n-*. Cela s'explique par le fait que, dans certains cas, les adjectifs construits avec le suffixe *-n-* peuvent acquérir un sens qualificatif (81).

(81) ДРУГ 'ami' → ДРУЖНЫЙ 'amical, uni'
DRUG DRUŽNYJ
ЧЕЛОВЕК 'homme' → ЧЕЛОВЕЧНЫЙ 'humain'
ČELOVEK ČELOVEČNYJ

La forme courte (82a) et le comparatif (82b) sont habituellement disponibles pour les adjectifs en *-n-*, mais pas pour ceux en *-sk-* ou *-Ov-*.

(82) a. ЧЕЛОВЕЧНЫЙ 'humain' : *человечен* 'humain$_{\text{M.NOM.SG}}$ (court)'
ČELOVEČNYJ *čelovečen*
КОНСКИЙ 'cheval$_{\text{REL}}$' : **конск* 'cheval$_{\text{REL.M.NOM.SG}}$ (court)'
KONSKIJ *konsk*
ГОДОВОЙ 'annuel' : **годов* 'annuel$_{\text{M.NOM.SG}}$ (court)'
GODOVOJ *godov*

b. ЧЕЛОВЕЧНЫЙ 'humain' : *человечн-ее* 'humain-COMP'
ČELOVEČNYJ *čelovečn-ee*
КОНСКИЙ 'cheval$_{\text{REL}}$' : **конск-ее* 'cheval$_{\text{REL}}$-COMP'
KONSKIJ *konsk-ee*
ГОДОВОЙ 'annuel' : **годов-ее* 'annuel-COMP'
GODOVOJ *godov-ee*

En ce qui concerne la dérivation d'adverbes, les adjectifs formés à l'aide du suffixe *-n-* sont facilement transformables en adverbes, tandis que les adjectifs formés à l'aide des suffixes *-sk-* et *-Ov-* ne le sont jamais (83).

(83) ЧЕЛОВЕЧНЫЙ 'humain' → ЧЕЛОВЕЧНО 'humainement'
ČELOVEČNYJ ČELOVEČNO
CHEVAL 'humain' → *КОНСКИ
KONSKIJ KONSKI
ГОДОВОЙ 'annuel' → *ГОДОВО
GODOVOJ GODOVO

Les adverbes en *-ski* sont admis tels quels dans certains contextes : lorsqu'ils modifient un adjectif (84a) ou un verbe (84b), ou lorsque la formation d'un adverbe est accompagnée par l'ajout simultané d'un préfixe *po-* (84c). Il existe des formations parallèles utilisant le suffixe *-n-* pour les cas décrits dans (84a) et (84b), tels que ФАНТАСТИЧНО (FANTASTIČNO), ТРАГИЧНО (TRAGIČNO).

(84) a. *эти хоккеисты фантастически хороши*
èti xokkeisty fantastičeski xoroši
'ces joueurs de hockey sont fantastiquement bons'

b. *его жизнь трагически оборвалась*
ego žizn' tragičeski oborvalas'
'sa vie s'est interrompue tragiquement'

c. ЧЕЛОВЕЧЕСКИЙ 'humain' → ПО-ЧЕЛОВЕЧЕСКИ 'humainement'
ČELOVEČESKIJ PO-ČELOVEČESKI

La formation de noms abstraits en utilisant le suffixe *-ost'* est productive pour les adjectifs construits avec le suffixe *-n-*, toutefois, pour les adjectifs en *-sk-* ou *-Ov-*, cette forme de dérivation est très restreinte (85).

(85) ЧЕЛОВЕЧНЫЙ 'humain' → ЧЕЛОВЕЧНОСТЬ 'humanité'
ČELOVEČNYJ ČELOVEČNOST'
КОНСКИЙ 'cheval' → *КОНСКОСТЬ
KONSKIJ KONSKOST'
ГОДОВОЙ 'annuel' → *ГОДОВОСТЬ
GODOVOJ GODOVOST'

Les cas étudiés montrent que les suffixes *-Ov-* et *-sk-* peuvent former des adjectifs utilisés comme épithètes. Le suffixe *-n-* offre une plus grande variété de propriétés morphologiques, dérivationnelles et syntaxiques pour les adjectifs (présence de formes courtes et de comparatifs, formation d'adverbes et de noms abstraits). En conséquence, ce suffixe permet de former des adjectifs plus prototypiques, avec un inventaire de propriétés plus étendu que celui des adjectifs formés avec les suffixes *-sk-* et *-Ov-*.

2.3.2 Propriétés sémantiques

Dans la section 2.1, nous avons examiné la sémantique des différents types d'adjectifs et constaté qu'une partie d'entre eux peut avoir une sémantique précise et prédéfinie en fonction du suffixe utilisé pour les former. La sémantique des autres adjectifs dépend fortement du nom de base, du nom recteur et du contexte dans lequel ils sont utilisés. Les adjectifs russes formés avec les suffixes *-n-*, *-sk-* et *-Ov-* appartiennent à ce dernier groupe, et ont des liens variés et complexes avec leurs noms de base.

Le sens d'un adjectif se concrétise généralement dans le contexte, lorsque l'adjectif modifie différents noms de base. Les exemples d'adjectifs formés à partir de КИРПИЧ 'brique' sont présentés en (86) avec leur contexte respectif.

(86) *кирпичный дом* ‘maison [faite] en briques’
kirpičnyj dom
кирпичный завод ‘usine [de production] de briques’
kirpičnyj zavod
кирпичный цвет ‘couleur brique’
kirpičnyj cvet

Selon Zemskaja (2015, pp.230-233), un adjectif avec le suffixe -*n*-, en l’occurrence КИРПИЧНЫЙ ‘brique’, peut avoir une variété de sens. Cependant, les adjectifs en -*n*- ne peuvent pas exprimer une relation impliquant une personne ou un être animé (Vinogradov et Švedova, 1964, pp.343, 350-354).

Švedova (1980, pp.269-270) recense plusieurs significations distinctes des adjectifs formés avec le suffixe -*n*- :

- Propre ou appartenant à X
картофельная ботва ‘tiges de pomme de terre’,
kartofel’naja botva
колхозные поля ‘champs collectifs’ ;
kolxoznye polja

- Composé de, fait de X
кирпичный дом ‘maison en brique’,
kirpičnyj dom
бархатное платье ‘robe en velours’ ;
barxatnoe plat’e

- Ayant, contenant X
яблочный пирог ‘tarte aux pommes’,
jabločnyj pirog
парусное судно ‘navire à voiles’ ;
parusnoe sudno

- Destiné à X
книжный шкаф ‘bibliothèque’,
knižnyj škaf
праздничное платье ‘robe de fête’ ;
prazdničnoe plat’e

- Se trouvant dans l’endroit désigné par X
древесный червь ‘ver de bois’,
drevesnyj červ’
южные города ‘villes du Sud’ ;
južnye goroda

- Existant dans la période désignée par X
  ночные страхи ‘peurs nocturnes’,
  nočnye straxi
  воскресная прогулка ‘promenade du dimanche’ ;
  voskresnaja progulka

- Lié à la production de X
  молочная ферма ‘ferme laitière’,
  moločnaja ferma
  консервная промышленность ‘industrie de la conserve’ ;
  konservnaja promyšlennost’

- Ressemblant à, évoquant X
  стальные мускулы ‘muscles d’acier’,
  stal’nye muskuly
  изумрудная трава ‘herbe d’émeraude’.
  izumrudnaja trava

Le nom de base seul n’est pas toujours suffisant pour déterminer le sens de l’adjectif dérivé. Dans certains cas, un contexte plus large est nécessaire. Ainsi, Zemskaja donne l’exemple de *автомобильные деньги* (*avtomobil’nye den’gi*) ‘argent de voiture’ qui, selon le contexte, peut prendre des significations différentes : cela peut correspondre à l’argent mis de côté pour l’achat d’une voiture, l’argent reçu suite à la vente d’une voiture, l’argent retrouvé dans une voiture, l’argent oublié dans une voiture, etc.

Il est important de considérer le sens des adjectifs en fonction de la pragmatique et du contexte. Un exemple donné par Zemskaja (2015, pp.230-233) montre qu’un adjectif peut avoir un sens spécifique qui ne peut pas être déterminé uniquement à partir du nom de base : *Как твоя капустная нога?* (*Kak tvoja kapustnaja noga*) ‘Comment va ta jambe de chou ?’ renvoie à une jambe blessée lors de la récolte du chou. Selon Uluxanov (1977, pp.105-107), il est possible de distinguer entre les noms recteurs un noyau dur utilisé dans un discours stylistiquement neutre et les noms recteurs périphériques utilisés dans un discours stylistiquement marqué. En exemple, les adjectifs ОВОЩНОЙ (OVOŠČNOJ) ‘légume’ ou МУСОРНЫЙ (MUSORNYJ) ‘poubelle’ modifient typiquement les noms de base qui désignent des objets inanimés (87a) ; les combinaisons avec les noms d’actions sont moins courantes (87b).

(87) a. *овощной суп* ‘soupe de légumes’
  ovoščnoj sup
  овощной магазин ‘magasin de légumes’
  ovoščnoj magazin
  мусорный пакет ‘sac poubelle’
  musornyj paket
  мусорная яма ‘fosse à ordures’
  musornaja jama

b. *Для московских речников открылась* ***овощная*** *навигация*
Dlja moskovskix rečnikov otkrylas' ***ovoščnaja*** *navigacija*
'La navigation [transport fluvial] **de légumes** a été ouverte pour les mariniers moscovites'
*Следовало давно произвести «****мусорный*** *аврал» – повыбрасывать всё ненужное*
*Sledovalo davno proizvesti «****musornyj*** *avral» – povybrasyvat' vsë nenužnoe*
'Cela fait longtemps qu'il aurait fallu organiser une urgence **ordures** – jeter tout ce qui n'est plus utile'

Malgré le nombre de significations possibles illimité des adjectifs, ceux formés avec le suffixe *-n-* sont le plus souvent dotés d'une signification qualificative et lexicalisée (88) (Lopatin, 1977, p.99 ; Uluxanov, 1977, pp.105-107).

(88) УМ 'esprit' → УМНЫЙ 'intelligent'
UM UMNYJ
ВЕРА 'foie' → ВЕРНЫЙ 'fidèle'
VERA VERNYJ
СИЛА 'force' → СИЛЬНЫЙ 'fort'
SILA SIL'NYJ

Les adjectifs formés avec le suffixe *-sk-* ont généralement un sens 'relatif à, propre à *X*' (Švedova, 1980, pp.278-279). Les noms de base peuvent désigner des individus (89a), des objets (89b), des groupes ethniques (89c), des toponymes (89d), des noms propres (89e), des noms d'organisations, de sociétés ou d'institutions, etc. (89f).

(89) a. ПРЕПОДАВАТЕЛЬ 'enseignant' → ПРЕПОДАВАТЕЛЬСКИЙ
PREPODAVATEL' PREPODAVATEL'SKIJ

b. УНИВЕРСИТЕТ 'université' → УНИВЕРСИТЕТСКИЙ
UNIVERSITET UNIVERSITETSKIJ

c. ГОТ 'gothique$_N$' → ГОТСКИЙ
GOT GOTSKIJ

d. Париж 'Paris' → ПАРИЖСКИЙ
Pariž PARIŽSKIJ

e. Дарвин 'Darwin' → ДАРВИНСКИЙ
Darvin DARVINSKIJ

f. ПАРЛАМЕНТ 'parlement' → ПАРЛАМЕНТСКИЙ
PARLAMENT PARLAMENTSKIJ

Les adjectifs formés à partir des noms désignant les humains et ayant le suffixe *-sk-* peuvent exprimer une grande variété de sens, allant de l'appartenance à une personne définie (90a) à l'appartenance à un courant idéologique précis (90b).

(90) a. БРАТ ‘frère’ → БРАТСКИЙ
BRAT BRATSKIJ
b. Ленин ‘Lénine’ → ЛЕНИНСКИЙ
Lenin LENINSKIJ

Dans le cas présenté en (88), il est question de la lexicalisation des adjectifs formés à l’aide du suffixe *-n-*. Toutefois, ce phénomène peut également se produire pour les adjectifs formés à l’aide du suffixe *-sk-* (91).

(91) БРАТ ‘frère’ → БРАТСКИЙ ‘fraternel’
BRAT BRATSKIJ
АНГЕЛ ‘ange’ → АНГЕЛЬСКИЙ ‘angélique’
ANGEL ANGEL’SKIJ
ДЬЯВОЛ ‘diable’ → ДЬЯВОЛЬСКИЙ ‘diabolique’
D’JAVOL D’JAVOL’SKIJ

Le suffixe *-sk-* se combine généralement avec des noms de base animés (avec un sens générique ‘relatif à un être animé ou à un groupe d’êtres animés’) ou des toponymes, selon l’étude de Zemskaja (2011, pp.253-256). Il diffère du suffixe *-n-* en donnant un sens plus restreint à l’adjectif formé. Le suffixe *-Ov-* se rapproche davantage du suffixe *-n-* quant à l’inventaire de noms de base avec lesquels il peut se combiner.

Les adjectifs formés avec le suffixe *-Ov-* ont un sens général ‘qui se réfère à, propre à *X*’. Les noms de base utilisés pour former ces adjectifs sont généralement non animés et concrets (92a) ; ils peuvent également désigner des animaux (92b) et des humains (92c).

(92) a. ПОЛЕ ‘champs’ → ПОЛЕВОЙ
POLE POLEVOJ
b. ТИГР ‘tigre’ → ТИГРОВЫЙ
TIGR TIGROVYJ
c. СТРЕЛОК ‘tireur’ → СТРЕЛКОВЫЙ
STRELOK STRELKOVYJ

Les adjectifs formés à l’aide du suffixe *-Ov-*, tout comme ceux formés à l’aide du suffixe *-n-*, peuvent développer des sens très variés et concrets, en fonction du contexte et du nom recteur qu’ils modifient. L’exemple en (93) montre des adjectifs formés à partir du nom de base совок (sovok) ‘pelle’.

(93) *совковая рукоятка* ‘poignée de [appartenant à une] pelle’
sovkovaja rukojatka
совковая лопата ‘bêche ressemblant à une pelle’
sovkovaja lopata

Comme nous l'avons mentionné plus haut, l'expression du sens qualificatif n'est pas un privilège exclusif du suffixe *-n-*. Ce sens peut être également exprimé, bien que plus rarement, par des adjectifs relationnels déjà existants avec le suffixe *-sk-* (94a) ou *-Ov-* (94b) selon Zemskaja (2000, pp.125-126).

(94) a. *нуворишское высокомерие*
nuvorišskoe vysokomerie
'orgueil de nouveau-riche [comme chez les nouveaux-riches]'

b. *самые рейтинговые политические программы*
samye rejtingovye političeskie programmy
'émissions politiques ayant les meilleurs scores'

La disponibilité des trois suffixes en question pour former les adjectifs dénominaux peut entraîner l'apparition de doublets (95a), voire même de triplets (95b), lorsque différents suffixes sont ajoutés à la même base nominale.

(95) a. КОНЬ 'cheval' → КОННЫЙ / КОНСКИЙ
KON' KONNYJ KONSKIJ

b. СЛЕСАРЬ 'serrurier' →
SLESAR'
СЛЕСАРНЫЙ / СЛЕСАРСКИЙ / СЛЕСАРЕВЫЙ
SLESARNYJ SLESARSKIJ SLESAREVYJ

Les doublets suffixaux en *-n-*, *-sk-* et *-Ov-* peuvent avoir des significations similaires et différentes. Bien que tous ces suffixes aient un sens relationnel générique, certains types de relations ne peuvent être exprimés qu'à l'aide d'un seul suffixe. Le suffixe *-n-* sert à transmettre un sens relationnel et qualificatif. Les liens entre l'adjectif et le nom recteur peuvent ainsi être très différents. Comme nous l'avons constaté plus haut, ce suffixe ne sert pas à transmettre de relation à une personne ou un groupe de personnes, c'est le suffixe *-sk-* qui s'en charge. Le suffixe *-Ov-*, quant à lui, transmet un sens uniquement relationnel, selon Uluxanov (1977, pp.204-206).

Un adjectif formé avec le suffixe *-n-* possède le plus souvent un sens lexicalisé, tandis qu'un adjectif formé avec le suffixe *-Ov-* possède un sens relationnel générique. Les adjectifs présentés dans (96a) peuvent avoir des significations différentes en fonction du contexte. Cependant, le sens des adjectifs présentés dans (96b) est toujours constant et identique dans tous les contextes.

(96) a. ВКУС 'goût' → ВКУСОВОЙ
VKUS VKUSOVOJ
ГРЯЗЬ 'saleté' → ГРЯЗЕВОЙ
GRJAZ' GRJAZEVOJ
ЖИР 'gras' → ЖИРОВОЙ
ŽIR ŽIROVOJ

b. ВКУС ‘goût’ → ВКУСНЫЙ ‘délicieux’
VKUS VKUSNYJ
ГРЯЗЬ ‘saleté’ → ГРЯЗНЫЙ ‘sale’
GRJAZ’ GRJAZNYJ
ЖИР ‘gras’ → ЖИРНЫЙ ‘gros’
ŽIR ŽIRNYJ

L’utilisation des suffixes -*n*- et -*Ov*- sert principalement dans le langue moderne pour la différenciation sémantique ou pour éviter et prévenir l’homonymie. Ainsi, selon Vinogradov (1952, pp.178-179), les adjectifs formés avec le suffixe -*n*- ont souvent un usage familier et sont plus proches des constructions avec le génitif (97a), tandis que les adjectifs formés avec le suffixe -*Ov*- ne le sont pas (97b)[17].

(97) a. *керосинн-ый* *запах* ‘odeur de kérosène’
kerosinn-yj *zapax*
kérosène$_{\text{REL}}$-M.NOM.SG odeur$_{\text{M.NOM.SG}}$
∼*запах* *керосин-а* ‘odeur de kérosène’
zapax *kerosin-a*
odeur$_{\text{M.NOM.SG}}$ kérosène-M.GEN.SG
керосинн-ая *бутылк-а* ‘bouteille de kérosène’
kerosinn-aja *butylk-a*
kérosène$_{\text{REL}}$-F.NOM.SG bouteille-F.NOM.SG
∼*бутылк-а* *керосин-а* ‘bouteille de kérosène’
butylk-a *kerosin-a*
bouteille-F.NOM.SG kérosène-M.GEN.SG

b. *керосинов-ое* *освещен-ие* ‘lumière de kérosène’
kerosinov-oe *osveščen-ie*
kérosène$_{\text{REL}}$-N.NOM.SG lumière-N.NOM.SG
**освещен-ие* *керосин-а* ‘lumière de kérosène’
osveščen-ie *kerosin-a*
lumière-N.NOM.SG kérosène-M.GEN.SG
керосинов-ый *двигатель* ‘moteur de kérosène’
kerosinov-yj *dvigatel’*
kérosène$_{\text{REL}}$-M.NOM.SG moteur$_{\text{M.NOM.SG}}$
**двигатель* *керосин-а* ‘moteur de kérosène’
dvigatel’ *kerosin-a*
moteur$_{\text{M.NOM.SG}}$ kérosène-M.GEN.SG

De plus, l’emploi des adjectifs avec des suffixes différents, notamment les doublets, peut varier selon les styles. Il est alors nécessaire d’examiner les aspects discursifs.

[17]Si ces constructions avec un adjectif ne peuvent pas être remplacées par des constructions avec le génitif, d’autres cas sont possibles : *освещение керосином* (*osveščenie kerosinom*), instrumental ; *двигатель на керосине* (*dvigatel’ na kerosine*), locatif.

2.3.3 Pragmatique et discours

Les particularités d'utilisation des adjectifs dénominaux en fonction de différents styles ont été étudiés par Zemskaja (2015, pp.208-218, 223-230). Cet auteur remarque que, premièrement, l'utilisation des adjectifs a une histoire plus ancienne que celle des syntagmes prépositionnels homonymes ; et, deuxièmement, que les adjectifs dénominaux ne sont pas utilisés de manière homogène.

Dans le style littéraire, ces adjectifs ne dénotent que l'ensemble de relations restreintes (appartenance, métonymie, matériel, finalité, etc.). La traduction de relations de lieu n'est pas typique et est marquée par le caractère individuel dans le contexte. Ainsi, les premières constructions en (98a-98c) sont acceptables, tandis que les deuxièmes constructions, leurs analogues, sont acceptables uniquement dans certaines conditions et dans certains contextes.

(98) a. *поездк-а* *в* *Крым*
poezdk-a *v* *Krym*
voyage-F.NOM.SG PREP Crimée$_{\text{M.ACC.SG}}$
'voyage en Crimée'
?*крымск-ая* *поездк-а*
krymsk-aja *poezdk-a*
Crimée$_{\text{REL}}$-F.NOM.SG voyage-F.NOM.SG
'voyage en Crimée'

b. *путешеств-ие* *вокруг* *Европ-ы*
putešestv-ie *vokrug* *Evrop-y*
voyage-N.NOM.SG PREP Europe-F.GEN.SG
'voyage autour d'Europe'
?*европейск-ое* *путешеств-ие*
evropejsk-oe *putešestv-ie*
européen-N.NOM.SG voyage-N.NOM.SG
'voyage autour d'Europe'

c. *поход* *по* *Кавказ-у*
pohod *po* *Kavkaz-u*
randonnée$_{\text{M.NOM.SG}}$ PREP Caucase-M.DAT.SG
'randonnée au Caucase'
?*кавказск-ий* *поход*
kavkazsk-ij *pohod*
caucasien-M.NOM.SG randonnée$_{\text{M.NOM.SG}}$
'randonnée au Caucase'

La même chose s'applique à l'expression des relations d'instrument ou de manière : ce type de relation est traduit par des noms à l'instrumental ; l'utilisation des adjectifs dénominaux est difficilement acceptable (99a-99b).

(99) a. *рисован-ие* *масл-ом*
risovan-ie *masl-om*
peinture-N.NOM.SG huile-N.INS.SG
'peinture à l'huile'
?*масленн-ое* *рисован-ие*
maslenn-oe *risovan-ie*
huileux-N.NOM.SG peinture-N.NOM.SG
'peinture à l'huile'

b. *уборк-а* *хлеб-а* *комбайн-ом*
ubork-a *xleb-a* *kombajn-om*
récolte-F.NOM.SG pain-M.GEN.SG moissonneuse-batteuse-M.INS.SG
'récolte du pain avec une moissonneuse-batteuse'
?*комбайнов-ая* *уборк-а* *хлеб-а*
kombajnov-aja *ubork-a* *xleb-a*
moissonneuse-batteuse$_{\text{REL}}$-F.NOM.SG récolte-F.NOM.SG pain-M.GEN.SG
'récolte du pain avec une moissonneuse-batteuse'

Cependant, l'utilisation des adjectifs au lieu de syntagmes nominaux est très courante dans la terminologie, les professionnalismes, le discours familier et parlé, et le style poétique. Selon Zemskaja, la conservation de l'utilisation des adjectifs dans des constructions terminologiques est expliquée par deux facteurs : ce type de constructions est devenu figé et ces constructions servent d'exemples pour de nouvelles formations qui se caractérisent par le même type de relations sémantiques entre le nom recteur et l'adjectif dénominal. Ainsi, la deuxième construction adjectivale en (99b) est présente et acceptable dans la terminologie (100a). Ici, l'adjectif exprime une relation d'instrument. De la même manière, dans la terminologie, l'adjectif peut exprimer une relation de lieu (100b), de sujet (100c) ou d'objet (100d) de l'action, des relations de métonymie (100e), de manière (100f), et des relations temporelles (100g).

(100) a. *комбайнов-ая* *уборк-а* *хлеб-а*
kombajnov-aja *ubork-a* *xleb-a*
moissonneuse-batteuse$_{\text{REL}}$-F.NOM.SG récolte-F.NOM.SG pain-M.GEN.SG
'récolte du pain avec une moissonneuse-batteuse'
∼*уборк-а* *хлеб-а* *комбайн-ом*
ubork-a *xleb-a* *kombajn-om*
récolte-F.NOM.SG pain-M.GEN.SG moissonneuse-batteuse-M.INS.SG
'récolte du pain par une moissonneuse-batteuse'

b. *клеточн-ое* *содержан-ие*
kletočn-oe *soderžan-ie*
cellulaire-N.NOM.SG contenu-N.NOM.SG
'contenu cellulaire'
∼*содержан-ие* *в* *клетк-ах*
soderžan-ie *v* *kletk-ax*
contenue-N.NOM.SG PREP cellule-F.LOC.PL

‘contenu dans les cellules’

c. *опухолев-ый* *рост*
opuxolev-yj *rost*
tumorale-M.NOM.SG croissance$_{\text{M.NOM.SG}}$
‘croissance tumorale’
∼*рост* *опухол-и*
rost *opuxol-i*
croissance$_{\text{M.NOM.SG}}$ tumeur-F.GEN.SG
‘croissance de la tumeur’

d. *траекторн-ое* *изменен-ие*
traektorn-oe *izmenen-ie*
trajectoire$_{\text{REL}}$-N.NOM.SG changement-N.NOM.SG
‘changement de trajectoire’
∼*изменен-ие* *траектор-ии*
izmenen-ie *traektor-ii*
changement-N.NOM.SG trajectoire-F.GEN.SG
‘changement de trajectoire’

e. *маятников-ые* *час-ы*
majatnikov-ye *čas-y*
pendule$_{\text{REL}}$-M.NOM.PL horloge-M.NOM.PL
‘horloge à pendule’
∼*час-ы* *с* *маятник-ом*
čas-y *s* *majatnik-om*
horloge-M.NOM.PL PREP pendule-M.INS.SG
‘horloge avec une pendule’

f. *фазов-ый* *анализ*
fazov-yj *analiz*
phase$_{\text{REL}}$-M.NOM.SG analyse$_{\text{M.NOM.SG}}$
‘analyse par phases’
∼*анализ* *фаз-ами*
analiz *faz-ami*
analyse$_{\text{M.NOM.SG}}$ phase-F.INS.PL
‘analyse par phases’

g. *дневн-ое* *дежурств-о*
dnevn-oe *dežurstv-o*
journalier-N.NOM.SG service-N.NOM.SG
‘service journalier’
∼*дежурств-о* *дн-ём*
dežurstv-o *dn-ëm*
service-N.NOM.SG jour-M.INS.SG
‘service de jour’

Dans le style familier ou parlé, les constructions adjectivales sont souvent remplacées

par un seul substantif (101a-101c). La prédominance des constructions adjectivales dans la terminologie peut être expliquée par leur homogénéité : ainsi, les constructions types ont toujours un nom recteur et un adjectif qui le modifie (Zemskaja, 2015, pp.223-230).

(101) a. *ливнев-ый* *дождь* ∼ ЛИВЕНЬ ‘averse’
livnev-yj *dožd’* *liven’*
averse$_{\text{REL}}$-M.NOM.SG pluie$_{\text{M.NOM.SG}}$
‘averse de pluie’

b. *рассадн-ые* *культур-ы* ∼ РАССАДА ‘semis’
rassadn-ye *kul’tur-y* RASSADA
semis$_{\text{REL}}$-F.NOM.PL culture-F.NOM.PL
‘cultures de semis’

c. *транспортн-ое* *средств-о* ∼ ТРАНСПОРТ ‘transport’
transportn-oe *sredstv-o* TRANSPORT
transport$_{\text{REL}}$-N.NOM.SG moyen$_{\text{N}}$-N.NOM.SG
‘moyen de transport’

Le style parlé est cependant caractérisé par la présence d’adjectifs dénominaux qui expriment une variété de relations dans différents contextes : relations métonymiques (102a), relations temporelles (102b) ou relations génériques paraphrasées avec une construction ‘lié à’ (102c).

(102) a. *вельветов-ые* *барышн-и*
vel’vetov-ye *baryšn-i*
velours$_{\text{REL}}$-F.NOM.PL demoiselle-F.NOM.PL
‘demoiselles en velours’
∼*барышн-и* *оде-т-ые* *в* *вельвет*
baryšn-i *ode-t-ye* *v* *vel’vet*
demoiselle-F.NOM.PL habiller-PTCP-F.NOM.PL PREP velours$_{\text{M.ACC.SG}}$
‘demoiselles habillées en velours’

b. *турнирн-ые* *будн-и*
turnirn-ye *budn-i*
tournois$_{\text{REL}}$-N.NOM.PL semaine-N.NOM.PL
‘semaine du tournois’
∼*будн-и* *на* *турнир-е*
budn-i *na* *turnir-e*
semaine-N.NOM.PL PREP tournois-M.LOC.SG
‘semaine au tournois’

c. *подростков-ая* *проблем-а*
podrostkova-ja *problem-a*
adolescent$_{\text{REL}}$-F.NOM.SG problème-F.NOM.SG
‘problème d’adolescents’

∼*проблем-а*	*связ-анн-ая*	*с*	*подростк-ами*
problem-a	*svjaz-ann-aja*	*s*	*podrostk-ami*
problème-F.NOM.SG	lier-PTCP-F.NOM.SG	PREP	adolescent-M.INS.PL

'problème lié aux adolescents'

Des adjectifs formés à partir de noms propres expriment une propriété individuelle dans le style familier (103a), mais peuvent également développer une relation de similitude et donc un sens qualificatif (103b). En ce qui concerne les adjectifs dénominaux dérivés à partir de noms communs, ils expriment un sens de relation générique (103c).

(103) a. *дягилевск-ий* *балет*
djagilevsk-ij *balet*
Diaghilev$_{\text{REL}}$-M.NOM.SG ballet$_{\text{M.NOM.SG}}$
'ballet de Diaghilev'

∼*балет* *Дягилев-а*
balet *Djagilev-a*
ballet$_{\text{M.NOM.SG}}$ Diaghilev-M.GEN.SG
'ballet de Diaghilev'

b. *пушкинск-ая* *ясность*
puškinsk-aja *jasnost'*
Pouchkine$_{\text{REL}}$-F.NOM.SG clarté$_{\text{F.NOM.SG}}$
'clarté de Pouchkine'

∼*ясность*	*как*	*у*	*Пушкин-а*
jasnost'	*kak*	*u*	*Puškin-a*
clarté$_{\text{F.NOM.SG}}$	comme	PREP	Pouchkine-M.GEN.SG

'clarté comme chez Pouchkine'

c. *женск-ая* *походк-а*
ženska-ja *pohodk-a*
féminin-F.NOM.SG allure-F.NOM.SG
'allure féminine'

∼*походк-а* *женщин*
pohodk-a *ženščin*
allure-F.NOM.SG femme$_{\text{F.GEN.PL}}$
'allure des femmes'

Conclusion

La classe lexicale des adjectifs présente une distribution hétérogène de propriétés parmi ses membres. La distinction canonique en adjectifs de qualité et adjectifs de relation ne suffit pas à décrire complètement ces sous-classes, leurs frontières étant floues. Cependant, dans certains contextes, il est crucial de comprendre les raisons qui ont mené certains adjectifs relationnels à acquérir un sens qualificatif au lieu d'avoir

pour objectif leur classification. De plus, l'analyse propre à la sémantique formelle, par exemple, permet une lecture intersective qui englobe les deux sous classes adjectivales. Il est donc opportun de s'affranchir de la distinction canonique des adjectifs et de se concentrer sur d'autres sous-classes.

Dans cette étude, nous examinons la concurrence suffixale, nous nous intéressons donc aux adjectifs construits. Ces adjectifs peuvent être dérivés à partir de noms, verbes, autres adjectifs, etc. Les adjectifs dénominaux sont les plus nombreux d'entre eux, permettant des analyses statistiques diverses. Le russe présente une grande variété de suffixes pour construire les adjectifs à partir des noms. Cependant, certains suffixes construisent toujours des adjectifs qui ne transmettent que certains types de relations (par exemple, l'appartenance), soit des adjectifs purement qualificatifs. La dérivation des adjectifs de qualité en russe n'est pas très active. Cependant, on ne peut pas affirmer que cette classe adjectivale n'est plus enrichie. Le sens qualitatif apparaît constamment chez les adjectifs de relation.

Quant aux suffixes -*n*-, -*sk*- et -*Ov*-, ils servent à former des adjectifs dénominaux avec un sens générique relationnel, qui ne peut être concrétisé qu'en fonction du nom recteur ainsi que du contexte. Les adjectifs dénominaux expriment une signification générale de la relation et peuvent également développer une signification qualitative. De ce fait, les adjectifs formés avec les trois suffixes en question constituent un groupe assez homogène.

Chapitre 3

La concurrence en morphologie

Sommaire

Introduction

La recherche sur la concurrence en linguistique a récemment suscité un grand intérêt dans la littérature sur la formation des mots. Des éléments d'une approche théorique qui s'intéresse à la concurrence ont été élaborés, notamment dans les travaux récents de Mark Aronoff (Aronoff, 2016, 2019). La concurrence entre les affixes est une manifestation particulière de la concurrence en morphologie qui implique l'analyse des relations formelles et sémantiques entre les bases et les dérivés, ainsi que l'étude de la productivité des affixes, des moyens d'éviter la synonymie dérivationnelle et, dans le cas où les doublets sont présents dans la langue, des conditions de leur coexistence. Dans la section 3.1, nous aborderons ces sujets. La section 3.2 sera consacrée aux différentes modalités de concurrence entre les suffixes -*n*-, -*sk*- et -*Ov*-.

3.1 Le phénomène de la concurrence

Dans cette section, nous présenterons les études actuelles de la concurrence en morphologie[1], en nous appuyant sur les travaux d'Aronoff, qui, à son tour, s'est inspiré de l'évolution biologique. Les phénomènes que cet auteur traite incluent l'analyse de la synonymie, du blocage lexical, de la productivité des affixes, de la spécialisation et de l'extinction des doublets.

3.1.1 La concurrence dans les études actuelles

Selon Lindsay et Aronoff (2013), les langues sont considérées comme des systèmes complexes et continus : tout comme les systèmes biologiques, les langues s'auto-organisent et évoluent en réponse à de nombreux changements mineurs dans leur fonctionnement. En plus des propriétés purement linguistiques, les langues se caractérisent également par d'autres types de propriétés qui sont universelles pour tout système organisé.

Les observations d'Aronoff sur la concurrence en morphologie (Aronoff, 2016, 2019) sont inspirées par la biologie de l'évolution, en particulier par le principe de Gause, également connu sous le nom de principe d'exclusion réciproque, formulé dans Gause (1934). Ce principe de dynamique des populations stipule qu'il est impossible pour deux populations de subsister sur des niches écologiques similaires en équilibre parfait. Lorsque les deux espèces sont en concurrence dans les mêmes conditions externes, l'une d'entre elles sera plus efficace que l'autre et se reproduira plus rapidement. Le principe de Gause repose sur les théories mathématiques développées antérieurement par Lotka (1925) et Volterra (1926). Selon les prédictions mathématiques, une population moins efficace est destinée à disparaître. Cependant, la concurrence ne se traduit pas nécessairement par l'extinction d'une espèce : il est également possible qu'elle s'adapte en occupant une autre niche biologique.

Il est important de noter que la concurrence en biologie décrite ci-dessus est différente de celle entre les humains (Aronoff, 2019), car cette dernière est caractérisée par la connaissance mutuelle: les deux compétiteurs sont conscients des conditions de compétition. En revanche, la compétition biologique, ainsi que la compétition linguistique, sont dépourvues de cette conscience.

Dans la grande majorité des cas, le terme *concurrence* s'applique à la linguistique lorsque les locuteurs sont exposés au choix entre différentes manières d'exprimer un certain concept[2]. La concurrence peut ainsi apparaître à différents niveaux de la langue. Au niveau morphologique, elle concerne à la fois la morphologie flexionnelle et dérivationnelle. Dans ce qui suit, nous nous concentrerons sur la morphologie dérivationnelle. La notion de concurrence dans ce domaine se concrétise et représente

[1]Une analyse diachronique de la concurrence est également disponible dans les travaux de Gardani *et al.* (2019).

[2]Cf. Gardani *et al.* (2019, p.4) pour les lectures différentes du terme *concurrence*.

la relation entre deux ou plusieurs affixes associés à des modèles de formation de mots équivalents ou similaires (Huyghe et Varvara, 2023).

Les règles de formation de mots concurrents doivent utiliser les mêmes catégories lexicales en entrée et en sortie, ainsi que réaliser la même opération sémantique. À cet égard, les suffixes *-n-*, *-sk-* et *-Ov-* que nous avons analysés sémantiquement dans la section 2.3.2 sont des candidats pertinents pour une étude de la concurrence morphologique.

La concurrence absolue, dans le sens strict du terme, entraîne l'apparition de formes interchangeables et de doublets qui ne peuvent être distingués que sur le plan formel. Cependant, les affixes concurrents qui forment les lexèmes avec une sémantique strictement équivalente sont rares et leur coexistence semble hautement improbable. De plus, les affixes ne suivent pas toujours des modèles sémantiques rigides. Ils peuvent être soumis à une certaine variation par rapport à une opération sémantique prototypique, ce qui entraîne éventuellement une équivalence locale entre affixes. Quant à l'apparition de doublets, elle n'est pas non plus obligatoire. Un des doublets peut être bloqué lexicalement pour des raisons liées au fonctionnement du lexique.

La difficulté à évaluer l'identité stricte entre deux processus de dérivation conduit à une extension de la notion de concurrence affixale à ce qui pourrait être considéré comme une concurrence au sens large ou partiel. Les affixes sont alors considérés comme concurrents lorsqu'ils se combinent avec des bases de la même catégorie lexicale et dérivent des mots de la même classe lexicale, avec globalement le même sens. Dans cette définition de concurrence l'équivalence absolue n'est généralement pas requise au niveau sémantique des lexèmes en entrée ni pour la sémantique des lexèmes en sortie des règles morphologiques. En d'autres termes, les affixes concurrents ne sont pas associés à des règles de morphologie dérivationnelle identiques, mais similaires, et un certain nombre de différences possibles peuvent être observées entre les affixes concurrents (Huyghe et Varvara, 2023). Dans la présente étude, nous allons adopter la notion de concurrence au sens large.

Pour résumer, nous pouvons nous référer à Fradin (2016), qui énumère les conditions nécessaires pour qu'une concurrence en morphologie ait lieu :

- L'existence de deux exposants différents ;
- Leur combinaison potentielle avec le même lexème de base ;
- La corrélation avec le concept similaire au niveau sémantique ;
- La même distribution syntaxique.

3.1.2 Productivité, synonymie et blocage lexical

L'identité des catégories lexicales et la similarité sémantique des lexèmes construits ne sont pas les seuls critères de la concurrence. D'autres critères sont importants, tels que la productivité des affixes, la synonymie et le blocage lexical.

3.1.2.1 Productivité

La notion de productivité a plusieurs interprétations en linguistique. Švedova (1980, p.135) et Zemskaja (2011, pp.217-225) définissent la productivité des suffixes comme leur capacité de servir de modèles pour la production de nouveaux mots. Zemskaya précise que dans ce cas, il s'agit de la productivité empirique. Lindsay et Aronoff (2013) illustrent cette idée avec les suffixes -*ity* et -*ness* qui transforment les adjectifs en noms en anglais. Ces suffixes ont des propriétés sémantiques similaires, mais le choix entre eux est influencé par les usages précédents que le locuteur a faits ou a rencontrés dans sa pratique linguistique. Ce choix, à son tour, influencerait les usages futurs. Cependant, l'usage n'est pas le seul facteur qui détermine le choix d'un affixe. Selon Lindsay et Aronoff, le critère principal est l'intolérance à la synonymie dans les langues. En conséquence, les affixes productifs sont en concurrence pour la formation de nouveaux lexèmes. Si un affixe ne se différencie pas de l'autre pour éviter la synonymie, il risque de perdre sa productivité.

D'autres chercheurs définissent la productivité comme le degré d'utilisation d'un affixe dans la construction de mots nouveaux (Booij, 2012 ; Bauer, 2001 ; Plag, 2006). Arutjunova (1961, pp.54-64) postule que les suffixes les plus productifs forment les nouveaux mots à partir de mots qui sont également récents dans la langue, tels que les néologismes ou les emprunts. Ce type de productivité des affixes est considérée comme systémique par Zemskaja (2011, pp.217-225). Lindsay et Aronoff (2013) ajoutent que les affixes productifs non seulement maximisent le nombre de lexèmes de base avec lesquels ils se combinent, mais minimisent aussi les restrictions de nature phonologique, sémantique et morphologique. Toujours en référence à l'exemple des suffixes -*ity* et -*ness*, ce dernier se combine avec des bases variées presque sans restrictions phonologiques ni morphologiques[3] et est choisi par défaut pour la formation des noms. Le suffixe -*ity*, en revanche, est productif uniquement dans certaines niches morphologiques.

La capacité des suffixes à se spécialiser (phonologiquement, morphologiquement, sémantiquement, etc.) joue un rôle essentiel dans la concurrence entre certains suffixes. En étudiant la productivité des suffixes concurrents anglais, Lindsay et Aronoff (2013) mettent en évidence le fait que des suffixes moins productifs tels que -*ical* peuvent concurrencer des suffixes plus productifs tels que -*ic* en sélectionnant des bases dotés de propriétés linguistiques spécifiques. Ainsi, lorsque des contraintes phonologiques s'appliquent aux deux suffixes et favorisent le plus court d'entre eux, le suffixe moins favorable peut maintenir sa productivité en se spécialisant différemment, au niveau phonologique, morphologique, sémantique ou pragmatique.

Pour résumer la discussion sur la productivité il conviendrait de faire le point sur sa nature. La productivité peut être considérée comme une notion discrète, où un affixe donné peut potentiellement être utilisé dans la formation de nouveaux mots. Dans ce cas, la productivité a une nature binaire, que Corbin (2012) appelle *disponibilité*.

[3]La seule restriction de nature morphologique concerne la non-combinaison avec les adjectifs construits de type X*ible* (Aronoff et Anshen, 2017).

Cependant, nous avons observé que les termes 'degré', 'maximiser' et 'minimiser' sont présents dans les définitions de la productivité. Cela implique que la productivité est plutôt scalaire, soit *rentabilité* dans la terminologie de Corbin (2012). Ainsi, Bauer (2001, pp.125-163) souligne que la productivité en tant que notion discrète n'est plus acceptable, étant donné qu'il y a un grand nombre de degrés de productivité entre les patrons très productifs et ceux qui ne sont pas productifs du tout : peu productifs, peu actifs, semi-productifs, etc. La productivité discrète doit alors céder la place à la productivité continue. La productivité peut ainsi être représentée comme un éventail de probabilités : d'un côté se trouvent des affixes à des degrés de productivité non restreints ; les affixes non productifs sont placés à l'opposé ; entre ces deux extrémités se retrouvent les affixes à productivité différentes (Aronoff et Anshen, 2017, pp.241-243 ; Aronoff et Lindsay, 2014, pp.70-73).

3.1.2.2 Synonymie

Nous avons mentionné dans la discussion sur la productivité que, selon Aronoff, les langues ne tolèrent pas la synonymie absolue (Aronoff, 2016, 2019). La synonymie se réfère à la correspondance totale ou partielle des significations principales les mots qui ne diffèrent que par des nuances de sens ou de style. Pour déterminer les synonymes, Apresjan (1974, p.223) se base sur le critère distributif. Dans le cas de synonymie totale, il s'agit des deux lexèmes concurrents qui ont exactement le même sens et la même distribution, et peuvent donc être utilisés de façon interchangeable. Illustrée ainsi, la synonymie représente un cas spécifique de la concurrence gausienne, et l'absence de synonymie est le résultat de l'application du principe de Gause. Cependant, comme Aronoff le souligne à plusieurs reprises, une telle synonymie n'existe pas dans les langues : si les deux mots ont exactement le même sens et la même distribution, l'un d'entre eux finirait par disparaître ou se différencier sémantiquement. Même les termes quasi synonymes, tels que HAZELNUT 'noisette' et FILBERT 'aveline', ont des distributions distinctes (104a-104b) : HAZELNUT est beaucoup plus probable dans certains contextes que FILBERT.

(104) a. *hazelnut spread* 'pâte à tartiner aux noisettes'
**filbert spread* 'pâte à tartiner aux avelines'

b. *hazelnut praline* 'praliné noisette'
**filbert praline* 'praliné aveline'

Un autre exemple de synonymie presque absolue donné par Aronoff est le triplet de mots HURRICANE 'ouragan', TYPHOON 'typhon' et CYCLONE 'cyclone ' : ils désignent le même type de tempête, mais dans des régions géographiques différentes, et apparaissent donc dans des contextes distincts.

Dans le sens strict du terme, les synonymes doivent avoir une sémantique identique. Cependant, comme discuté dans la section précédente, l'identité absolue n'est généralement pas requise dans les études de la concurrence. En outre, dans la

réalité linguistique, les synonymes absolus sont rares ; la plupart des synonymes sont des quasi-synonymes, avec des nuances de sens différentes. Tout comme la productivité, la synonymie peut être considérée comme un phénomène gradable. En fonction du nombre de contextes qu'ils partagent et des sens en commun qu'ils expriment, les deux lexèmes peuvent être plus ou moins synonymes : certains d'entre eux ont des sens et des distributions identiques, tandis que d'autres manifestent des variations (Baayen *et al.*, 2013). Présentée de cette manière, la synonymie correspondrait également à un éventail de probabilités : les deux extrémités contiennent des synonymes absolus d'un côté, et des mots ayant des sens et des distributions sans chevauchement de l'autre.

3.1.2.3 Blocage lexical

L'un des moyens pour prévenir l'apparition de synonymes est de bloquer une forme émergeante. Il s'agit alors d'un blocage lexical, défini comme le non-occurrence d'une forme qui est due à l'existence d'une autre forme. Cette définition a été formulée par Aronoff (1976), qui se base sur une représentation du lexique organisée par cases de paradigmes, où chaque case ne peut contenir qu'une seule forme. Ce phénomène est également traité par Paul (1897) dans la perspective psycholinguistique : l'effet de blocage peut être expliqué par le fait que le locuteur privilégie de récupérer dans son stock lexical un lexème déjà existant plutôt que de former un nouveau (Gardani *et al.*, 2019, 17-21). Plus tard, Aronoff (2016) révise la notion de blocage en la divisant en deux : le blocage d'un seul mot et le blocage d'un patron. Le blocage d'un mot individuel est alors vu comme le résultat de la concurrence gausienne entre deux mots potentiellement synonymes pour un seul sens. Selon le principe de Gause, un seul mot remporterait cette compétition.

En ce qui concerne la morphologie dérivationnelle, le terme de blocage peut être défini de manière plus précise comme le fait qu'un lexème a peu de chances d'être construit s'il existe un lexème qui lui est sémantiquement similaire. Le blocage conçu de cette manière est alors discret : une forme concurrente existe tandis qu'une autre n'existe pas. Généralement, c'est une forme qui existe déjà qui remporte cette compétition ; le contenu sémantique du dérivé sera alors associé à un seul suffixe. Cependant, cette 'victoire' peut être temporaire (Aronoff, 2016) : la nouvelle construction peut exister, mais à une très basse fréquence et ne s'utiliser que dans des contextes très particuliers (GLORY 'gloire' est un mot communément utilisé ; GLORIOSITY 'gloire' se retrouve dans les noms de quelques salons de coiffure). Roché (1997) exemplifie le même phénomène en français (105), en indiquant les fréquences des doublons sur le Web entre parenthèses.

(105) CAMION → CAMIONNEUR (470 000) / CAMIONNIER (10)

De plus, étant donné que le blocage est un phénomène psycholinguistique, il peut dépendre des conditions mentales des locuteurs : si un mot construit existant est temporairement oublié, un mot construit avec un suffixe distinct serait alors utilisé,

tandis que dans des conditions optimales ce même mot aurait été bloqué (Aronoff et Anshen, 2017). Enfin, si un mot est bloqué par un autre, il peut toutefois apparaître, mais avec un sens distinct comparé à celui qu'il devait acquérir suite à l'opération morphologique de base (Aronoff et Lindsay, 2014). Dans ce cas de différenciation sémantique, les deux lexèmes ne seraient plus en concurrence. Dans la terminologie écologique, chaque mot a ainsi trouvé sa propre niche.

Compte tenu des particularités des modes de concurrence, la nature binaire n'est pas en mesure de refléter les subtilités des interactions entre les lexèmes rivaux, telles que productivité et synonymie discutées plus haut (Aronoff et Lindsay, 2014). La nature du blocage semble également être scalaire. Il peut se manifester avec une intensité différente en fonction de la fréquence des lexèmes : pour les lexèmes de haute fréquence, le blocage se produit avec une régularité élevée (ce qui le rapproche du phénomène binaire) ; les lexèmes de basse fréquence coexistent souvent avec leurs rivaux (Gardani *et al.*, 2019, pp.17-21). De ce fait, le blocage lexical a lieu plus souvent en morphologie dérivationnelle, comparé à la morphologie flexionnelle. L'éventail de probabilités pour le blocage peut ainsi se situer entre les deux cas extrêmes – blocage total et coexistence des rivaux – avec plusieurs scénarios intermédiaires.

Dans la section suivante nous allons présenter les moyens dont disposent les langues pour éviter l'apparition des synonymes dérivationnels et les moyens de différenciation dans les cas où de tels synonymes apparaissent toutefois.

3.1.3 Solutions pour la concurrence affixale

Du point de vue de l'évolution biologique, la concurrence entre deux espèces peut être résolue de deux manières : la disparition ou l'adaptation d'une des espèces, selon le principe de Gause. En morphologie, les mêmes mécanismes ont lieu au niveau des mots individuels, discutés dans la section précédente, et au niveau des affixes. Ces mécanismes agissent en accord avec le principe d'économie linguistique (Bréal, 1904), selon lequel un des affixes équivalents est voué à se spécialiser ou à disparaître.

Le cas le plus simple de résolution de la concurrence consiste en la disparition d'un des suffixes. C'est le cas, par exemple, du suffixe *-ment* en anglais, qui a cédé la place au suffixe *-ation*. Les deux ont été empruntés du français, mais le suffixe *-ation* a gagné le terrain de la formation des noms abstraits déverbaux à partir du XVIIe siècle (Lindsay et Aronoff, 2013). Un autre exemple est le suffixe agentif français *-on* qui était un concurrent très actif de *-eur* en ancien et moyen français. Cependant, il est progressivement devenu peu productif et a abandonné sa fonction agentive au profit de *-eur* (Huyghe et Varvara, 2023). Dans la plupart des cas, la concurrence entre les deux suffixes résulte en une perte progressive de la productivité d'un des suffixes. Ce dernier restera toujours dans la langue, mais ne sera pas utilisé pour la formation de nouveaux mots, ou bien son utilisation serait très limitée, par exemple, au discours terminologique (Aronoff et Lindsay, 2014).

Le suffixe le moins compétitif en termes de concurrence n'est pas obligatoirement voué à disparaître ou à perdre sa productivité. Il peut plutôt trouver une niche

linguistique et se différencier d'un autre suffixe plus compétitif sur des critères tels que la phonologie, la morphologie, la sémantique, la pragmatique, etc. (Aronoff, 2016).

Les phénomènes phonologiques peuvent influencer l'utilisation d'un suffixe particulier. Par exemple, la position de l'accent ou la longueur du nom de base en syllabes peuvent limiter l'utilisation d'un suffixe donné. Ces phénomènes ont été étudiés en français, notamment par Lignon, 2013 et Bonami et Thuilier, 2019 pour les suffixes -*iser* et -*ifier*, et en anglais, par Lindsay et Aronoff (2013), pour les suffixes -*ize* et -*ify*.

Il est possible qu'un des suffixes concurrents s'attache à des lexèmes de base formés à l'aide d'un suffixe spécifique, comme le montrent Lindsay et Aronoff (2013) pour les suffixes adjectivales anglais -*ic* et -*ical*. Ainsi, les affixes concurrents trouvent une niche morphologique dans laquelle ils développent des propriétés distinctives : bien que -*ical* soit moins productif que -*ic* en anglais contemporain, il est largement préféré à -*ic* pour les radicaux se terminant par -*olog*-, ce qui peut être considéré comme un modèle de sa subsistance dans le système dérivationnel. Les propriétés morphologiques ont également été étudiées par Missud et Villoing (2020) pour -*age*, -*ion* et -*ment* en français, Varvara (2020) pour -*mento* et -*zione* en italien et Bonami et Thuilier (2019) pour l'impact de la famille morphologique sur le choix entre les suffixes -*iser* et -*ifier* en français.

Du point de vue pragmatique, le choix peut varier en fonction du registre de langue. Lindsay et Aronoff (2013) donnent l'exemple de -*esque* ou -*ian* en anglais qui sont utilisés dans le discours formel, alors que le suffixe -*ish* est plus courant dans le discours familier.

Certaines études examinent également les différences externes. Les facteurs stylistiques et les fréquences d'utilisation sont analysés notamment par Baayen (1994) sur l'exemple des préfixes *in*- et *un*- et des suffixes -*ity* et -*ness* en anglais ou par Naccarato (2019) sur l'exemple du russe. Romaine (1983) étudie l'âge les locuteurs en fonction de l'utilisation des suffixes -*ité* et -*ness*. Dressler *et al.* (2019) font le point sur les diminutifs en allemand et italien. Les préférences sociolinguistiques sont traitées par Säily (2011) sur l'exemple des suffixes -*ity* et -*ness*. L'évolution diachronique des affixes et les changements de leur productivité ont également été étudiés par Lindsay et Aronoff, 2013 ; Arndt-Lappe, 2014 et Aronoff, 2016 (Huyghe et Varvara, 2023).

Du point de vue sémantique, il est possible qu'un des deux lexèmes en concurrence développe un sens distinct pour s'adapter, auquel cas les deux lexèmes ne sont plus considérés comme concurrents (Krysin, 1965, p.12-13). Dans les cas exceptionnels où les lexèmes sont des synonymes absolus, ils peuvent progressivement devenir des quasi-synonymes, avec une perte de sens partagés ou une spécialisation (Azarx, 1987, p.65). Selon Aronoff et Cho (2001), les suffixes -*hood* et -*ship* en anglais ont originellement servi à dériver des adjectifs avec le même sens 'état ou condition', mais -*ship* est actuellement utilisé uniquement pour des termes génériques, tandis que -*hood* peut être utilisé pour des termes à la fois génériques et individuels. Des différences sémantiques possibles entre les dérivés sont également discutées. Les suffixes anglais -*ity* et -*ness*

ainsi que -*age* et -*ery* sont notamment traités par Baeskow (2012) et Schulte (2015), repectivement. Fradin (2016) s'intéresse à la sémantique des suffixes français -*age*, -*ion*, -*ment*. Lehrer (2000), à son tour, examine les propriétés connotatives des suffixes -*ster*, -*eer* et -*eur* en anglais (Huyghe et Varvara, 2023).

Nous avons examiné les différentes dimensions de la concurrence, en mettant en évidence que les propriétés phonologiques, morphologiques et sémantiques des lexèmes de base peuvent être déterminantes pour résoudre la concurrence entre les deux affixes. Ces contraintes peuvent parfois s'imposer et primer sur les contraintes liées aux règles morphologiques. Par exemple, selon Roché (1997), le terme AVIONNEUR désigne 'une personne qui construit des avions'. Dans ce cas, le suffixe -*eur* est peu attendu car il est habituellement utilisé pour construire des noms d'agents à partir de bases verbales. Cependant, Roché explique que le suffixe -*eur* est tout de même utilisé au lieu de -*ier* ici pour des raisons d'euphonie : ainsi, il permet d'éviter des séquences sonores difficilement acceptables. Les cas où un suffixe inattendu est utilisé pour éviter des séquences phonologiques problématiques ont été appelés *échangisme suffixal* par Lignon et Plénat, 2009 et par Plénat, 2011.

De manière générale, le principe de Gause et la loi de différenciation sémantique impliquent la solution inévitable de la concurrence. Cependant, ce processus est diachronique, il prend généralement du temps, la durée ne peut pas être estimée à l'avance. Ainsi, il peut y avoir des étapes transitoires, où la recherche d'une forme appropriée est en cours, et donc les deux concurrents coexistent en état d'équilibre (Aronoff, 2016). En termes de linguistique, il s'agit des doublets.

A la fin de la section 3.1.1, nous avons énuméré les conditions nécessaires pour la concurrence affixale, à savoir : l'existence de deux exposants différents, leur combinaison potentielle avec le même lexème de base, la corrélation avec le même concept au niveau sémantique, la même distribution syntaxique. Ces conditions évoluent lorsque des doublets émergent : les bases ne sont pas toujours identiques et le contenu sémantique n'est pas déterminé à l'avance et peut être lié à au moins deux exposants.

Pour illustrer ces points, Fradin (2014, 2016) donne l'exemple des verbes RESSORTIR$_1$ et RESSORTIR$_2$, qui n'ont pas le même paradigme (106). Il s'agit donc de deux verbes distincts morphologiquement, malgré l'identité formelle au niveau des formes de citations.

(106) a. RESSORTIR$_1$: *il ressort*$_{\text{PRS.3.SG}}$; *il ressort-ait*$_{\text{PST.3.SG}}$
b. RESSORTIR$_2$: *il ressort-it*$_{\text{PRS.3.SG}}$; *il ressortiss-ait*$_{\text{PST.3.SG}}$

Dans le cas des verbes PERLER$_1$ et PERLER$_2$, il ne s'agit pas non plus du même lexème. Bien que tous les deux aient le même paradigme, leur sens n'est pas le même (107) : l'un est transitif, l'autre est intransitif. De nouveau il s'agit ici de lexèmes différents ; PERLAGE et PERLEMENT se serait alors pas considérés comme doublets.

(107) a. PERLER$_1$: *Claudine perlait un sac*
→ PERLAGE

b. PERLER$_2$: *Une larme perla sur son cil*
→ PERLEMENT

Un autre exemple concerne les lexèmes qui ont le même paradigme et le même sens, mais une distribution syntaxique différente (108). Comme dans l'exemple précédent, ENTERREMENT et ENTERRAGE ne sont pas en concurrence.

(108) a. ENTERRER$_1$: *un humain*
→ ENTERREMENT

b. ENTERRER$_2$: *un objet*
→ ENTERRAGE

Pour résumer, seuls les lexèmes qui ont la même forme, le même paradigme et le même contenu sémantique peuvent être considérés comme des doublets morphologiques. Si l'une des contraintes de cette liste est violée, il ne s'agit pas de doublets.

Selon Fradin (2016), l'une des propriétés des doublets consiste en ce que leur sens peut être strictement identique ou complètement divergent, avec plusieurs cas se situant entre ces deux extrêmes, ce que Azarx (1987, p.64) appelle les doublets complets et incomplets. L'identité est illustrée par Fradin en (109). Les lexèmes PAVAGE et PAVEMENT sont en effet des doublets en concurrence.

(109) a. PAVER$_1$
→ PAVAGE$_1$:
Le pavage de la cour devait s'achever avant Noël.
→ PAVEMENT$_1$:
Le pavement de la plateforme du tramway progresse.

b. PAVER$_2$
→ PAVAGE$_2$:
Le pavage de la cour est concentrique.
→ PAVEMENT$_2$:
Les visiteurs découvrent le pavement de la cathédrale de Sienne.

Dans le cas inverse de divergence sémantique, Fradin propose des exemples en (110). Les deux lexèmes – RASAGE et RASEMENT – ne sont pas en concurrence puisqu'ils sont dérivés des deux verbes distincts sémantiquement.

(110) a. RASER$_1$
→ RASAGE$_1$:
Le rasage des aisselles.

b. RASER$_2$
→ RASEMENT$_1$:
Le rasement de la ville et du château.

L'étude des doublets impliquerait ainsi une étude des bases et de leurs sémantiques, ainsi que l'étude des contextes dans lesquels les dérivés apparaissent, pour déduire s'il s'agit de vrais doublets ou non. Le décompte des doublets ne devrait donc pas se faire sur la forme mais sur le sens. Ainsi, le problème de la polysémie s'ajoute à l'étude des doublets. De plus, Fradin remarque qu'il existe des cas de confusion, où les vrais et faux doublets sont employés de manière interchangeable.

En plus du contenu sémantique des bases, les doublets peuvent être différenciés selon la norme, qui consiste en un ensemble de règles fondées sur les usages considérés comme des modèles à suivre. Un doublon peut être en accord avec la norme linguistique ou bien être une création personnelle (111a). Lorsqu'on est face à une forme inattendue, on peut l'analyser comme un cas de distorsion entre un suffixe attendu et le suffixe attesté. Dans ces distorsions, certaines contraintes (phonologiques, morphologiques, sémantiques) peuvent être violées (Roché, 1997). Selon le contexte, un doublet peut faire partie du lexique commun ou être spécifique à un domaine (111b). Enfin, selon leur emploi géographique, l'utilisation des doublets peut varier selon les régions ou les pays des locuteurs (111c).

(111) a. TUTOIEMENT / ?TUTOYAGE
b. PLAFONNEMENT / PLAFONNAGE (terminol.)
c. ENCAVEMENT / ENCAVAGE (Suisse)

3.2 La concurrence suffixale en russe

Dans la section 3.1, nous avons présenté le phénomène de la concurrence suffixale dans son ensemble. Nous nous focaliserons ici sur la concurrence des adjectifs dénominaux russes formés à l'aide des suffixes -*n*-, -*sk*- et -*Ov*-, en particulier sur trois aspects de cette concurrence : la productivité des affixes, les contraintes de différents ordres linguistiques qui favorisent le choix d'un suffixe donné et l'existence de doublets.

3.2.1 Productivité des suffixes

En ce qui concerne la productivité des suffixes construisant différents types adjectivaux, Zemskaja (2015, p.124) souligne que l'inventaire des adjectifs qualificatifs s'élargit plus lentement et de manière plus rare que celui des adjectifs de relation. Cet auteur considère la dérivation des adjectifs de relation comme étant en deuxième position en termes de productivité et de construction de nouveaux mots, juste après la dérivation des substantifs. Les adjectifs de relation sont particulièrement fréquents dans la terminologie, comme nous l'avons constaté dans la section 2.3.3. De plus, les suffixes -*n*-, -*sk*- et -*Ov*- se combinent avec un inventaire de plus en plus vaste de bases nominales et les restrictions phonologiques, morphologiques ou sémantiques sont de moins en moins nombreuses (Zemskaja, 1965, pp. 21-22).

Selon Zverkovskaja (1986), les suffixes -*n*- et -*Ov*- ont conservé une productivité élevée tout au long de l'histoire de la langue russe. Dans la langue russe contemporaine,

ce sont les suffixes -*n*-, -*sk*- et -*Ov*- qui construisent la majorité des adjectifs et sont considérés comme étant productifs.

Des travaux en slavistique tentent de hiérarchiser la productivité des trois suffixes considérés. Ainsi, Nemčenko (1976, pp.11-12) analyse le nombre d'adjectifs néologiques construits, ce qui permet de conclure que le suffixe le plus productif est -*sk*-. Cependant, une étude menée par Alekseeva (2011, 65-94), effectuée trente cinq ans plus tard, aboutit à une conclusion différente, selon laquelle le suffixe -*Ov*- aurait gagné en productivité et aurait dépassé le suffixe -*sk*-.

3.2.2 Dimensions de la concurrence

Dans la section 3.1.3, nous avons abordé l'existence de différents types de contraintes qui influencent le choix d'un affixe lors de la dérivation lexicale, telles que les contraintes formelles, stylistiques, lexicales, dérivationnelles, etc. Dans cette section, nous présenterons l'état actuel des connaissances sur les contraintes associées aux suffixes -*n*-, -*sk*- et -*Ov*-[4].

3.2.2.1 Structure syllabique et position de l'accent

Il n'y a pas d'indications particulières concernant la position de l'accent du nom de base ni sur la structure syllabique des noms. Les adjectifs avec les trois suffixes peuvent être motivés par les noms où l'accent est sur le thème (112a-112f) ; ou sur la flexion (112g-112k).

(112) a. ЯГОДА /ˈjagoda/ /jágoda/ 'fruit rouge' → ЯГОДНЫЙ
JAGODA JAGODNYJ

b. КАПКАН /kapˈkan/ /kapkán/ 'piège' → КАПКАННЫЙ
KAPKAN KAPKANNYJ

c. МАЧТА /ˈmatʃ͡ʲta/ /máčta/ 'mât' → МАЧТОВЫЙ
MAČTA MAČTOVYJ

d. ГОРНОСТАЙ /gornoˈstaj/ /gornostáj/ 'hermine' → ГОРНОСТАЕВЫЙ
GORNOSTAJ GORNOSTAEVYJ

e. ЗАВОД /zaˈvod/ /zavód/ 'usine' → ЗАВОДСКОЙ
ZAVOD ZAVODSKOJ

f. МЕДИЦИНА /mʲedʲiˈt͡sina/ /mˈedˈicína/ 'médecine' →
MEDICINA
МЕДИЦИНСКИЙ
MEDICINSKIJ

g. СКОВОРОДА /skovoroˈda/ /skovorodá/ 'poêle' → СКОВОРОДНЫЙ
SKOVORODA SKOVORODNYJ

[4]Tous les exemples proviennent de Švedova (1980), à l'exception de ceux où une référence explicite est mentionnée.

h. БАХЧА /bax'tʃʲa/ /baxčá/ 'melon' → БАХЧЕВОЙ
BAXČA BAXČEVOJ

i. БУЗИНА /buzi'na/ /buziná/ 'sureau' → БУЗИНОВЫЙ
BUZINA BUZINOVYJ

j. СЕЛО /sʲe'lo/ /s'eló/ 'village' → СЕЛЬСКИЙ
SELO SEL'SKIJ

k. СЛОБОДА /slobo'da/ /slobodá/ 'sloboda' → СЛОБОДСКОЙ
SLOBODA SLOBODSKOJ

Les noms de base des adjectifs en question peuvent avoir une structure polysyllabique, comme exemplifié en (112a-112k) ou monosyllabique : *-n-* (113a), *-Ov-* (113b), *-sk-* (113c). Il est généralement admis que le suffixe *-Ov-* se combine davantage avec les bases monosyllabiques.

(113) a. ЗУБ /zub/ /zub/ 'dent' → ЗУБНОЙ
ZUB ZUBNOJ

b. ТКАНЬ /tkanʲ/ /tkan'/ 'tissu' → ТКАНЕВЫЙ
TKAN' TKANEVYJ

c. Дон /don/ /don/ 'Don (fleuve)' → ДОНСКОЙ
Don DONSKOJ

Ainsi, la structure syllabique et la position de l'accent dans le nom de base ne sont pas considérées comme des propriétés discriminantes pour le choix du suffixe (à l'exception du suffixe *-Ov-*). Des informations plus spécifiques sont disponibles dans la littérature concernant le dernier phonème du thème.

3.2.2.2 Dernier phonème des radicaux

Le suffixe *-sk-* se combine avec les thèmes terminant par les consonnes labiales et dentales (114a-114d), ainsi que par /n/, /r/, /lʲ/, /j/ (114e-114h). Ce suffixe peut se retrouver après les consonnes vélaires et post-alvéolaires uniquement pour les toponymes (114i-114k). Le suffixe *-sk-* se combine plus rarement avec les thèmes finissant par /rʲ/, /nʲ/, /l/ (114l-114n) – pour les noms de base qui désignent les mois et les toponymes d'origine étrangère. Il peut choisir les thèmes finissant par les voyelles si le nom de base désigne un toponyme étranger (114o).

(114) a. СЕРБ 'serbe' → СЕРБСКИЙ
SERB SERBSKIJ

b. ШЕФ 'chef' → ШЕФСКИЙ
ŠEF ŠEFSKIJ

c. СОСЕД 'voisin' → СОСЕДСКИЙ
SOSED SOSEDSKIJ

d. эскимос 'esquimau' → эскимосский
ÈSKIMOS ÈSKIMOSSKIJ

e. океан 'océan' → океанский
OKEAN OKEANSKIJ

f. организатор 'organisateur' → организаторский
ORGANIZATOR ORGANIZATORSKIJ

g. февраль 'février' → февральский
FEVRAL' FEVRAL'SKIJ

h. май 'mai' → майский
MAJ MAJSKIJ

i. Лейпциг 'Leipzig' → лейпцигский
LEJPCIG LEJPCIGSKIJ

j. Углич 'Ouglitch (ville)' → угличский
UGLIČ UGLIČSKIJ

k. Париж 'Paris' → парижский
PARIŽ PARIŽSKIJ

l. сентябрь 'septembre' → сентябрьский
SENTJABR' SENTJABR'SKIJ

m. июнь 'juin' → июньский
IJUN' IJUN'SKIJ

n. Ямал 'péninsule de Yamal' → ямалский
JAMAL JAMALSKIJ

o. Бордо 'Bordeaux' → бордоский
BORDO BORDOSKIJ

Les emprunts qui se terminent par une consonne de toute nature forment aussi facilement les adjectifs avec le suffixe -*n*- (Vinogradov, 1952, p.173). Par contre, les noms de base qui se terminent par les séquences 'consonne (sauf *l*) + k' ne font pas d'adjectifs avec le suffixe -*n*-, d'autres suffixes sont privilégiés (115a). Les noms se terminant par la séquence -*čk(a)* font exception à cette règle. Dans ce cas, une alternance voyelle/∅ est observée (115b) (Zemskaja, 2011, p.89).

(115) a. парк 'parc' → парковый
PARK PARKOVYJ
воск 'cire' → восковой
VOSK VOSKOVOJ

b. спичка 'allumette' → спичечный
SPIČKA SPIČEČNYJ

3.2.2.3 Structure morphologique

Les trois suffixes en question ont des préférences distinctes au niveau du suffixe dérivationnel du nom de base (dans les cas où ce dernier est construit). Ainsi, la dérivation adjectivale en *-n-* est très productive à partir des noms avec les suffixes *-k(a)*, *-ic(a)*, *-ok*, *-stv(o)*, *-ost'*, *-ot(a)*, *-in(a)* (116a-116g).

(116) a. КЛУБНИКА 'fraise' → КЛУБНИЧНЫЙ
KLUBNIKA KLUBNIČNYJ

b. БОЛЬНИЦА 'hôpital' → БОЛЬНИЧНЫЙ
BOL'NICA BOL'NIČNYJ

c. ОБМОРОК 'évanouissement' → ОБМОРОЧНЫЙ
OBMOROK OBMOROČNYJ

d. ГОСТЕПРИИМСТВО 'hospitalité' → ГОСТЕПРИИМНЫЙ
GOSTEPRIIMSTVO GOSTEPRIIMNYJ

e. РЕДКОСТЬ 'rareté' → РЕДКОСТНЫЙ
REDKOST' REDKOSTNYJ

f. ВЫСОТА 'hauteur' → ВЫСОТНЫЙ
VYSOTA VYSOTNYJ

g. ГЛУБИНА 'profondeur' → ГЛУБИННЫЙ
GLUBINA GLUBINNYJ

Parmi les noms de base l'on retrouve les noms composés (117a) et les noms composés suffixés en *-k(a)* (117b), ainsi que des composés avec des éléments néoclassiques (117c-117l): *-bus*, *-gramm(a)*, *-graf*, *-plan*, *-skop*, *-drom*, *-tek(a)*, *-tip(ija)*, *-fon(ija)*, *-gam(ija)*.

(117) a. ЧЕРНОЗЁМ 'terre végétale (terre noire)' → ЧЕРНОЗЁМНЫЙ
ČERNOZËM ČERNOZËMNYJ

b. КОФЕВАРКА 'cafetière ([machine à] infuser le café)' →
KOFEVARKA
КОФЕВАРОЧНЫЙ
KOFEVAROČNYJ

c. АВТОБУС 'bus' → АВТОБУСНЫЙ
AVTOBUS AVTOBUSNYJ

d. ДИАГРАММА 'diagramme' → ДИАГРАММНЫЙ
DIAGRAMMA DIAGRAMMNYJ

e. ТЕЛЕГРАФ 'télégraphe' → ТЕЛЕГРАФНЫЙ
TELEGRAF TELEGRAFNYJ

f. АЭРОПЛАН 'aéroplane' → АЭРОПЛАННЫЙ
AÈROPLAN AÈROPLANNYJ

g. ПЕРИСКОП ‘périscope’ → ПЕРИСКОПНЫЙ
PERISKOP PERISKOPNYJ

h. ВЕЛОДРОМ ‘vélodrome’ → ВЕЛОДРОМНЫЙ
VELODROM VELODROMNYJ

i. ДИСКОТЕКА ‘discothèque’ → ДИСКОТЕЧНЫЙ
DISKOTEKA DISKOTEČNYJ

j. СТЕРЕОТИП ‘stéréotype’ → СТЕРЕОТИПНЫЙ
STEREOTIP STEREOTIPNYJ

k. ТЕЛЕФОН ‘téléphone’ → ТЕЛЕФОННЫЙ
TELEFON TELEFONNYJ

l. ПОЛИГАМИЯ ‘polygamie’ → ПОЛИГАМНЫЙ
POLIGAMIJA POLIGAMNYJ

Les noms suffixés en *-ni(e)*, *-ti(e)*, *-nik*, *-ščin(a)* servent de bases pour les adjectifs en *-n-* plus rarement (118a-118d).

(118) a. ПОСЕЛЕНИЕ ‘colonie’ → ПОСЕЛЕННЫЙ
POSELENIE POSELENNYJ

b. СОБЫТИЕ ‘évènement’ → СОБЫТИЙНЫЙ
SOBYTIE SOBYTIJNYJ

c. РУДНИК ‘mine’ → РУДНИЧНЫЙ
RUDNIK RUDNIČNYJ

d. БАРЩИНА ‘servage’ → БАРЩИННЫЙ
BARŠČINA BARŠČINNYJ

En ce qui concerne le suffixe *-Ov-*, les noms de base peuvent être construits avec les suffixes suivants : *-ok*, *-ik*[5], *-nik*, *-čik*, *-ak*, *-njak*, *-k(a)*, *-očk-*, *-ušk-*, *-ec*, *-in(a)*, *-in*, *-ol*, *-it*, *-b(a)*, *-yš*, *-en’*, *-l’*, *-’ë* (119a-119s) (Vinogradov, 1952, p.177).

(119) a. ГРИБОК ‘mycose’ → ГРИБКОВЫЙ
GRIBOK GRIBKOVYJ

b. ШАРИК ‘petit ballon’ → ШАРИКОВЫЙ
ŠARIK ŠARIKOVYJ

c. ПЛАВНИК ‘aileron’ → ПЛАВНИКОВЫЙ
PLAVNIK PLAVNIKOVYJ

d. КОЛОКОЛЬЧИК ‘grelot’ → КОЛОКОЛЬЧИКОВЫЙ
KOLOKOL’ČIK KOLOKOL’ČIKOVYJ

e. СОЛОНЧАК ‘désert de sel’ → СОЛОНЧАКОВЫЙ
SOLONČAK SOLONČAKOVYJ

[5]Dans leurs étude des ordres des affixes, Sims et Parker (2015, p.171) présentent la combinaison de *-ik* et *-Ov-* comme une des plus fréquentes.

f. ИЗВЕСТНЯК ‘calcaire’ → ИЗВЕСТНЯКОВЫЙ
IZVESTNJAK IZVESTNJAKOVYJ

g. ГРЯДКА ‘potager’ → ГРЯДКОВЫЙ
GRJADKA GRJADKOVYJ

h. КОСТОЧКА ‘noyau’ → КОСТОЧКОВЫЙ
KOSTOČKA KOSTOČKOVYJ

i. РАКУШКА ‘coquillage’ → РАКУШКОВЫЙ
RAKUŠKA RAKUŠKOVYJ

j. РЕЗЕЦ ‘incisive’ → РЕЗЦОВЫЙ
REZEC REZCOVYJ

k. ПАРУСИНА ‘drap de toile’ → ПАРУСИНОВЫЙ
PARUSINA PARUSINOVYJ

l. ВАТИН ‘rembourrage’ → ВАТИНОВЫЙ
VATIN VATINOVYJ

m. СТИРОЛ ‘styrène’ → СТИРОЛОВЫЙ
STIROL STIROLOVYJ

n. КАМЫШИТ ‘roseaux’ → КАМЫШИТОВЫЙ
KAMYŠIT KAMYŠITOVYJ

o. РЕЗЬБА ‘gravure’ → РЕЗЬБОВОЙ
REZ’BA REZ’BOVOJ

p. ЗАРОДЫШ ‘foetus’ → ЗАРОДЫШЕВЫЙ
ZARODYŠ ZARODYŠEVYJ

q. ОПОЛЗЕНЬ ‘glissement de terrain’ → ОПОЛЗНЕВЫЙ
OPOLZEN’ OPOLZNEVYJ

r. ПОРОСЛЬ ‘arbuste’ → ПОРОСЛЕВЫЙ
POROSL’ POROSLEVYJ

s. СЫРЬЁ ‘matière première’ → СЫРЬЕВОЙ
SYR’Ë SYR’EVOJ

Finalement, le suffixe -*sk*- privilégie des noms de base, généralement animés, avec les suffixes -/nij/- ou -/tij/- (120a-120b) ainsi que les noms composés avec la deuxième partie -*gorod* / -*grad* ‘-ville’ (120c). Les noms animés désignant les personnes qui se terminent par -*ist*[6] ou -*tel’* forment en général des adjectifs avec le suffixe -*sk*- (120d-120e) (Lopatin, 1977, p.30 ; Vinogradov, 1952, p.177).

(120) a. ПРАВЛЕНИЕ ‘conseil’ → ПРАВЛЕНСКИЙ
PRAVLENIE PRAVLENSKIJ

[6]D’après Sims et Parker (2015, p.171) la combinaison de -*ist* et -*sk*- est également une des plus fréquentes entre les deux suffixes.

b. житьё 'vie' → житейский
ŽIT'Ë ŽITEJSKIJ

c. Ленинград 'Léninegrad' → ленинградский
LENINGRAD LENINGRADSKIJ

d. журналист 'journaliste' → журналистский
ŽURNALIST ŽURNALISTSKIJ

e. просветитель 'éclaireur' → просветительский
PROSVETITEL' PROSVETITEL'SKIJ

3.2.2.4 Genres et classes flexionnelles

Les informations morphologiques supplémentaires telles que les classes flexionnelles ou les genres sont rarement abordées. Sorokina (1984, p.69) examine le suffixe -*Ov*-, qui peut se combiner avec des noms de genre féminin, masculin et neutre, et conclut qu'il est le plus productif avec des noms masculins. Les noms indéclinables se caractérisent généralement par l'absence de dérivés (Krysin, 2008, p.482). Toutefois, cela dépend grandement du degré d'utilisation du mot. Par exemple, le mot манто (manto) 'manteau' n'a pas de dérivés enregistrés, alors que пальто (pal'to) 'manteau' a des dérivés tels que пальтовый (pal'tovyj) (*пальтовая ткань* (*pal'tovaja tkan'*) 'tissue de manteau'). Le mot кофе (kofe) 'café' a un dérivé кофейный (kofejnyj), tandis que какао (kakao) 'chocolat chaud' n'en a pas (à l'exception de l'adjectif какаовый (kakaovyj) utilisé dans le langage professionnel).

Les allomorphies thématiques semblent plus discriminatoires.

3.2.2.5 Voyelle mobile

En ce qui concerne l'alternance voyelle/∅, de manière générale, la voyelle est gardée dans le radical de l'adjectif formé avec le suffixe -*n*- (121a-121b). Cependant, même si dans l'espace thématique du nom de base l'alternance voyelle/∅ n'a pas lieu, elle peut émerger dans les thèmes dérivationnels (121c-121d). Une autre alternance /i/ ∼ /e/, non typique pour le paradigme nominal, peut être observée dans les noms qui se terminent par -/ij/- (121e).

(121) a. семья 'famille' → семейный
SEM'JA SEMEJNYJ
семь-я 'famille-F.NOM.SG' ∼ *сем-ей* 'famille-F.GEN.PL'
sem'-ja *sem-ej*

b. упадок 'déclin' → упадочный
UPADOK UPADOČNYJ
упадок 'déclin$_{\text{M.NOM.SG}}$' ∼ *упадк-ов* 'déclin-M.GEN.PL'
upadok *upadk-ov*

c. ИГЛА ‘aiguille’ → ИГ**О**ЛЬНЫЙ
IGLA IGOL’NYJ
игл-а ‘aiguille-F.NOM.SG’ ∼ *игл* ‘aiguille$_{\text{F.GEN.PL}}$’
igl-a *igl*

d. КОРАБЛЬ ‘navire’ → КОРАБ**Е**ЛЬНЫЙ
KORABL’ KORABEL’NYJ
корабль ‘navire$_{\text{M.NOM.SG}}$’ ∼ *корабл-ей* ‘navire-M.GEN.PL’
korabl’ *korabl-ej*

e. ЛИН**И**Я ‘ligne’ → ЛИН**Е**ЙНЫЙ
LINIJA LINEJNYJ
лини-я ‘ligne-F.NOM.SG’ ∼ *линий* ‘ligne$_{\text{F.GEN.PL}}$’
lini-ja *linij*

Contrairement à *-n-*, si dans l’espace thématique du nom de base il y a une alternance voyelle/∅, le thème sans voyelle est préféré par le suffixe *-Ov-* (122).

(122) СОСНА ‘pin’ → СОСНОВЫЙ
SOSNA SOSNOVYJ
сосн-а ‘pin-F.NOM.SG’ ∼ *сосен* ‘pin$_{\text{F.GEN.PL}}$’
sosn-a *sosen*

En ce qui concerne *-sk-*, la voyelle est généralement préservée dans le radical de l’adjectif (123a). Cependant, comme pour les suffixes *-n-* et *-Ov-* l’ajout des séquences qui ne sont pas présentes dans l’espace thématique du nom de base peut aussi avoir lieu dans le radical de l’adjectif dérivé : ∅-/e/ ou ∅-/o/ (123b-123c).

(123) a. ДЕРЕВНЯ ‘village’ → ДЕРЕВ**Е**НСКИЙ
DEREVNJA DEREV**E**NSKIJ
деревн-я ‘village-F.NOM.SG’ /dʲeˈrʲevnʲa/ /d'er'évn'a/ ∼
derevn-ja
деревень ‘village$_{\text{F.GEN.PL}}$’ /dʲerʲeˈvʲenʲ/ /d'er'ev'én'/
dereven’

b. МИНИСТР ‘ministre’ → МИНИСТ**Е**РСКИЙ
MINISTR MINISTERSKIJ
министр ‘ministre$_{\text{M.NOM.SG}}$’ ∼ *министр-ов* ‘ministre-M.GEN.PL’
ministr *ministr-ov*

c. УГР ‘ougrien (peuple)’ → УГ**О**РСКИЙ
UGR UGORSKIJ
угр ‘ougrien$_{\text{N.M.NOM.SG}}$’ ∼ *угр-ов* ‘ougrien$_{\text{N}}$-M.GEN.PL’
ugr *ugr-ov*

3.2.2.6 Mouillure

La mouillure est présente dans les radicaux des adjectifs en *-n-* (124a). Cependant, si la base se termine par *-sk-*, l'alternance généralement n'a pas lieu (124b) (Zemskaja, 2011, p.89). Le suffixe *-Ov-* choisit également dans l'espace thématique un radical avec une consonne finale molle (124c-124d). Le suffixe *-sk-*, à son tour, n'a pas de préférence : les radicaux utilisés pour la dérivation peuvent avoir une consonne finale molle (124e) ou dure (124f-124g).

(124)
a. БАЛ /bal/ /bal/ 'bal' → БАЛЬНЫЙ /ˈbalʲnɨj/ /bál'noj/
BAL BAL'NYJ

b. ПРОПУСК 'passe' → ПРОПУСКНОЙ
PROPUSK PROPUSKNOJ
ГРОТЕСК 'grotesque' → ГРОТЕСКНЫЙ
GROTESK GROTESKNYJ

c. СМОЛА /smoˈla/ /smolá/ 'résine' →
SMOLA
СМОЛЕВОЙ /smolʲeˈvoj/ /smol'evój/
SMOLEVOJ

d. ЕЛЬ /jelʲ/ /jel'/ 'sapin' → ЕЛОВЫЙ /jeˈlovɨj/ /jelóvoj/
EL' ELOVYJ

e. Урал /uˈral/ /urál/ 'Oural (montagnes)' →
Ural
УРАЛЬСКИЙ /uˈralʲskʲij/ /urál'skoj/
URAL'SKIJ

f. КОНЬ /konʲ/ /kon'/ 'cheval' → КОНСКИЙ /ˈkonskʲij/ /kónskoj/
KON' KONSKIJ

g. МОРЕ /ˈmorʲe/ /mór'e/ 'mer' → МОРСКОЙ /morsˈkoj/ /morskój/
MORE MORSKOJ

3.2.2.7 Palatalisation

La présence des consonnes palatalisées du type /k/ ∼ /č/, /g/ ∼ /ž/, /x/ ∼ /š/, /c/ ∼ /č/ dans l'espace thématique des noms de base a une influence sur le choix du suffixe. Le suffixe *-n-*, en l'occurrence, choisit le thème avec une consonne palatalisée (125a-125d). Zemskaja (2011, p.89) postule que la consonne non palatalisée ne peut pas apparaître devant le suffixe *-n-*. Une mutation du type /c/ ∼ /t/ est présente dans le radical des adjectifs en *-n-* (125e). Des mutations non régulières du type /st/ ∼ /šč/, /b/ ∼ /bl/ peuvent aussi avoir lieu (125f-125g) pour le suffixe *-Ov-*. La consonne palatalisée dans les couples /g/ ∼ /ž/ et /x/ ∼ /š/ est privilégiée par *-sk-* (125h-125i). Les cas où la dernière consonne du radical n'est pas palatalisée sont plus rares et

concernent les formations plus récentes (125j) (Zemskaja, 2011, p.90). Les mutations du type /č/ ∼ /c/ sont plus rares (125k).

(125) a. БРАК /brak/ /brak/ ‘mariage’ → БРАЧНЫЙ /ˈbrat͡ʃʲnɨj/ /bráčnoj/
BRAK BRAČNYJ

b. КНИГА /ˈknʲiga/ /kn'íga/ ‘livre’ → КНИЖНЫЙ /ˈknʲiʒnɨj/ /kn'ížnoj/
KNIGA KNIŽNYJ

c. ГРЕЧИХА /grʲeˈt͡ʃʲixa/ /gr'ečíxa/ ‘sarrasin’ →
GREČIXA
ГРЕЧИШНЫЙ /grʲeˈt͡ʃʲiʃnɨj/ /gr'ečišnoj/
GREČIŠNYJ

d. ПУГОВИЦА /ˈpugovʲit͡sa/ /púgov'ica/ ‘bouton (vêtement)’ →
PUGOVICA
ПУГОВИЧНЫЙ /ˈpugovʲit͡ʃʲnɨj/ /púgov'ičnoj/
PUGOVIČNYJ

e. ИНЕРЦИЯ /iˈnert͡sija/ /inércija/ ‘inertie’ →
INERCIJA
ИНЕРТНЫЙ /iˈnertnɨj/ /inértnoj/
INERTNYJ

f. ХОЛСТ /xolst/ /xolst/ ‘toile’ → ХОЛЩЁВЫЙ /xolˈʃʲːovɨj/ /xolščóvoj/
XOLST XOLŠČOVYJ

g. ЗЯБЬ /zʲabʲ/ /z'ab'/ ‘champs labouré’ →
ZJAB’
ЗЯБЛЕВЫЙ /ˈzʲablʲevɨj/ /z'ábl'evoj/
ZJABLEVYJ

h. ВОЛГА /ˈvolga/ /vólga/ ‘Volga (fleuve)’ →
VOLGA
ВОЛЖСКИЙ /ˈvolʒskʲij/ /vólžskoj/
VOLŽSKIJ

i. ЧЕХИЯ /ˈt͡ʃʲexʲija/ /čéx'ija/ ‘Tchéquie’ →
ČEXIJA
ЧЕШСКИЙ /ˈt͡ʃʲeʃskʲij/ /čéšskoj/
ČEŠSKIJ

j. ЛЕЙПЦИГ ‘Leipzig’ → ЛЕЙПЦИГСКИЙ
LEJPCIG LEJPCIGSKIJ

k. ТКАЧ /tkat͡ʃ/ /tkač/ ‘tisseur’ → ТКАЦКИЙ /ˈtkat͡sʲkʲij/ /tkáckoj/
TKAČ TKACKIJ

3.2.2.8 Allomorphies segmentales

La formation des adjectifs avec le suffixe -*n*- peut se faire sur les radicaux tronqués : /ij/ (126a-126c) ou /enij/ (126d), /k/ (126e). Les voyelles finales /o/ et /e/ des noms étrangers indéclinables sont aussi supprimées (126f-126g).

(126) a. мгновение 'instant' → мгновенный
MGNOVENIE MGNOVENNYJ

b. сценарий 'scénario' → сценарный
SCENARIJ SCENARNYJ

c. лаборатория 'laboratoire' → лабораторный
LABORATORIJA LABORATORNYJ

d. воскресение 'dimanche' → воскресный
VOSKRESENIE VOSKRESNYJ

e. цепочка 'chaîne' → цепочный
СЕРОČКА СЕРОČNYJ

f. факсимиле 'fax' → факсимильный
FAKSIMILE FAKSIMIL'NYJ

g. барокко 'baroque' → барочный
BAROKKO BAROČNYJ

Dans l'exemple (114o) vu précédemment, la voyelle finale est préservée dans le radical de l'adjectif dérivé ; cependant, le suffixe -*sk*- s'attache à des thèmes tronqués des noms indéclinables (127a). Les autres troncations sont possibles : la séquence -/ij/- pour les noms qui finissent par -*enie*, -*tie* (127b-127c) ou -*ija* (127d) ; la séquence -/j/- des bases désignant les toponymes (127e) ; -*ec* pour les noms désignant les êtres animés (127f) ; -*k*-/-*ok*- des toponymes mais aussi des noms communs (127g-127h) ; -*ščin(a)*/-*čin(a)*, -*ek* et -*š(a)* - dans certains cas (127i-127l).

(127) a. Чикаго 'Chicago' → чикагский
Čikago čikagskij

b. отделение 'département' → отделенский
otdelenie otdelenskij

c. общежитие 'foyer' → общежитский
obščežitie obščežitskij

d. типография 'typographie' → типографский
tipografija tipografskij

e. Болонья 'Bologne' → болонский
Bolon'ja bolonskij

f. беженец 'réfugié' → беженский
beženec beženskij

g. Ямайка ‘Jamaïque’ → ямайский
Jamajka jamajskij

h. уравниловка ‘nivellement’ → уравниловский
uravnilovka uravnilovskij

i. женщина ‘femme’ → женский
ženščina ženskij

j. мужчина ‘homme’ → мужской
mužčina mužskoj

k. Терек ‘Terek (fleuve)’ → терский
Terek terskij

l. Польша ‘Pologne’ → польский
Pol’ša pol’skij

Les noms indéclinables finissant par -*u* ou -*i* sont aussi tronqués (128a-128b) devant -*Ov*-. La troncation de -*nik* peut avoir lieu mais n’est pas systématique (128c).

(128) a. кенгур**у** ‘kangourou’ → кенгуровый
kenguru kengurovyj

b. джерс**и** ‘Jersey’ → джерсевый
džersi džersevyj

c. можжевель**ник** ‘genévrier’ → можжевеловый
možževel’nik možževelovyj

L’ajout des séquences -/j/- et -/ij/- n’est pas régulier (129a-129b) devant -*Ov*-. En ce qui concerne -*n*-, l’ajout du matériel phonologique supplémentaire a lieu surtout pour les noms d’origine étrangère qui sont indéclinables : /t/, /z/, /ɟ/ (129c-129e). L’ajout d’une séquence /*es*/ est observé pour les noms d’origine slave (129f).

(129) a. голубь ‘pigeon’ → голубьевый
golub’ golub’evyj

b. герань ‘géranium’ → гераниевый
geran’ geranievyj

c. кабаре ‘cabaret’ → кабаретный
kabare kabaretnyj

d. буржуа ‘bourgeois’ → буржуазный
buržua buržuaznyj

e. регби ‘rugby’ → регбийный
regbi regbijnyj

f. слово ‘mot’ → словесный
slovo slovesnyj

Un autre phénomène qui peut avoir lieu, c'est le remplacement des suffixes. Ainsi, *-ok* peut se transformer en *-ec* devant *-sk-* (130a). Les noms qui désignent les toponymes d'origine slave et qui se terminent pas *-sk* subissent une fusion entre le *-sk* final du thème et le suffixe dérivationnel *-sk-*. Cette fusion peut être totale (130b) ou partielle (130c).

(130) a. ИГРОК 'joueur' → ИГРЕЦКИЙ
IGROK IGRECKIJ

b. Брянск 'Briansk (ville)' → БРЯНСКИЙ
Brjansk BRJANSKIJ

c. Дамаск 'Damas' → ДАМАССКИЙ
Damask DAMASSKIJ

Une autre propriété qui détermine des allomorphies consiste en ce que le suffixe *-n-* peut s'attacher à un thème qui est distinct de celui du nominatif singulier (131a-131b). La même chose est valable pour *-Ov-* (131c-131d).

(131) a. ВРЕМЯ 'temps' → ВРЕМЕННЫЙ
VREMJA VREMENNYJ
врем-я 'temps-N.NOM.SG' ∼ *времён* 'temps$_{\text{N.GEN.PL}}$'
vrem-ja *vremën*

b. НЕБО 'ciel' → НЕБЕСНЫЙ
NEBO NEBESNYJ
неб-о 'ciel-N.NOM.SG' ∼ *небес* 'ciel$_{\text{N.GEN.PL}}$'
neb-o *nebes*

c. ЗВЕНО 'maillon' → ЗВЕНЬЕВОЙ
ZVENO ZVEN'EVOJ
звен-о 'maillon-N.NOM.SG' ∼ *звень-ев* 'maillon-N.GEN.PL'
zven-o *zven'-ev*

d. СУДНО 'bateau' → СУДОВОЙ
SUDNO SUDOVOJ
судн-о 'bateau-N.NOM.SG' ∼ *суд-ов* 'bateau-N.GEN.PL'
sudn-o *sud-ov*

3.2.2.9 Sémantique

La sémantique des adjectifs dérivés est étroitement liée à celle des noms de base, comme nous l'avons établi dans les sections 2.2.3 et 2.3.2. Dans cette dernière, nous avons également introduit les propriétés sémantiques des noms de base, préférées par les suffixes *-n-*, *-sk-* et *-Ov-*.

Le suffixe *-n-* est généralement combiné avec des noms communs, principalement inanimés, qui désignent à la fois des objets concrets (132a) et des phénomènes abstraits

(132b). Les noms désignant des êtres animés (132c) et les noms désignant des animaux (132d) sont rares.

(132) a. хлеб 'pain' → хлебный
XLEB XLEBNYJ

b. радость 'joie' → радостный
RADOST' RADOSTNYJ

c. инженер 'ingénieur' → инженерный
INŽENER INŽENERNYJ

d. рыба 'poisson' → рыбный
RYBA RYBNYJ

Les noms désignant les être animés sont limités aux noms suffixés en *-ar*, *-ar'*, *-or*, *-er*, *-ir*, *-ent*, *-ik* (133a-133g).

(133) a. пекарь 'boulanger' → пекарный
PEKAR' PEKARNYJ

b. гусляр 'harpiste' → гуслярный
GUSLJAR GUSLJARNYJ

c. гравёр 'graveur' → гравёрный
GRAVËR GRAVËRNYJ

d. акционер 'actionnaire' → акционерный
AKCIONER AKCIONERNYJ

e. фуражир 'butineur' → фуражирный
FURAŽIR FURAŽIRNYJ

f. абонент 'abonné' → абонентный
ABONENT ABONENTNYJ

g. фанатик 'fanatique' → фанатичный
FANATIK FANATIČNYJ

Le suffixe *-Ov-* se combine avec des bases nominales désignant également des objets non animés concrets, mais il a une préférence pour les noms désignant des plantes et des minéraux (Zemskaja, 1991, p.153).

Quant à *-sk-*, il est majoritairement combiné avec des noms communs désignant des toponymes (134a)[7], des instituts et des entreprises (134b-134c) (Vinogradov, 1952, p.176).

(134) a. Иран 'Iran' → иранский
IRAN IRANSKIJ

b. комитет 'comité' → комитетский
KOMITET KOMITETSKIJ

[7]Le suffixe *-sk-* est aussi le seul à construire des adjectifs ethniques en russe.

c. ЗАВОД ‘usine’ → ЗАВОДСКОЙ
ZAVOD ZAVODSKOJ

3.2.2.10 Étymologie

En ce qui concerne l'étymologie, nous retrouvons les mêmes spécificités sémantiques que celles mentionnées précédemment pour les suffixes -*n*- et -*sk*- qui se combinent avec les noms d'origine étrangère. Les noms non animés concrets désignant des objets ou des substances forment les adjectifs avec le suffixe -*n*- (135a-135c). Les emprunts correspondant à des noms animés concrets qui désignent des personnes forment les adjectifs avec le suffixe -*sk*- (135d-135e) (Krysin, 2008, p.484).

(135) a. АБАЖУР ‘abat-jour’ → АБАЖУРНЫЙ
ABAŽUR ABAŽURNYJ

b. КАПКАН ‘piège’ → КАПКАННЫЙ
KAPKAN KAPKANNYJ

c. МАЙОНЕЗ ‘mayonnaise’ → МАЙОНЕЗНЫЙ
MAJONEZ MAJONEZNYJ

d. ПРЕЗИДЕНТ ‘président’ → ПРЕЗИДЕНТСКИЙ
PREZIDENT PREZIDENTSKIJ

e. ДАМА ‘dame’ → ДАМСКИЙ
DAMA DAMSKIJ

Les adjectifs construits avec le suffixe -*Ov*- peuvent être dérivés de noms d'origine étrangère abstraits (136a-136c).

(136) a. ИНТЕРФЕЙС ‘interface’ → ИНТЕРФЕЙСОВЫЙ
INTERFEJS INTERFEJSOVYJ

b. ЛАУНДЖ ‘lounge’ → ЛАУНДЖЕВЫЙ
LAUNDŽ LAUNDŽEVYJ

c. КЛИНИНГ ‘clinning’ → КЛИНИНГОВЫЙ
KLINING KLININGOVYJ

3.2.3 Doublets suffixaux

Avant de présenter le sujet des doublets, il convient de faire une parenthèse sur la formation des adjectifs à partir d'homonymes[8]. Les homonymes dérivent de nouveaux homonymes en quantité très faible (137a). En règle générale, pour prévenir l'homonymie entre les dérivés, des suffixes distincts sont employés (137b-137c). D'après l'analyse de Fradin (2014, 2016) exposée dans la section 3.1.3, ces deux derniers exemples ne sont pas considérés comme doublets.

[8]Les questions d'homonymie et de polysémie seront traitées plus en détails dans la section 5.1.2.

(137) a. КЛЮЧ$_1$ 'clé' → КЛЮЧЕВОЙ$_1$
KLJUČ KLJUČEVOJ
ключевая позиция 'position clé'
ključevaja pozicija
КЛЮЧ$_2$ 'source' → КЛЮЧЕВОЙ$_2$
KLJUČ KLJUČEVOJ
ключевая вода 'eau de source'
ključevaja voda

b. ЗАВОД$_1$ 'usine' → ЗАВОДСКОЙ
ZAVOD ZAVODSKOJ
ЗАВОД$_2$ 'mise en marche' → ЗАВОДНОЙ
ZAVOD ZAVODNOJ

c. ЛУК$_1$ 'arc' → ЛУЧНЫЙ
LUK LUČNYJ
ЛУК$_2$ 'oignon' → ЛУКОВЫЙ
LUK LUKOVYJ

Concernant les doublets réels, c'est-à-dire les adjectifs formés à partir de la même base nominale, dans la plupart des cas, en particulier lorsqu'il s'agit de néologismes, ces adjectifs peuvent être considérés comme quasi-synonymes (Hénault et Sakhno, 2015).

Comme mentionné précédemment dans la section 3.2.2, le suffixe -*sk*- semble se différencier davantage des suffixes -*n*- et -*Ov*- en relation à la sémantique des noms de base avec lesquels il se combine. Vinogradov et Švedova (1964, p.299) notent que le suffixe -*sk*- est souvent en concurrence avec sa variante étendue -*Ovsk*- (138). Les constructions avec le suffixe -*sk*- sont plus anciennes, tandis que le suffixe -*Ovsk*- a gagné en productivité dans la formation d'adjectifs dénominaux entre la fin du XVIIIe siècle et le début du XIXe siècle.

(138) a. ФАНАТ 'supporteur' →
FANAT
ФАНАТСКИЙ / ФАНАТОВСКИЙ
FANATSKIJ FANATOVSKIJ

b. ТИНЕЙДЖЕР 'adolescent' →
TINEJDŽER
ТИНЕЙДЖЕРСКИЙ / ТИНЕЙДЖЕРОВСКИЙ
TINEJDŽERSKIJ TINEJDŽEROVSKIJ

Selon Hénault et Sakhno (2015) et Sakhno (2022), les adjectifs formés avec le suffixe -*sk*- sont souvent liés à la désignation d'une certaine société. Cela signifie que le sens mobilisé du nom de base est attribué au nom recteur non pas comme intrinsèquement inhérent, mais en termes de localisation sociale, institutionnelle, idéologique, politique, géographique, ethnique, etc. Par contre, les adjectifs formés

avec le suffixe -*n*- mobilisent souvent le sens le plus profond et/ou le plus général du nom de base et marquent souvent une relation étroite de type conceptuel entre la sémantique du nom de base et la sémantique du nom recteur. Des doublets suffixaux en -*sk*- et -*n*- sont exemplifiés en (139). Tandis que ГОРНЫЙ (GORNYJ) se rapporte aux montagnes en tant que phénomène naturel et géophysique, ГОРСКИЙ (GORSKIJ) fait référence aux montagnes en tant que lieu de résidence.

(139) ГОРА 'montagne' →
GORA
ГОРНЫЙ : *горный воздух* 'air de montagne' /
GORNYJ *gornyj vozdux*
ГОРСКИЙ : *горские обычаи* 'coutumes montagnardes'
GORSKIJ *gorskie obyčai*

Les adjectifs suffixés en -*Ov*-, quant à eux, mobilisent plus souvent un sens plus particulier et secondaire (Hénault et Sakhno, 2015) du nom de base et marquent souvent une relation de type situationnel entre la sémantique du nom de base et celle du nom recteur. Les doublets avec les suffixes -*n*- et -*Ov*- sont présentés en (140). СНЕЖНЫЙ (SNEŽNYJ) désigne un hiver pendant lequel il tombe une grande quantité de neige, dans la conscience linguistique russe, un hiver normal est conceptuellement inséparable de la neige, tandis que СНЕГОВОЙ (SNEGOVOJ) est un terme scientifique ou technique.

(140) СНЕГ 'neige' →
SNEG
СНЕЖНЫЙ : *снежная зима* 'hiver enneigé' /
SNEŽNYJ *snežnaja zima*
СНЕГОВОЙ : *снеговая линия* 'ligne de neige'
SNEGOVOJ *snegovaja linija*

C'est la compétition entre les suffixes -*n*- et -*Ov*-, en raison de leurs préférences sémantiques similaires par rapport à la base nominale, qui engendre le plus grand nombre de doublets adjectivaux. Ce phénomène, du fait de son ampleur, est l'objet d'une attention particulière dans les études linguistiques (Alekseeva, 2011, p.73).

Zemskaja (1965, pp.142-148) observe que l'existence des doublets en -*n*- et -*Ov*- a une longue histoire. Ce phénomène est attesté tout le long de l'histoire de la langue russe pour les noms concrets qui désignent des plantes (141a), des substances (141b), des matériaux (141c). En général, les adjectifs formés avec le suffixe -*n*- sont plus anciens que ceux dérivés avec le suffixe -*Ov*-.

(141) a. АПЕЛЬСИН 'orange' → АПЕЛЬСИНОВЫЙ / АПЕЛЬСИННЫЙ
APEL'SIN APEL'SINOVYJ APEL'SINNYJ
b. КУПОРОС 'vitriol' → КУПОРОСОВЫЙ / КУПОРОСНЫЙ
KUPOROS KUPOROSOVYJ KUPOROSNYJ

c. МРАМОР ‘marbre’ → МРАМОРОВЫЙ / МРАМОРНЫЙ
MRAMOR MRAMOROVYJ MRAMORNYJ

Au XIXe siècle, les adjectifs formés avec le suffixe -*n*- ont développé une connotation qualificative (Zemskaja, 2000, pp.125-126). Cette évolution n'est pas due au hasard, car les adjectifs en -*n*- sont plus enclins à former des adjectifs de qualité en raison de leurs propriétés morphologiques et syntaxiques. Ils disposent de formes courtes, de formes comparatives synthétiques, de la possibilité de former des adverbes en -*o*, et de la capacité de créer des noms abstraits en -*ost'* (cf. la section 1.3.2). Les adjectifs formés avec le suffixe -*Ov*-, en revanche, ne possèdent pas ces propriétés.

Lorsqu'un adjectif acquiert un sens qualificatif, les liens sémantiques avec son nom de base s'affaiblissent et l'adjectif ne désigne plus une relation entre le nom recteur et le nom de base, mais désigne une qualité propre à ce dernier (142a-142b). La désemantisation des adjectifs peut aussi avoir lieu, dans ce cas, les locuteurs ne relient plus l'adjectif à son nom de base étymologique (142c).

(142) a. ОПЫТ ‘expérience’ → ОПЫТНЫЙ
OPYT OPYTNYJ

b. ГРОЗА ‘orage’ → ГРОЗНЫЙ
GROZA GROZNYJ

c. ЗЛАК ‘céréale’ → ЗЛАЧНЫЙ [9]
ZLAK ZLAČNYJ

L'affaiblissement ou disparition des liens sémantiques entre le nom de base et l'adjectif en -*n*- renforce la dérivation des adjectifs de relation avec le suffixe -*Ov*- (143).

(143) a. СРОК ‘délai’ → СРОЧНЫЙ ‘urgent’ / СРОКОВЫЙ
SROK SROČNYJ SROKOVYJ

b. ГРОЗА ‘orage’ → ГРОЗНЫЙ ‘terrible’ / ГРОЗОВОЙ
GROZA GROZNYJ GROZOVOJ

c. ГРУЗ ‘poids’ → ГРУЗНЫЙ ‘lourd’ / ГРУЗОВОЙ
GRUZ GRUZNYJ GRUZOVOJ

De nos jours, ce processus de dérivation des adjectifs de relation avec le suffixe -*Ov*- a pris de l'ampleur en raison du développement scientifique et technologique et de l'enrichissement du vocabulaire technique. Le suffixe -*Ov*- est très actif dans le discours professionnel en raison du besoin de nouvelles nominalisations (voir la discussion dans la section 2.3.3). Les adjectifs en -*Ov*- transmettent un sens relationnel générique, tandis que les adjectifs en -*n*- sont plutôt qualificatifs.

[9] *Злачное место* (*zlačnoe mesto*) peut être traduit en français par ‘endroit fertile‘. Dans le contexte de la prière orthodoxe pour les défunts, cela désigne l'endroit de repos paisible pour les justes, c'est-à-dire le paradis. Cependant, avec le temps, cette expression est devenue ironique et a pris un sens opposé, se référant à un endroit de débauche, d'ivresse et de divertissements douteux et interdits.

L'adjectivation avec le suffixe -*Ov*- permet de résoudre un problème de l'existence d'un adjectif qualificatif formé sur la même base en offrant un adjectif purement relationnel. Ainsi, l'adjectif ПЫЛЬНЫЙ (PYL'NYJ) signifie 'poussiéreux, couvert de poussière' dans les phrases de type (144a) et ne peut pas transmettre le même sens dans les constructions casuelles et/ou prépositionnelles (144b). Pour ce dernier, l'adjectif ПЫЛЕВОЙ (PYLEVOJ) est utilisé (144c).

(144) a. *пыльн-ые* *частиц-ы* 'particules poussiéreuses'
pyl'n-ye *častic-y*
poussiéreux-F.NOM.PL particule-F.NOM.PL
пыльн-ые *очк-и* 'lunettes poussiéreuses'
pyl'n-ye *očk-i*
poussiéreux-N.NOM.PL lunettes-N.NOM.PL

b. *частиц-ы* *пыл-и*
častic-y *pyl-i*
particule-F.NOM.PL poussière-F.GEN.SG
'particules de poussière'
очк-и *от* *пыл-и*
očk-i *ot* *pyl-i*
lunettes-N.NOM.PL PREP poussière-F.GEN.SG
'lunettes contre la poussière'

c. *пылев-ые* *частиц-ы* 'particules de poussière'
pylev-ye *častic-y*
poussière$_{\text{REL}}$-F.NOM.PL particule-F.NOM.PL
пылев-ые *очк-и* 'lunettes contre la poussière'
pylev-ye *očk-i*
poussière$_{\text{REL}}$-N.NOM.PL lunettes-N.NOM.PL

Alekseeva (2011, p.77,99) suggère également que l'émergence des doublets en -*n*- et -*Ov*- est relativement récente, mais justifie ce fait par la tendance à former des adjectifs à partir de noms de base empruntés. Cette tendance témoigne d'une certaine instabilité dans le processus de formation des adjectifs. Avec l'apparition d'emprunts massifs, de nombreux adjectifs dérivés émergent initialement dans la langue parlée, souvent de manière informelle, avant de s'introduire dans la langue publique et les textes littéraires, créant ainsi des doublets. En outre, Zemskaja (1965, pp.142-148) note que l'inventaire des noms de base avec lesquels le suffixe -*Ov*- peut se combiner s'est élargie, passant des noms concrets privilégiés historiquement aux noms abstraits et déverbaux (145).

(145) a. ПУСК 'lancement' → ПУСКОВОЙ
PUSK PUSKOVOJ

b. РЫВОК 'élan' → РЫВКОВОЙ
RYVOK RYVKOVOJ

Il est à noter que tous les adjectifs formés à l'aide du suffixe -*n*- ne développent pas de sens qualificatif. Certaines formes sont toujours purement relationnelles. Cependant, les adjectifs formés à l'aide du suffixe -*Ov*- apparaissent simultanément à leurs côtés et les deux peuvent exister en tant que doublets dotés du même sens. Zemskaja (1965, pp.142-148) souligne que l'utilisation de ces formes varie en fonction du discours : les adjectifs en -*n*- sont plus fréquemment utilisés dans un style neutre, alors que leurs doublets en -*Ov*- prédominent dans la terminologie professionnelle.

Ainsi, la langue se retrouve en période de stabilisation et de coexistence de deux formes. La possibilité de formation de doublets n'est cependant pas la même pour les trois suffixes en question.

Conclusion

Le phénomène de la concurrence suffixale concerne les cas où plusieurs suffixes peuvent être combinés avec la même base pour produire deux lexèmes distincts formellement mais similaires sémantiquement. Cependant, les langues ne tolèrent pas la synonymie absolue, et même si les deux formes concurrentes peuvent coexister temporairement, la situation de concurrence a tendance à être résolue. La concurrence a généralement lieu lorsque les suffixes sont productifs ; à long terme, un suffixe moins concurrentiel peut perdre sa productivité. Un des suffixes ne doit pas nécessairement être éliminé, il peut également se spécialiser par rapport à l'autre dans un discours spécifique. De plus, même si plusieurs suffixes sont disponibles pour la dérivation avec le même sens lexical-référentiel, l'une des formes peut être bloquée en raison de contraintes phonologiques, morphologiques ou sémantiques.

Les suffixes -*n*-, -*sk*- et -*Ov*- sont considérés comme les plus productifs dans la dérivation des adjectifs. Ils semblent diverger en grande partie grâce à leurs préférences sémantiques pour les noms de base. Le suffixe -*n*- est utilisé pour former des adjectifs à partir des noms dénombrables, en majorité inanimés. Le suffixe -*sk*- sert principalement à former des adjectifs à partir des noms de personnes, des noms géographiques et ethniques, des noms d'organisations et d'institutions. Avec le suffixe -*Ov*-, des adjectifs peuvent être formés à partir des noms d'animaux, ainsi que de certains noms inanimés. Il a été constaté que, dans le russe moderne, le suffixe -*Ov*- se combine avec un inventaire de bases nominales de plus en plus large. Certains cas présentent un intérêt supplémentaire puisqu'il existe des adjectifs qui sont formées sur une même base avec les suffixes -*n*-, -*sk*- et -*Ov*-, dont les derniers sont généralement plus récent et ont une signification purement relative, tandis que l'adjectif avec -*n*- est souvent qualificatif.

Partie II

Données et annotations

Chapitre 4

Données des adjectifs

Sommaire

Introduction

Les études quantitatives de la concurrence entre les suffixes qui forment des adjectifs dénominaux en russe requièrent des ressources lexicales considérables qui fournissent à la fois des informations sur les fréquences ainsi que des annotations détaillées des noms de base pour chaque adjectif. Ce chapitre portera sur la création d'une base de données adjectivale RuDénom qui répertorie les adjectifs dénominaux russes avec leurs fréquences dans RusCorpora. La question des noms de base sera abordée dans le chapitre suivant.

Dans la section 4.1, nous exposerons les diverses sources pouvant être utilisées pour établir un corpus d'adjectifs dénominaux, les méthodes qui peuvent être utilisées pour sa constitution ainsi que les outils numériques permettant de l'analyser. Nous

présenterons aussi la constitution de la base de données RuDénom. La section 4.2 portera sur trois sous-ensembles de données : données de haute et de basse fréquence et les données de doublets.

4.1 Constitution d'une base de données adjectivales

Comme discuté dans la partie I, nous nous focaliserons uniquement sur les adjectifs dénominaux formés à l'aide des suffixes *-n-*, *-sk-*, *-Ov-*. Ces suffixes sont considérés comme les plus productifs en synchronie et construisent la majorité des adjectifs dérivés à partir des noms en russe.

4.1.1 Ressources et méthodes disponibles

4.1.1.1 Etudes qualitatives et quantitatives

Les études menées en morphologie par les slavistes ont généralement un objectif précis ; les bases de données sont construites progressivement pour couvrir certains phénomènes mais pas pour fournir une base de données exhaustive sur les adjectifs. Par exemple, Alekseeva (2011) se concentre uniquement sur les néologismes adjectivaux, Antipina (2012b) étudie les couples de doublets et les paronymes, Pèrènlèj (2006) analyse les adjectifs dénominaux à sémantique spatiale dans une perspective sociologique. En général, les données proviennent de dictionnaires (Antipina, 2012a ; Pèrènlèj, 2006) ou de la presse (Alekseeva, 2011). Par conséquent, ces données sont peu nombreuses (374 néologismes, 606 couples de doublets, 573 adjectifs dénominaux respectivement). D'autres études, comme Ljaševskaja et Šarov (2009), sont basées sur des corpus et fournissent, par exemple, une liste des adjectifs les plus fréquents dans RusCorpora[1]. Cependant, comme cette dernière approche est basée sur la fréquence des adjectifs dans le corpus, le nombre de cas recensés est également limité (1 002 adjectifs), les adjectifs peu fréquents et les hapax étant exclus.

Les mêmes problématiques s'observent pour le français par Bonami et Tribout (2021) : les ressources disponibles ont toutes été construites pour documenter un ou plusieurs processus de formation de mots spécifiques, mais en conséquence l'échantillon du lexique français qui est rassemblé a des caractéristiques variables qui dépendent des objectifs de la recherche. *Lexeur* (Wauquier *et al.*, 2020) ou *Dénom* (Strnadová, 2014), par exemple, contiennent de nombreux éléments non documentés dans les dictionnaires suite à une collecte de données à partir des corpus ; en revanche, *Converts* (Tribout, 2010) se concentre sur les éléments documentés dans un dictionnaire, en l'absence d'une bonne méthode pour extraire les converts à partir d'un corpus. Il est donc difficile de combiner les ressources pour faire des études statistiques significatives.

La nécessité d'un corpus volumineux et le plus exhaustif possible est justifiée par la croissance des études quantitatives appliquées à la linguistique. Non seulement une

[1]Le dictionnaire de fréquences est accessible à l'adresse suivante : `http://dict.ruslang.ru/freq.php`.

telle ressource peut contribuer à des études statistiques ou basées sur l'apprentissage automatique, mais elle peut également faciliter l'analyse des processus productifs en russe à travers l'étude des hapax, des néologismes et des doublets.

Dans la présente recherche, nous avons opté pour une étude quantitative, complétée par une étude qualitative. Une variété de méthodes quantitatives et leur application à l'étude de la concurrence en russe sera présentée dans la section 4.1.2.

4.1.1.2 Ressources disponibles

Traditionnellement, les linguistes ont étudié les mots en se basant sur des dictionnaires, corpus tirés des œuvres littéraires ou de la presse. Auparavant, ces ressources étaient construites de manière entièrement manuelle, mais leur informatisation a permis d'automatiser la tâche. Cependant, les dictionnaires présentent des limites en tant que source d'information, étant soumis aux biais des éditeurs qui ont tendance à privilégier les œuvres littéraires classiques considérés comme les principales sources de citations. De plus, ils contiennent souvent des mots archaïques. Il est donc erroné de supposer que tous les mots existants sont présents dans le dictionnaire, ou que tous les mots présents dans le dictionnaire sont actuellement en usage. Cela résulte en un corpus relativement petit dont le contenu est trop normalisé pour permettre l'observation de constructions rares qui intéressent les morphologues (Hathout, 2009 ; Aronoff et Lindsay, 2014). Les journaux et les œuvres littéraires modernes, quant à eux, offrent la possibilité de découvrir des exemples rares et exceptionnels, tels que des néologismes ou des occasionalismes (Dal et Namer, 2012), qui reflètent l'utilisation active de la langue et la créativité lexicale des écrivains ou des journalistes. Bien que toutes ces unités ne soient pas nécessairement incluses dans les sources lexicographiques, elles peuvent potentiellement être référencées dans des corpus en ligne.

Un corpus est une collection de textes électroniques qui peut être constituée de différents types de contenus, tels que des œuvres littéraires et des journaux, mais aussi des transcriptions de conversations et d'émissions de radio, des sites web, des forums, etc. L'utilisation de corpus en ligne, par rapport aux dictionnaires, est justifiée par le fait que la morphologie y est employée de façon plus 'libérale' et plus variée que la morphologie présente dans les grammaires normatives (Ljaševskaja *et al.*, 2003). Les corpus linguistiques sont généralement conçus pour permettre différents types de recherches et possèdent à cet égard deux caractéristiques principales.

Premièrement, un corpus est caractérisé par sa représentativité, c'est-à-dire un contenu de textes équilibré, car un corpus ne peut pas contenir tous les textes disponibles dans une langue, mais il doit refléter au maximum l'utilisation de celle-ci dans toute sa variété de genres et de styles (Savčuk, 2011). Cela signifie que le corpus doit contenir, autant que possible, tous les types de textes, écrits ou parlés (littéraires, journalistiques, scientifiques, oraux, etc.) et que tous ces textes doivent être représentés proportionnellement à leur importance dans la langue à une période donnée.

Deuxièmement, un corpus est caractérisé par la présence d'informations supplémentaires sur les propriétés des textes, c'est-à-dire une annotation linguistique

(méta-textuelle, morphologique, syntaxique, phonologique, sémantique, etc.) qui le distingue des bibliothèques en ligne. Un corpus étiqueté de cette manière permet une recherche rapide de mots et de constructions en fonction de critères définis (Plungjan, 2005).

La troisième source de données, qui est encore plus volumineuse, est le Web. Contrairement aux corpus disponibles en ligne qui présentent une répartition équilibrée des genres textuels, le Web n'a pas été conçu pour être utilisé comme corpus pour les recherches linguistiques (Timberlake, 2004, p.7). Cependant, l'arrivée du Web a permis aux linguistes d'accéder à une masse de données d'une ampleur jamais atteinte auparavant. Le Web a ainsi été utilisé avec succès pour découvrir des données authentiques et des mots considérés auparavant comme peu probables, voire impossibles (Uth, 2010 ; Dal *et al.*, 2018 ; Fradin, 2019). Les avantages du Web incluent sa taille, la présence de nombreux types de textes différents (y compris certains qui ne peuvent être trouvés nulle part ailleurs), la variété des niveaux de langue utilisés, son évolution constante, etc. Une question souvent soulevée en linguistique de corpus concerne le rapport quantité/qualité : une plus grande quantité de données peut correspondre à un corpus moins contrôlé, et une augmentation de la quantité de données disponibles entraîne une augmentation du bruit (toutes les données inexactes, inutiles ou non pertinentes qui peuvent interférer avec l'interprétation des données significatives) (Hathout et Tanguy, 2002 ; Hathout *et al.*, 2008). Un moteur de recherche tel que Google, par exemple, représente une source de données très bruyante, contenant des faux positifs dus à des fautes d'orthographe, à des spams, à l'absence de marquage grammatical. Hathout et Tanguy (2002) ainsi que Aronoff et Lindsay (2014) recommandent de faire preuve de prudence lors du calcul des fréquences d'utilisation et des occurrences sur le Web. Aronoff et Lindsay (2014) proposent de ne pas accorder beaucoup d'importance aux chiffres bruts eux-mêmes et de limiter les enquêtes aux mots extraits. De plus, Resnik *et al.* (2005) et Hathout et Tanguy (2002) soutiennent qu'il est nécessaire d'utiliser des outils de recherche sur le Web suffisamment sophistiqués pour constituer une base de données fiable. Considérer le Web comme un corpus linguistique générique serait ainsi une erreur.

Dans la présente thèse nous allons utiliser le Corpus National de la Langue Russe (Plungjan *et al.*, 2005)[2], désormais RusCorpora, qui sera décrit en détail dans la section 4.1.3.

4.1.1.3 Méthodes de constitution de la base de données

Il y a deux méthodes pour constituer automatiquement une base de données pour les études morphologiques (Strnadová, 2014, pp.18,19). La première méthode pour récupérer des données lexicales consiste à utiliser une approche inductive pour chercher des patrons spécifiques. Cette méthode permet de récupérer une grande quantité de

[2]Il existe une variété de corpus russes disponibles en ligne tel que ruWaC ou Taïga, pour plus de détails, cf. Kopotev *et al.* (2017).

données, mais présente l'inconvénient de devoir attribuer, *a posteriori*, une base au dérivé. Il existe des outils automatisés pour attribuer une base, mais cela n'est pas très précis sans information sémantique.

La deuxième méthode pour récupérer des données lexicales consiste à utiliser une approche hypothético-déductive pour générer des candidats à partir d'un lexique de référence en utilisant des règles de création lexicale. Il existe deux stratégies pour générer des candidats morphologiquement corrects. La première, appelée stratégie minimale, consiste à générer uniquement des candidats qui respectent des contraintes morphophonologiques de bonne formation (Roché et Plénat, 2014). La deuxième, appelée stratégie maximale, à pour but de générer toutes les formes possibles, indépendamment de leur conformité aux principes de bonne formation. La première stratégie est basée sur la Théorie de l'Optimalité (Prince et Smolensky, 2004) et la seconde est basée sur l'hypothèse selon laquelle un locuteur peut ignorer ces contraintes dans une situation d'expression écrite spontanée (Huguin, 2021). L'existence des candidats est généralement vérifiée sur Internet. L'inconvénient de cette méthode est qu'elle ne prend en compte que les dérivés de façon régulière et dépend entièrement des règles utilisées pour générer les candidats.

Dans la présente étude, nous adopterons une méthodologie inductive et utilisons les données provenant de RusCorpora, pour extraire les adjectifs contenant les suffixes *-n-*, *-sk-* et *-Ov-* ainsi que leurs fréquences. Le processus d'extraction et de nettoyage des données sera abordé dans la section 4.1.4.

4.1.2 Études quantitatives

Dans la section 3.1, nous avons conclu que les différents aspects de la concurrence, tels que la productivité, la synonymie, le blocage – considérés auparavant comme catégoriels – sont maintenant plutôt considérés comme probabilistes. La linguistique informatique voit les langues naturelles comme des entités pouvant donner lieu à des calculs (Dal *et al.*, 2004). Les méthodes quantitatives semblent alors particulièrement appropriées pour analyser la concurrence affixale, compte tenu de sa nature scalaire et de ses aspects multidimensionnels.

Dans la littérature linguistique, la concurrence affixale est généralement étudiée à partir d'une perspective théorique, parfois à l'aide de données volumineuses, mais pas nécessairement accompagnée d'analyses quantitatives détaillées. Certaines études présentent des résultats statistiques descriptifs, basés sur des données spécifiquement collectées et analysées linguistiquement. Cependant, des études récentes ont de plus en plus recours à des méthodes quantitatives pour étudier les phénomènes de concurrence affixale. Les statistiques peuvent être basées sur une variété de modèles, tels que des modèles analogiques (Chapman et Skousen, 2005 ; Arndt-Lappe, 2014), des régressions logistiques (Bonami et Thuilier, 2019 ; Bobkova, 2022b), des arbres de décision (Naccarato, 2019) ainsi que des ensembles d'arbres (Gries, 2019 ; Bonami et Pellegrini, 2022), la sémantique distributionnelle (Varvara, 2020 ; Wauquier, 2020 ; Huyghe et Wauquier, 2021 ; Missud et Villoing, 2021 ; Bonami et Guzmán Naranjo, 2023), les

réseaux de neurones artificiels (Guzmán Naranjo, 2019 ; King *et al.*, 2020). D'autres auteurs font une comparaison entre les différentes méthodes (Baayen *et al.*, 2013 ; Allassonnière-Tang *et al.*, 2021). Les approches quantitatives semblent particulièrement appropriées pour décrire la concurrence affixale, car elles permettent de fournir des méthodes relatives et fréquentistes pour compléter les approches plus descriptives (Huyghe et Varvara, 2023). En particulier, les méthodes quantitatives permettent d'évaluer les degrés de concurrence, l'influence de différents facteurs sur la sélection d'affixes rivaux et proposent des analyses multifactorielles des situations de concurrence en morphologie, ce qui permet une meilleure compréhension de la notion de concurrence affixale.

La slavistique a également connu une croissance des études basées sur les corpus et des approches expérimentales. Par exemple, la productivité morphologique a été étudiée par Baayen *et al.*, 2013 ; Antic, 2012 et Sims et Parker, 2015. L'étude du profil lexical a été faite par Kuznetsova, 2015, où la distribution lexèmes dans le corpus est analysée. La distribution des sens des mots à travers un corpus a été analysée par Endresen *et al.*, 2012. Les profils grammaticaux des verbes russes ont été étudiés avec les méthodes de linguistique de corpus (Janda et Lyashevskaya, 2011). De nombreuses études ont également analysé le comportement linguistique des locuteurs, comme Kapatsinski, 2010 ou Sims, 2006.

Dans les études de la morphologie du russe, les méthodes numériques ont été également mises en place pour traiter la concurrence. Le modèle de régression logistique binaire (Sokolova *et al.*, 2012) a permis de prédire l'une des constructions verbales alternatives. Divjak et Arppe (2013) utilisent la régression logistique pour prédire le choix entre des verbes synonymes[3]. Baayen *et al.* (2013) analysent un modèle d'arbre d'inférence conditionnel pour choisir entre deux constructions verbales alternatives en russe. Bobkova (2022a,b) utilise la régression logistique pour identifier les propriétés des noms de base qui influencent le choix d'un suffixe adjectival et les méthodes basées sur les arbres pour analyser les erreurs de ces modèles. De plus, Bobkova et Montermini (2023) étudient les données des adjectifs dénominaux, y compris les doublets, à l'aide des forêts aléatoires.

Les méthodes numériques décrites précédemment permettent une analyse plus approfondie du langage naturel en utilisant des outils statistiques et informatiques. Comme cela a déjà été mentionné, de telles méthodes requièrent l'utilisation de corpus volumineux.

4.1.3 RusCorpora

RusCorpora, ou le Corpus National de la Langue Russe (Plungjan *et al.*, 2005), est une collection de textes électroniques reflétant les différents styles de la langue russe. Ces textes contiennent une grande quantité d'informations linguistiques et

[3]La régression logistique binaire s'applique dans le cas où il s'agit d'un choix entre deux catégories distinctes. Cette même méthode peut aussi être appliquée lorsqu'il y a plus de deux catégories possibles dans la classification.

méta-textuelles. La disponibilité de ces informations distingue ce corpus d'une simple collection de textes accessibles sur Internet. Le niveau de granularité et l'authenticité de ces informations, ainsi que la description des phénomènes linguistiques présents dans les textes constituent la valeur principale de RusCorpora (Poljakov, 2003). RusCorpora est un corpus en ligne qui, dans sa version actuelle, contient plus d'un milliard d'occurrences.

Comme tout corpus linguistique, RusCorpora est caractérisé par une représentativité de tous types de textes, écrits ou parlés et contient des textes littéraires, journalistiques, scientifiques, transcription de l'oral. De plus, RusCorpora contient cinq types d'annotation linguistique : méta-textuelle, morphologique, syntaxique, accentuelle et sémantique. Pour les objectifs de ce travail, nous nous intéresserons particulièrement à l'annotation morphologique.

RusCorpora couvre une période allant du milieu du XVIIIe siècle au début du XXIe siècle. Le corpus contient des œuvres littéraires d'importance culturelle, mais aussi d'autres textes d'importance linguistique : mémoires, essais, articles de journaux, textes scientifiques et de vulgarisation scientifique, correspondances, journaux intimes, documents, discours publics, corpus de dialectes, corpus syntaxique, corpus accentologique, corpus éducatif, corpus de l'ancien slave, ainsi que des corpus parallèles. Pour les besoins de cette étude, nous nous intéressons au russe standard, écrit et parlé. Par conséquent, les adjectifs ont été extraits de cinq sous-corpus suivants :

- Le corpus général représente le russe standard: il comprend des textes écrits allant des années 1950 à aujourd'hui (dont 40% de textes littéraires), des enregistrements réels de discours russes de la même période (dont la transcription a été vérifiée manuellement) et des textes plus anciens datant du milieu du XVIIIe siècle au milieu du XXe siècle, avec une proportion plus élevée d'œuvres littéraires en raison de leur plus grande disponibilité (Savčuk et Sičinava, 2006 ; Dič, 2003 ; Oskol'skaja, 2006 ; Savčuk, 2009) ;

- Le corpus média comprend des articles de journaux publiés entre 1990 et les années 2000. Les textes journalistiques font partie d'un corpus à part, puisque les textes de médias ne peuvent pas être intégralement inclus dans le sous-corpus général sans compromettre sa représentativité, tant thématique que chronologique. Les volumes des sous-corpus général et média sont similaires. Le corpus contient les textes des sept médias, en proportion égale : *Izvestija*, *Sovetskij Sport*, *Trud*, *Komsomol'skaja pravda*, *RIA Novosti*, *RBK*, *Novyj Region* ;

- Le corpus multimédia contient des extraits de films datant de 1930 à 2000 (Grišina, 2005 ; Grišina, 2009) ;

- Le corpus oral contient des transcriptions audio de discours publics et privés, ainsi que des films russes datant de 1930 à 2000. Les textes varient en fonction de leurs genres mais aussi de leur provenance géographique, les enregistrements ayant été réalisés dans différentes régions de la Russie. L'intérêt d'inclure les données du

corpus oral réside dans la volonté de découvrir les structures plus dynamiques de la langue vivante, le corpus écrit étant généralement plus conservateur (Grišina, 2003) ;

- Le corpus poétique. Ce corpus couvre la période entre 1750 et début des années 2000 (Plungjan *et al.*, 2009).

Actuellement, RusCorpora existe dans une nouvelle version[4], enrichie et améliorée. Cependant, la recherche des données pour cette thèse a été réalisée entre février 2017 et août 2022, lorsque l'ancienne version de RusCorpora a été maintenue[5]. Comme le processus d'extraction et de vérification des données est fastidieux, nous avons choisi de maintenir le travail avec les données extraites de l'ancienne version de RusCorpora qui, quant à elle, contenait plus de six cent millions d'occurrences.

Comme mentionné précédemment, les textes de RusCorpora ont été annotés automatiquement avec des tags morphologiques, sémantiques et méta-textuels. Ces étiquettes permettent une recherche plus précise dans le corpus. Les tags méta-textuels contiennent des informations sur les textes en entier, telles que l'auteur, le genre, le style, etc. (Savčuk, 2005). Les informations morphologiques incluent la forme de citation d'un lexème pour une forme fléchie donnée, les informations sémantiques et syntaxiques (animé/inanimé pour les noms, transitif/intransitif pour les verbes, etc.), ainsi que les valeurs morpho-syntaxiques des mots-formes (Ljaševskaja *et al.*, 2003). Les propriétés sémantiques incluent les informations intrinsèques aux lexèmes (genre des noms, sous-classes des adjectifs, etc.), les classes taxonomiques pour les noms concrets et abstraits (personnes, plantes, matériaux, etc.), ainsi que les spécifications sur les liens dérivationnels. L'ensemble de ces étiquettes permet une analyse linguistique plus approfondie (Kustova *et al.*, 2005).

Du point de vue des annotations, le corpus est divisé en deux parties. La majorité du corpus contient des textes annotés automatiquement, avec la désambiguïsation entre les lemmes, les mots-formes, les catégories lexicales et les propriétés grammaticales effectuée par la machine. Cependant, l'attribution des tags homonymes est moins fiable. En russe, il existe une homonymie courante entre certaines formes des noms, des adjectifs et des adverbes[6]. Cela peut causer des ambiguïtés lors de l'attribution automatique des tags morphologiques. Il ne faut cependant pas considérer les tags attribués automatiquement comme des erreurs, mais plutôt comme des hypothèses. Si certaines formes grammaticales sont homonymes, tous les tags possibles pour ces formes leur sont attribués (Ljaševskaja *et al.*, 2003). Ce choix permet une plus large couverture des résultats, mais introduit beaucoup de bruit. Une minorité du corpus (environ six millions de mots) a été désambiguïsée manuellement, les annotations sont ainsi plus détaillées et plus précises en cas d'homonymie grammaticale ou sémantique.

[4]Disponible à l'adresse suivante : `https://ruscorpora.ru/` ; dernier accès : avril, 2023.

[5]Cette version a été accessible à l'adresse suivante : `https://ruscorpora.ru/old/search-main.html` ; dernier accès : août 2022 ; la recherche n'est plus maintenue dans la majorité de corpus.

[6]Nous allons revenir sur les différents types d'homonymie dans la section 4.1.5.

Cette partie du corpus peut être référencée en tant qu'étalon-or (Poljakov, 2003). Cependant, si une forme est peu fréquente, il y a une faible probabilité de la retrouver dans ce sous-corpus restreint.

Malgré l'avantage que cet échantillon restreint aurait apporté pour la constitution de la base des adjectifs, il n'aurait pas été très grand en termes de volume. L'objectif de la présente recherche est de constituer un corpus exhaustif des adjectifs dénominaux, il est donc nécessaire d'interroger l'intégralité de RusCorpora.

4.1.4 Recueil des données

L'interface de RusCorpora permet de limiter la recherche à un sous-ensemble qui correspond à des caractéristiques grammaticales et/ou sémantiques précises. Ainsi, nous pouvons limiter la recherche à la classe grammaticale des adjectifs[7].

L'utilisateur peut rechercher des tokens exacts (une forme fléchie telle quelle) ou des lemmes (toutes les formes fléchies qui peuvent y correspondre seront alors présentées). Pour faciliter la recherche, l'ancienne version de RusCorpora permettait d'utiliser des expressions régulières. Dans le présent travail, nous utiliserons la recherche par lemmes (dans leur forme de citation) qui retourne une liste de phrases contenant les formes fléchies correspondantes à ces lemmes, ainsi qu'un tableau récapitulatif des lemmes avec la fréquence cumulée des occurrences de leurs tokens. Un exemple simplifié de recherche contenant le suffixe dérivationnel -*n*- et le suffixe flexionnel -*yj* est présenté dans le tableau 4.1.

C'est le tableau des lemmes avec leurs fréquences qui nous intéresse sur la page de recherche. Cependant, il existe un défi lié à la quantité de résultats obtenus. L'ancienne version de RusCorpora imposait une limite de 100 exemples par document. Lorsque la recherche est suffisamment large (comme dans le cas d'une recherche visant à identifier tous les adjectifs se terminant par un suffixe dérivationnel et un suffixe flexionnel à l'aide d'une expression régulière), cette limite de 100 exemples par document peut facilement être atteinte pour les œuvres littéraires. Pour contourner ce problème, notre approche consiste à introduire un préfixe dans l'expression régulière (un caractère supplémentaire correspondant graphiquement à chaque lettre de l'alphabet cyrillique). Cela multiplie le nombre de requêtes par 33 (le nombre de caractères dans l'alphabet cyrillique), mais minimise le rappel en cas de présence de plus de 100 adjectifs avec un suffixe spécifique dans un document donné.

L'expression régulière finale est ainsi la suivante :

```
*[а-я]{н,ск,ов,ев}{ый,ий,ой}
```

Parcourir manuellement des centaines, voire des milliers de pages de recherches et récupérer des tableaux de lemmes est une tâche qui ne peut être réalisée dans le cadre d'une thèse. Il est donc nécessaire de mettre en place une automatisation. En

[7]Comme mentionné plus haut dans la section 4.1.3, la majorité de RusCorpora a été annotée automatiquement, ce qui implique la présence de bruit.

Recherche	**ный* '*nyj'		
Trouvé	1 587 433 documents		
	4 438 252 tokens		
Corpus	générale		
Document 1	'*Интерфакс-Запад' прекратит работу в Белоруссии с 1 января.* "Interfaks-Zapad' cessera ses activités en Biélorussie à partir du 1er janvier.'		
Source	*Ведомости*, 2021.12 'Vedomosti'		
Exemple 1	*Телеканал создал* ***резервный*** *аккаунт на время, пока доступ к старому не восстановят.* 'La chaîne de télévision a créé un compte de secours pour la période où l'accès à l'ancien compte ne sera pas rétabli.'		
Exemple 2	***Официальный*** *представитель Министерства* ***иностранных*** *дел России Мария Захарова тогда заявила, что ситуация с модерацией контента на* ***официальных*** *американских IT-платформах достигла смыслового и технологического тупика.* 'La représentante officielle du Ministère des Affaires étrangères de la Russie, Maria Zakharova, a déclaré à l'époque que la situation de la modération du contenu sur les plate-formes IT américaines officielles était bloquée tant en termes de sens que de technologie.'		
Tokens		Lemmes	
резервный	1	*резервный*	1
официальный	1	*официальный*	2
иностранных	1	*иностранный*	1
официальных	1		

Tableau 4.1: Résultat simplifié d'une recherche effectuée sur RusCorpora

l'absence d'une API officielle fournie par RusCorpora permettant l'accès aux données et métadonnées du corpus, nous avons recours à la technique du web scraping. Deux librairies Python sont notamment disponibles pour cette tâche : `requests`[8] (pour accéder au contenu des pages web) et `BeautifulSoup`[9] (pour analyser ces pages et extraire les informations nécessaires en se basant sur la structure sémantique des balises HTML).

En raison de la grande quantité de pages avec résultats de recherche, un traitement automatisé supplémentaire a été effectué après l'extraction de données à partir du Web. Ce traitement a permis de concaténer les données et de supprimer les doublons

[8]Disponible à l'adresse suivante : `https://pypi.org/project/requests/`.
[9]Disponible à l'adresse suivante : `https://pypi.org/project/beautifulsoup4/`.

en conservant un seul résultat pour chaque lemme et en additionnant les fréquences. Cette extraction brute a abouti à une liste de 207 665 lemmes accompagnés de leurs fréquences respectives.

Plusieurs filtrages automatiques ont été mis en place pour éliminer les lemmes mal référencés. Les adjectifs préfixés ainsi que les adjectifs composés (avec un trait d'union) ont été retirés. Un grand sous-ensemble de données contiennent notamment les adjectifs formés avec des variantes de *-n-* et de *-sk-* (discutées dans la section 2.2.2), ces entrées ont été mises de côté. Cependant, les 'variantes suffixales' extraites automatiquement ont nécessité une vérification manuelle supplémentaire, car certains adjectifs se terminant graphiquement par des variantes étendues présentent en réalité des adjectifs formés avec *-n-* ou *-sk-* (146).

(146)
a. ∼ *-Onn-*
ТЕЛЕФОН 'téléphone' → ТЕЛЕФОННЫЙ
TELEFON TELEFONNYJ

b. ∼ *-ijsk-*
ИНДИЯ 'Inde' → ИНДИЙСКИЙ
INDIJA INDIJSKIJ

c. ∼ *-arn-*
ТОВАР 'marchandise' → ТОВАРНЫЙ
TOVAR TOVARNYJ

d. ∼ *-Ovn-*
ОСНОВА 'base' → ОСНОВНОЙ
OSNOVA OSNOVNOJ

e. ∼ *-ansk-*
ИСПАНИЯ 'Espagne' → ИНСПАНСКИЙ
ISPANIJA ISPANSKIJ

f. ∼ *-insk-*
ВОИН 'guerrier' → ВОИНСКИЙ
VOIN VOINSKIJ

g. ∼ *-al'n-*
ИДЕАЛ 'idéal' → ИДЕАЛЬНЫЙ
IDEAL IDEAL'NYJ

h. ∼ *-ičn-*
ОТЛИЧИЕ 'différence' → ОТЛИЧНЫЙ
OTLIČIE OTLIČNYJ

i. ∼ *-ozn-*
МОРОЗ 'gèle' → МОРОЗНЫЙ
MOROZ MOROZNYJ

La liste finale après une extraction et filtrages automatiques se compose de 57 066 entrées.

Après avoir implémenté les filtrages automatiques, nous avons procédé à la vérification de la fréquence des lemmes en utilisant une liste précise de ces derniers plutôt que des expressions régulières. Il est courant de constater que les fréquences extraites avec des expressions régulières ne correspondent pas aux fréquences réelles des lemmes, d'où la nécessité de cette vérification.

4.1.5 Validation des résultats

La vérification manuelle successive a permis de parcourir la liste d'adjectifs extraits et de conserver uniquement les adjectifs grammaticalement corrects et construits sur des bases nominales.

La définition de 'grammaticalement correct' peut cependant poser des problèmes. En particulier, les hapax (adjectifs ayant une fréquence 1) en (147a) ont été conservés puisqu'il s'agit des formes qui ne violent pas les principes de la dérivation adjectivale. Cependant, l'exemple en (147b) a été écarté, bien que sa forme soit référencée. Il s'agit d'un emploi de l'adjectif en fonction métalinguistique où il est indiqué, justement, comme 'non existant' (*Взяв же слово 'клеветник', мы его не можем разложить так же: прилагательного 'клеветной' не существует.* (*Vzjav že slovo 'klevetnik', my ego ne možem razložit' tak že: prilagatel'nogo 'klevetnoj' ne suščestvuet.*) 'En prenant le mot КЛЕВЕТНИК 'celui qui diffame', nous ne pouvons pas le décomposer de la même manière : l'adjectif КЛЕВЕТНОЙ n'existe pas).

(147) a. ЛАКМУС 'lacmus' → ЛАКМУСНЫЙ
LAKMUS LAKMUSNYJ
ТАЛМУД 'Talmud' → ТАЛМУДСКИЙ
TALMUD TALMUDSKIJ

b. КЛЕВЕТНИК 'diffamateur' → *КЛЕВЕТНОЙ
KLEVETNIK KLEVETNOJ

Les adjectifs construits à partir de lexèmes d'autres catégories lexicales que les noms ont été éliminés : déverbaux (148a), déadverbiaux (148b), pronominaux (148c). Les converts dont le thème finit par *-n-*, *-k-* ou *-v-*, ce qui correspond graphiquement aux séquences recherchées, ont aussi été écartés (148d). Les composés sans trait d'union ont été supprimés (148e). Les adjectifs pour lesquels une base nominale ne pouvait pas être identifiée ont également été retirés. Cela concerne les adjectifs d'origine slave dont la base n'est plus restituable en synchronie (148f) ou les adjectifs empruntés (148g). Les adjectifs formés à partir de sigles lus et épelés ont également été éliminés (148h), sauf dans les cas où l'abréviation a été assimilée par la grammaire russe et fonctionne comme un nom, et peut donc avoir des formes fléchies en fonction des cas (148j).

(148) a. РВАТЬ 'déchirer' → РВАНЫЙ
RVAT' RVANYJ

b. ТОГДА 'auparavant' → ТОГДАШНИЙ
TOGDA TOGDAŠNIJ

c. ИХ 'leur' → ИХНИЙ
IX IXNIJ

d. КАБАН 'sanglier' → КАБАНИЙ
KABAN KABANIJ

e. ДИКИЙ ЗВЕРЬ 'animal sauvage' → ДИКОЗВЕРСКИЙ
DIKIJ ZVER' DIKOZVERSKIJ

f. ? → ЛЕВЫЙ 'gauche'
LEVYJ

g. ? → МОБИЛЬНЫЙ 'mobile'
MOBIL'NYJ

h. КПСС 'parti communiste de l'Union Soviétique' → КПССНЫЙ
KPSS KPSSNYJ

i. ГИБЕДЕДЕ (>ГИБДД) 'inspection de sécurité routière' →
GIBEDEDE
ГИБЕДЕДЕШНЫЙ
GIBEDEDEŠNYJ

j. ВУЗ 'institut' → ВУЗОВЫЙ
VUZ VUZOVYJ
вуз 'institut$_{\text{M.NOM.SG.}}$' ∼
vuz
вуз-ы 'institut-M.NOM.PL.' ∼ *вуз-ов* 'institut-M.GEN.PL.'
vuz-y *vuz-ov*

Les fautes d'orthographe ont également été écartées (149). Nous avons conservé les lexèmes adjectivaux transcrits uniquement avec les caractères cyrilliques de l'alphabet russe actuel et avons retiré les lexèmes comportant les graphèmes en vigueur avant la réforme orthographique de 1918[10].

(149) *КУКОЛНЫЙ < КУКОЛЬНЫЙ 'poupée'
KUKOLNYJ KUKOL'NYJ
*НОРМАЛВНЫЙ < НОРМАЛЬНЫЙ 'normal'
NORMALVNYJ NORMAL'NYJ
*ОЗДНИЙ < ПОЗДНИЙ 'tard'
OZDNIJ POZDNIJ
*ИРЕКЛАМНЫЙ < *и рекламныи* 'et publicitaire'
IREKLAMNYJ *i reklamnyj*

[10]La reforme orthographique de 1918 avait pour objectif la simplification de l'écriture et de la lecture du russe, ainsi que l'accès facilité au peuple. Quatre graphèmes ont été supprimés : *ѣ, ѳ, i, ѵ.* : dans le contexte *въ однихъ съверныхъ провинціяхъ* (*v odnix severnyx provincijax*) 'dans certaines régions du nord' le mot-forme *съверн-ыхъ*$_{\text{F.PL.LOC}}$, qui contient des graphèmes d'avant la révolution, est lemmatisé comme СЕВЕРНЫЙ (SEVERNYJ) par RusCorpora, avec l'orthographe actuelle.

Outre les problèmes mentionnés précédemment, la forme adjectivale présente dans le tableau des lemmes peut résulter d'une mauvaise restitution de la forme initiale en raison de la superposition des paradigmes. Plusieurs cas de figure peuvent être cités.

Premièrement, la liste de lemmes contient des adjectifs du type *одиссеевый (odisseevyj) 'Odyssée', alors qu'il s'agit d'une erreur de lemmatisation de l'adjectif d'appartenance одиссеев (odisseev). De manière générale, les adjectifs d'appartenance ont été lemmatisés avec un suffixe flexionnel adjectival *-ij/-oj* qui correspond au nominatif singulier. Des fois le contexte s'avère insuffisant pour désambiguïser la catégorie des adjectifs en question, car les paradigmes des adjectifs (qualificatifs ou relationnels) sont parfois identiques à ceux des adjectifs d'appartenance. Par exemple, les contextes en (150a) ne permettent pas de déterminer s'il s'agit des adjectifs d'appartenance ou pas. Cependant, si l'adjectif apparaît au nominatif, l'ambiguïté est levée (150b). Nous avons sélectionné uniquement les adjectifs dont le caractère qualificatif ou relationnel était non ambigu.

(150) a. Альпы 'Alpes' → альпов / альповый
Al'py al'pov al'povyj
у альповой лачуги 'près d'une cabane alpine'
u al'povoj lačugi
анаконда 'anaconda' → анакондов / анакондовый
anakonda anakondov anakondovyj
своим анакондовым зрением 'avec son regard d'anaconda'
svoim anakondovym zreniem
Бенеттон 'Benetton' → бенеттонов / бенеттоновый
Benetton benettonov benettonovyj
в лучах бенеттоновой витрины 'dans la lueur de vitrine Benetton'
v lučax benettonovoj vitriny

b. Стинг 'Sting (chanteur)' → стинговый
Sting stingovyj
стинговый хит 'hit de Sting'
stingovyj xit

Les adjectifs d'appartenance ne sont pas les seuls à contenir la séquence *-Ov* en fin de mot, qui est mal restituée comme un adjectif dans sa forme longue. Les noms propres en *-Ov* se trouvent également dans les résultats de recherche, cependant ils ne présentent pas d'intérêt pour notre étude (151a). Les mêmes problèmes d'extraction concernent les noms propres finissant par *-sk(ij)*, qui ont aussi été éliminés (151b).

(151) a. гречка 'sarrasin' → *гречковый <Гречков
grečka grečkovyj Grečkov
А в шахматы он играл обычно со своим ассистентом Гречковым
A v šaxmaty on igral obyčno so svoim assistentom Grečkovym
'Et aux échecs, il jouait habituellement avec son assistant Grečkov'

b. волокно 'fibre' → *волоконский <Волоконский
volokno volokonskij Volokonskij
рассказывает сосед Константин Волоконский
rasskazyvaet sosed Konstantin Volokonskij
'raconte le voisin Konstantine Volokonskij'

Les adjectifs mal restitués qui correspondent aux noms géographiques (souvent non dérivés) finissant par *-sk(ij)*, *-sk(oe)*, *-Ovo*, *-Ov(oe)* ont été également éliminés (152).

(152) *пресновский < Пресновский 'Presnovskij (région au Kazakhstan)'
presnovskij Presnovskij
*петровский < Петровское 'Petrovskoe (lieu-dit)'
petrovskij Petrovskoe
*дружковый < Дружково 'Družkovo (station)'
družkovyj Družkovo
*мелеховый < Мелеховое 'Melixovoe (village)'
melexovyj Melexovoe

Certains adjectifs mal restitués correspondent à la forme du génitif qui contient *-Ov-* (153a). De plus, des cas de gérondifs ont également été détectés (153b).

(153) a. *зятевой < *зятев-ей* 'gendre-M.GEN.PL'
zjatevoj *zjatev-ej*
[...] нет ли в доме снох, зятевей
[...] net li v dome snox, zjatevej
'[...] y a-t-il des belles-filles ou des gendres à la maison'
*апгрейдовый < *апгрейд-ов* 'upgrade-M.GEN.PL'
apgrejdovyj *apgrejd-ov*
[...] операционная система потребует новых компьютеров, новых апгрейдов
[...] operacionnaja sistema potrebuet novyx komp'juterov, novyx apgrejdov
'[...] le système d'opération nécessitera de nouveaux ordinateurs et des mises à niveau'

b. *омрачневый < *омрачневая* 'assombrir$_{GRD}$' 'assombrir'
omračnevyj *omračnevaja*
И вдруг, от слога к слогу всё более и более омрачневая и на последнем, как туча
I vdrug, ot sloga k slogu vsë bolee i bolee omračnevaja i na poslednem, kak tuča
'Et soudain, de syllabe en syllabe, tout devenait de plus en plus sombre, et sur le dernier – comme un nuage d'orage'

Un autre cas de figure concerne les adjectifs substantivés. Les adjectifs d'origine étrangère, non dérivés, désignent souvent des familles ou des classes de plantes ou

d'animaux (154a). Ces cas ont été éliminés de la liste. Cependant, certains adjectifs peuvent apparaître en tant que tels (154b) ou en tant qu'adjectifs substantivés (154c). Ces cas ont été préservés malgré la confusion potentielle des fréquences extraites.

(154) a. *АСКОБОЛОВЫЙ < АСКОБОЛОВЫЕ 'ascobolaceae'
ASKOBOLOVYJ ASKOBOLOVYE

b. СТОЛ 'table' → СТОЛОВЫЙ 'table$_{\text{REL}}$'
STOL STOLOVYJ
столовые приборы 'les couverts de table'
stolovye pribory

c. СТОЛОВАЯ КОМНАТА 'salle à manger' → СТОЛОВАЯ 'cantine'
STOLOVAJA KOMNATA STOLOVAJA
Я был в столовой 'J'ai été à la cantine'
Ja byl v stolovoj

Finalement, l'extraction brute présente quelques adjectifs formés avec le même suffixe dérivationnel mais dont le radical présente des allomorphies (155). L'adjectif qui est utilisé plus fréquemment est celui qui a été conservé dans la base de données.

(155) ВАРЯГ 'Varègue' → ВАРЯГСКИЙ / ВАРЯЖСКИЙ
VARJAG VARJAGSKIJ VARJAŽSKIJ

La composition de la base de données extraite et filtrée automatiquement est présentée dans le tableau 4.2. Les adjectifs dérivés à partir des noms qui nous intéressent pour cette étude représentent 21.32% des données (`OK`). Un pourcentage plus élevé est constitué par les noms propres (noms de familles, toponymes) – presque 25% (`PROP`). Le taux des adjectifs composés (`COMP`) et des adjectifs avec des particularités orthographiques (`ORTH`), tels que les fautes d'orthographe ou les transcriptions avec les caractères d'avant 1918, est également élevé (autour de 17% et 14% respectivement). Les adjectifs formés sur d'autres catégories lexicales ainsi que les converts (`AUTRE`), les noms déclinés (`DECL`), les adjectifs formés avec le même suffixe dérivationnel présentant des allomorphies (`VAR`) et les adjectifs non construits (`SIMPLE`), sont les moins représentés. Finalement, la catégorie `KO` regroupe les adjectifs invalides et les erreurs de lemmatisation.

La composition de la base de données RuDénom (adjectifs étiquetés `OK`) est présentée dans le tableau 4.3.

La base de données RuDénom qui contient la totalité des adjectifs et de leurs noms de base (12 592 entrées) sera examinée au chapitre 8. Avant de présenter les trois sous-ensembles de données, tirés de cette base générale, qui seront analysés dans les chapitres 7 et 9, nous allons aborder la question de couverture de RuDénom.

Afin d'évaluer la couverture de notre base de données des adjectifs dénominaux, nous l'avons comparée à deux sources : un dictionnaire de fréquences (Ljaševskaja et Šarov, 2009) où nous avons seulement considéré les données des adjectifs et une

Évaluation		Volume
OK	12 592	0.2132
PROP	14 741	0.2495
COMP	10 189	0.1724
ORTH	8 096	0.1370
AUTRE	3 426	0.0580
DECL	1 255	0.0212
VAR	186	0.0031
SIMPLE	553	0.0094
KO	6 028	0.1020
Total	57 066	

Tableau 4.2: Composition de la base de données après les filtrages

Suffixe	Total
-*n*-	5 039
-*sk*-	4 643
-*Ov*-	2 910
Total	12 592

Tableau 4.3: Composition de RuDénom

liste d'adjectifs néologiques (Alekseeva, 2011). Compte tenu de l'hétérogénéité des ressources, nous avons préalablement filtré les deux bases en enlevant les adjectifs simples, ainsi que les adjectifs dérivés ayant des suffixes différents de -*n*-, -*sk*- et -*Ov*-.

Le dictionnaire des fréquences de Ljaševskaja et Šarov (2009) a été élaboré à partir de données de RusCorpora et contient les lexèmes les plus fréquents figurant dans le corpus. Tous les adjectifs pertinents pour notre étude qui figurent dans ce dictionnaire sont également présents dans RuDénom, ce qui permet d'atteindre une couverture de 100% de notre base de données.

Les données de la liste des adjectifs néologiques proviennent, quant à elles, de diverses sources telles que des journaux et magazines imprimés et en ligne, des sites Web, des flyers publicitaires, des livres de fiction, des dictionnaires, de la télévision et de la radio. La couverture de RuDénom par rapport à Alekseeva (2011) est moins élevée comparé au dictionnaire des fréquences, avec 82.69% de données en commun. Parmi les cas non extraits de RusCorpora, il y a, par exemple, des données de doublets : риелтерский (rielterskij) 'agent immobilier' et тинейджерский (tinejdžerskij) 'adolescent' figurent dans notre liste tandis que leurs doublets en -*n*- cités par Alekseeva (2011) sont absents : риелтерный (rielternyj) et тинейджерный (tinejdžernyj).

	Dict. freq.	Liste adj. néo.
RuDénom	100.00%	82.69%

Tableau 4.4: Couverture de RuDénom

4.2 Sous-corpus de données

Les trois sous-ensembles de données que nous avons constitués sont formés par des adjectifs de haute et basse fréquence, ainsi que des données de doublets. L'importance de l'étude des doublets a été justifiée dans la section 3.1.3. Pour la suite, nous allons nous focaliser sur les lexèmes de haute et basse fréquence.

4.2.1 Considérations méthodologiques

Chaque locuteur d'une langue possède un stock de mots dans sa mémoire qui peut varier en fonction de plusieurs facteurs tels que l'âge, le lieu de résidence, le niveau d'instruction, etc.

Certaines théories linguistiques, comme celles défendues par les structuralistes et génératives américains (Bloomfield, 1933 ; Chomsky, 1965), adoptent une vision restrictive du lexique, qui se limite, selon eux, aux informations idiosyncrasiques. Ces théories ont été catégorisées comme 'à entrée appauvrie'[11] par Jackendoff (1975). D'autres approches, représentées par Jackendoff (1975), Booij (2010) et Bybee (1985), par exemple, acceptent la notion de redondance dans le lexique, considérant ce dernier comme le domaine où opèrent les règles morphologiques, qu'elles soient dérivationnelles ou flexionnelles. Ces perspectives sont qualifiées de théories 'à entrée complète'[12]. Selon cette vision, le lexique ne se contente pas de stocker des informations non redondantes, mais intègre également des facteurs tels que la fréquence d'utilisation. Halle (1973), par exemple, affirme que la présence d'un élément dans le lexique est une question de fréquence[13].

Beaucoup de théories linguistiques considèrent le lexique comme un ensemble de données passives stockées dans la mémoire à long terme. Les stocks individuels, soit les lexiques mentaux individuels, selon la terminologie utilisée par Aronoff et Anshen (2017), varient d'une personne à l'autre et au fil du temps. Cependant, il existe un noyau commun de mots connus par la majorité des locuteurs d'une langue, ce qui est nécessaire pour la communication entre les individus. Dans la modélisation du lexique en tant que dictionnaire proposée notamment par Pustejovsky (1998) une entrée lexicale est considérée comme une liste d'informations. Se fondant sur les critères de l'usage, de la richesse sémantique et de la productivité de dérivations, le noyau dur renferme les lexèmes les plus fréquents et permet ainsi d'avoir une vision approximative de la langue russe standard contemporaine. En ce qui concerne les lexèmes construits, les

[11] *Empoverished entry* (Jackendoff, 1975).
[12] *Full entry* (Jackendoff, 1975).
[13] Cf. Montermini (2019) sur la question des théories du lexique.

règles morphologiques productives qui s'appliquent à une grande partie d'entre eux ont tendance à être largement applicables lors de la formation de nouveaux lexèmes. Par conséquent, plus il y a de lexèmes distincts ayant un exposant donné, plus ils servent de modèles pour la formation de nouveaux mots. Les lexèmes de haute fréquence sont donc représentatifs d'une langue donnée, ils font généralement partie des dictionnaires.

Les effets cognitifs de lexèmes de haute fréquence sont étudiés notamment par Bybee (2002). Ainsi, l'effet de conservation dépend de la répétition des formes linguistiques, ce qui renforce leurs représentations dans la mémoire, les rendant plus facilement accessibles. Les mots à haute fréquence sont plus rapidement reconnus que les mots à basse fréquence. L'effet d'autonomie se manifeste lorsque l'on accède aux mots très fréquents indépendamment des éléments qui les constituent. En morphologie dérivationnelle cela se traduit par la mémorisation d'un mot construit en entier dans le lexique (Stemberger et MacWhinney, 2004) et, occasionnellement, par un écart sémantique entre les lexèmes construits et leurs bases (lexicalisation). Lindsay et Aronoff (2013) appuient cette idée avec les exemples de HISTORIC / HISTORICAL 'historique' et ELECTRIC / ELECTRICAL 'électrique', qui sont tous des lexèmes de haute fréquence en anglais. Le terme HISTORICAL est utilisé pour faire référence à un événement dans l'histoire passée, tandis que le terme HISTORIC peut également être utilisé pour décrire un moment historiquement significatif, même si l'événement est en cours. Le terme ELECTRIC décrit les éléments qui produisent, transmettent ou fonctionnent avec l'électricité, tandis que le terme ELECTRICAL peut également faire référence à des choses ou des personnes impliquées dans l'électricité. Dans cet exemple, deux affixes concurrents coexistent car les lexèmes construits se distinguent sémantiquement et se stabilisent en tant que formes distinctes. En résumé, les lexèmes de haute fréquence suivent le principe d'autonomie, c'est-à-dire qu'ils sont stockés dans le lexique en tant qu'unités entières, même s'ils sont construits morphologiquement. En revanche, les lexèmes de basse fréquence sont moins autonomes et sont donc atteints via leurs constituants.

Les lexèmes de haute et de basse fréquence ne posséderaient pas exactement les mêmes propriétés. Les lexèmes de haute fréquence, le lexique institutionnalisé, constituent des zones de grande stabilité morphologique, tandis que les lexèmes de basse fréquence et les hapax constituent des zones de grande variabilité. Les néologismes, les occasionnalismes, les emprunts introduisent constamment des changements aléatoires dans le système. Même si tous les lexèmes de basse fréquence ne sont pas des néologismes ou des occasionnalismes, ces derniers sont toutefois repérés majoritairement parmi les mots de très basse fréquence (Plag, 2018).

La distinction entre les occasionalismes et les néologismes ne fait pas l'unanimité. Selon Haspelmath et Sims (2013, p.71), les lexèmes inédits, non observés auparavant dans la langue, sont définis comme des néologismes. Ces derniers, lorsqu'ils ne trouvent pas une adoption généralisée et se cantonnent à des occurrences sporadiques, sont nommés occasionalismes. De son côté, Bauer (1983, p.45) met en avant que les occasionalismes sont des termes spontanément créés par les locuteurs pour répondre à

un besoin communicatif spécifique. Il s'agit de la première étape dans la vie d'un mot, à la suite de laquelle ce dernier peut ou non obtenir le statut de néologisme.

En slavistique, le distinction entre les termes 'néologisme' et 'occasionnalisme' repose sur la distinction entre la 'langue' et 'parole' (Lopatin, 1977 ; Zemskaja, 2011). Les néologismes sont des mots nouveaux introduits dans une langue, tandis que les occasionnalismes correspondent à des mots nouveaux utilisés dans un discours spécifique. Cette distinction est fondée sur l'inclusion ou non de ces mots dans les dictionnaires ainsi que sur leur conformité aux règles morphologiques en vigueur dans la langue. Ainsi, un occasionnalisme est une unité lexicale unique propre à l'usage individuel d'un locuteur, dénuée de reproductibilité. Il dépend fortement du contexte de son utilisation. Le néologisme est caractéristique d'une période donnée de développement de la langue et de la société. Une autre particularité des occasionnalismes consiste en ce qu'ils sont généralement formés en violant consciemment la norme linguistique dans un but expressif. Zemskaja (2011, pp.238-239) considère cette divergence de la norme comme un choix non conventionnel au niveau des lexèmes de base, avec des propriétés formelles, syntaxiques et sémantiques inhabituelles par rapport à la règle de construction choisie.

Si les mots de haute fréquence sont généralement référencés dans des dictionnaires, les néologismes et les occasionnalismes sont repérés uniquement dans des corpus massifs en ligne et sur le Web, ce qui permet d'observer des phénomènes qui seraient autrement passés inaperçus. Il est à noter que les néologismes finissent par rentrer dans les dictionnaires, mais cela dépend du choix des lexicographes et des éditeurs, qui décident quand un néologisme peut être référencé. Les corpus et le Web restent ainsi des sources primaires pour repérer les néologismes.

Ces mots ne font pas partie du lexique des locuteurs, et même si ces derniers ont été exposés à un lexème de basse fréquence dans leur pratique linguistique, le mot en question n'est pas mémorisé comme une unité et reste donc un mot potentiel (Aronoff et Anshen, 2017, p.181). Les mots potentiels sont définis par Vinokur (1959) comme des mots qui 'n'existent pas' en réalité, mais qui pourraient exister si la langue l'avait souhaité. La notion d'existence des mots est toutefois débattue. Ainsi, Rainer (2012) complète la dichotomie des mots réels et potentiels par un troisième mode d'existence : mots virtuels. Alors que les mots potentiels correspondent aux bases possibles des règles de formation de mots, les mots virtuels (soit les mots bloqués) ne peuvent jamais participer à la dérivation.

Les lexèmes de basse fréquence présentent un intérêt particulier pour une étude linguistique. Tout d'abord, ils reflètent la créativité linguistique des locuteurs, qui peuvent utiliser des constructions de manière non conventionnelle. Deuxièmement, ces lexèmes sont moins touchés par la lexicalisation, les glissements sémantiques, la perte de transparence et la polysémie. De plus, les lexèmes peu fréquents, en particulier les hapax, permettent une annotation basée sur des occurrences réelles plutôt que sur des lexèmes abstraits, ainsi qu'un examen en lien avec leur contexte de production (Dal et Namer, 2016).

Finalement, les hapax présentent un intérêt linguistique pour l'étude de productivité des suffixes. Par exemple, Baayen (1993) définit la productivité comme le nombre de hapax attestés dans un corpus pour un affixe donné : plus le nombre de hapax est élevé, plus la règle est productive. Baayen (2009) utilise également le nombre de hapax mais affine les calculs et définit la productivité comme le rapport entre le nombre de hapax et la fréquence totale de réalisations de lexèmes dans un corpus donné. Dans les deux cas, la liste des hapax est cruciale pour définir la productivité.

En raison des caractéristiques propres aux lexèmes de haute et de basse fréquence (dorénavant, nous considérerons les termes 'basse fréquence' et 'hapax' comme étant équivalents et interchangeables), nous opterons pour une analyse séparée des deux sous-ensembles de données.

4.2.2 Caractéristiques des données

L'objectif de cette section est de présenter trois sous-ensembles distincts : les adjectifs de haute fréquence, les adjectifs de basse fréquence et les adjectifs doublets. Ces derniers représentent 2 036 entrées de la base de données RuDénom, et nous allons les examiner plus en détail à la fin de cette section. Les 10 556 adjectifs restants sont dérivés d'une seule base. C'est à partir de cet ensemble que nous constituerons les sous-corpus d'adjectifs de haute et de basse fréquence.

En ce qui concerne les hapax, leur fréquence des types est identique à leur fréquence des tokens[14]. Cependant, il est encore nécessaire de déterminer la composition des données de haute fréquence.

En lexicographie, il est fréquent de définir un seuil pour la constitution des dictionnaires de fréquences, limitant ainsi les lexèmes en fonction de leur fréquence d'apparition dans un corpus (Ljaševskaja et Šarov, 2009, pp.1-21). Cette méthode permet de déterminer le noyau dur des lexèmes d'une langue donnée. Bybee (2002, p.17) met cependant en évidence que cette tâche n'est pas triviale. Si on cherche à diviser les tokens en deux groupes, l'un de haute fréquence et l'autre de basse fréquence, le groupe de haute fréquence peut comporter très peu de types distincts. À l'inverse, si l'on place la moitié des types dans le groupe de haute fréquence et l'autre moitié dans le groupe de basse fréquence, le nombre de tokens dans le groupe de haute fréquence sera considérablement plus élevé que dans le groupe de basse fréquence. La plupart des études adoptent un compromis entre ces deux positions en cherchant un écart naturel dans les classements de fréquence qui place environ 30 à 50% des tokens dans un groupe et 50 à 70% des tokens dans l'autre.

Cependant, le but de constitution des sous-ensembles de données de haute et basse fréquence n'est pas centré sur un examen des fréquences en elles-mêmes, mais plutôt sur les propriétés des noms de base qui déterminent le choix d'un suffixe particulier[15].

[14]La fréquence des types fait référence au nombre de lexèmes uniques contenant un affixe, tandis que la fréquence des tokens fait référence à la fréquence cumulée de tous les mots-formes de ces lexèmes.

[15]Le chapitre 8 se concentrera sur l'étude des fréquences sur l'ensemble de la base de données, sans prendre en compte les sous-corpus de haute et basse fréquence.

Ainsi, le nombre de tokens dans les deux groupes n'a pas une importance cruciale dans notre démarche. Pour constituer notre sous-ensemble de haute fréquence, nous avons opté pour un nombre de types plus ou moins équivalent à celui que l'on retrouve parmi les hapax. Le choix de 100 comme seuil de fréquence est justifié par le fait que les deux sous-ensembles de données sont représentés par un nombre approximativement équivalent de types. Les distributions des fréquences dans le corpus général sont présentées dans le tableau 4.5. Le sous-ensemble de haute fréquence est composé de 2 593 lexèmes distincts, tandis que le sous-ensemble de basse fréquence – de 2 525 lexèmes.

Sous-corpus	Types (V)	Tokens (N)
$F \geq 100$	2 593	11 288 286
$100 > F > 1$	5 438	98 136
$F = 1$	2 525	2 525
Total	10 556	11 398 947

Tableau 4.5: Distribution des types et des tokens

La distribution des suffixes dans les corpus de haute et basse fréquence, ainsi que dans les fréquences intermédiaires est présentée dans le tableau 4.6.

Suffixe	H Freq	$100 \geq F > 1$	B Freq	Total
-*n*-	1 234	2 135	708	4 077
-*sk*-	798	2 147	1 285	4 230
-*Ov*-	561	1 156	532	2 249
Total	2 593	5 438	2 525	10 556

Tableau 4.6: Distribution des types réelle par suffixe

Ce sont les données de haute et basse fréquence qui nous intéressent plus particulièrement. La distribution du suffixe -*Ov*- est pratiquement identique dans les deux sous-corpus. En revanche, les tendances pour les suffixes -*n*- et -*sk*- sont inversées : le suffixe -*n*- est le plus utilisé dans les données de haute fréquence (1 234 adjectifs), tandis que le suffixe -*sk*- prend de l'importance dans les données de basse fréquence (1 285 adjectifs). Cette observation peut être expliquée en partie par des raisons sémantiques, notamment la formation d'adjectifs hapaxiques à partir de toponymes. Ce phénomène sera examiné plus en détail dans la section 5.2.4.

Il convient de souligner que les chiffres présentés dans le tableau 4.6 reflètent la réalité de la distribution des adjectifs dans le corpus extrait à partir de RusCorpora. Par contre, ces chiffres sont préliminaires pour l'étude de la concurrence suffixale : dans la section 6.2, une méthode numérique basée sur les représentations vectorielles des mots sera présentée pour l'annotation sémantique. Cependant, selon les modèles utilisés, tous les substantifs qui correspondent aux noms de base des adjectifs de notre corpus n'y sont pas référencés, ce qui entraîne une réduction du nombre de données.

Les choix méthodologiques seront explicités dans la section 6.2. Le tableau 4.7 présente la distribution finale des données en anticipant sur ces choix méthodologiques.

Suffixe	H Freq	B Freq
-n-	1 192	439
-sk-	773	549
-Ov-	529	249
Total	2 494	1 237

Tableau 4.7: Distribution des types pour l'étude de la concurrence par suffixe

L'utilisation des lexèmes qui sont également référencés dans les modèles vectoriels pré-entraînés a résulté en une diminution considérable des données dans le sous-corpus des basses fréquences, ce qui a en particulier réduit la prédominance de l'utilisation du suffixe *-sk-*. En revanche, la distribution des suffixes dans les données de haute fréquence reste globalement identique.

Pour conclure la discussion sur les caractéristiques des données, il reste à aborder les données de doublets dont la distribution est présentée dans le tableau 4.8.

Suffixes	Bases uniques
-n-/*-Ov-*	584
-n-/*-sk-*	336
-sk-/*-Ov-*	45
-n-/*-sk-*/*-Ov-*	30
Total	995

Tableau 4.8: Distribution des données de doublets

Comme nous l'avons souligné dans la section 3.2, la concurrence entre les affixes *-n-* et *-Ov-* est mise en avant dans la littérature scientifique, tandis que la compétition entre les affixes *-n-* et *-sk-* est moins fréquemment abordée ; la concurrence entre les affixes *-sk-* et *-Ov-* est encore moins étudiée. Nous avons extrait de la base de données RuDénom tous les adjectifs construits à partir du même nom, et les volumes d'échantillons obtenus reflètent les tendances observées dans la littérature. En effet, les doublets formés avec les suffixes *-n-*/ *-Ov-* sont les plus fréquents (584 couples), suivis par les doublets en *-n-*/ *-sk-* (336 couples), tandis que les doublets en *-sk-*/ *-Ov-* sont les moins représentés (45 couples). Nous avons également observé les triplets formés à partir de 30 noms de base.

Ces données sont assez hétérogènes en termes de fréquences. Nous allons illustrer quelques cas sur l'exemple de triplets. On y retrouve les adjectifs qui sont tous très fréquents (156a) ou tous peu fréquents (156b). En majorité, cependant, les fréquences sont mixtes (156c).

(156) a. мир ‘monde/paix’
MIR
→ мирный (38 421) / мировой (75 067) / мирской (332)
MIRNYJ MIROVOJ MIRSKOJ

b. колледж ‘collège’
KOLLEDŽ
→ колледжный (1) / колледжевый (1) / колледжский (5)
KOLLEDŽNYJ KOLLEDŽEVYJ KOLLEDŽSKIJ

c. тюрьма ‘prison’
TJUR’MA
→ тюремный (10 816) / тюрьмовой (1) / тюремский (1)
TJUREMNYJ TJUR’MOVOJ TJUREMSKIJ
крокодил ‘crocodile’ →
KROKODIL
крокодиловый (907) / крокодильный (2) /
KROKODILOVYJ KROKODIL’NYJ
крокодильский (26)
KROKODIL’SKIJ

Le premier exemple en (156a) mérite une attention particulière. Nous avons marqué deux traductions possibles pour la base nominale мир : ‘monde’ et ‘paix’. Il s’agit d’homonymes, deux lexèmes distincts malgré leur identité formelle : мир$_1$ et мир$_2$. Nous avons évoqué ces cas dans la section 3.1.3 de l’état de l’art ; nous y reviendrons également dans la section 9.2.1 en les étudiant de manière plus approfondie. De manière générale, la question des fréquences de doublets sera discutée en détail dans la section 9.1.2.

Conclusion

Dans ce chapitre, nous avons présenté la méthode de collecte de données en décrivant différentes sources utilisées pour obtenir des adjectifs dénominaux. Nous avons choisi de travailler avec les corpus en ligne, notamment RusCorpora, plutôt que les dictionnaires en raison de leur inventaire limité de lexèmes et avons écarté le Web en raison de la recherche moins contrôlée qui peut conduire à une proportion de bruit considérable. L’utilisation d’un corpus en ligne permet d’accéder à un grand nombre de données, y compris les formations non attestées dans les dictionnaires, tout en minimisant le bruit par rapport au Web.

Nous avons adopté une méthode inductive pour extraire le maximum d’adjectifs avec les suffixes *-n-*, *-sk-* et *-Ov-* ainsi que des informations sur leurs fréquences à partir de RusCorpora. Malgré les filtrages automatiques mis en place, cette base de données a nécessité une vérification manuelle considérable pour écarter les faux positifs. Nous avons également comparé la couverture de notre base de données par rapport aux

autres sources lexicographiques, et les taux sont très élevés.

Enfin, nous avons constitué trois sous-corpus de données à partir de cette base de données générique : un corpus d'adjectifs de haute fréquence, avec une fréquence supérieure à 100 ; un corpus de basse fréquence, soit un corpus de hapax, où chaque adjectif n'est attesté qu'une seule fois ; et le corpus de doublets (ainsi que triplets), les adjectifs qui sont formés à partir de la même base nominale.

Les corpus de haute et basse fréquences seront examinés dans le chapitre 7 ; la base de données générique RuDénom sera examinée dans le chapitre 8 ; les doublets seront examinés dans le chapitre 9. Dans les chapitres 5 et 6 suivants, nous nous concentrerons sur les noms de base des adjectifs, l'annotation de leurs propriétés, à la fois manuelle et automatique.

Chapitre 5

Données des noms de base

Sommaire

Introduction

Nous avons recueilli des données adjectivales de haute et de basse fréquence, ainsi que des doublets. Ce chapitre portera sur les questions et les choix méthodologiques liés à leurs noms de base, ainsi que sur l'annotation des propriétés linguistiques de ces noms.

Comme nous étudions des adjectifs dénominaux, un nom de base peut être identifié pour chaque adjectif de notre corpus. Cependant, certains noms de base ne sont pas facilement identifiables en raison d'un décalage entre la forme et le sens du dérivé, ou parce qu'un adjectif peut être motivé selon plusieurs noms ou selon un nom et un verbe. Nous examinerons ces questions, ainsi que le sujet de l'homonymie et de la polysémie des noms de base dans la section 5.1. La section 5.2 sera consacrée à

l'annotation et à l'analyse des propriétés phonologiques, morphologiques, sémantiques et étymologiques des noms de base, ainsi qu'à des statistiques descriptives.

5.1 Identification des noms de base

La liste des adjectifs de la base de données RuDénom a été complétée automatiquement avec une liste de bases nominales. Cette liste a été créée en retirant progressivement un suffixe flexionnel et un suffixe dérivationnel des adjectifs. Dans certains cas, cette méthode fournit un résultat précis au niveau formel, car le radical obtenu correspond à un thème de la base nominale, qui, à son tour, est graphiquement identique à sa forme de citation (c'est le cas des noms masculins se terminant par une consonne au NOM.SG.). Toutefois, en raison des allomorphies typiques du russe, une modification manuelle a été parfois nécessaire. De plus, des ajouts manuels, tels qu'une flexion pour les noms féminins et neutres, ainsi qu'un signe mou pour certains noms masculins et féminins, ont été nécessaires pour obtenir la base nominale dans sa forme de citation. Cependant, la base formelle obtenue de cette manière n'est pas toujours unique ni suffisante.

5.1.1 Double motivation

La condition nécessaire pour une étude de la dérivation en synchronie est la présence d'une base pour chaque dérivée : seuls les mots qui peuvent être considérés comme morphologiquement construits en synchronie font l'objet d'étude (Zemskaja, 2011, pp.65-71 ; Lopatin, 1977, p.19). Dans certains cas, cependant, l'identification de la base peut être difficile.

5.1.1.1 Base formelle et sémantique

Premièrement, il peut y avoir un décalage entre la base formelle et la base sémantique : en français, par exemple, ROYALISTE dérive formellement de l'adjectif ROYAL par l'ajout du suffixe -*iste*, mais il est sémantiquement lié au nom ROI plutôt qu'à l'adjectif ROYAL (Roché, 2010). En parlant de ces cas sur l'exemple de russe, Švedova (1980, pp.132-133) se sert du terme 'motivation indirecte', contrairement à la 'motivation directe'. Les deux peuvent être exemplifiées par les gentilés. Certains d'entre eux sont dérivés directement d'un nom de lieu ; on n'observe alors aucun décalage sémantique ni formel. C'est le cas des gentilés formés avec les suffixes -*anin* ou -*ič* (157a). D'autres sont motivés indirectement par des toponymes et directement par des adjectifs ; dans ce cas, on observe un décalage entre la base formelle et la base sémantique. Les gentilés construits avec le suffixe -*ec* sont notamment concernés (157b).

(157) a. Египет ‘Égypte’ → египтянин ‘Egyptien’
Egipet egiptjanin
Варшава ‘Varsovie’ → варшавянин ‘Varsovien’
Varšava varšavjanin
Москва ‘Moscou’ → москвич ‘Moscovite’
Moskva moskvič

b. Америка ‘Amérique’
Amerika →
американский ‘américain$_A$’ → американец ‘Américain$_N$’
amerikanskij amerikanec
Европа ‘Europe’
Evropa →
европейский ‘européen$_A$’ → европеец ‘Européen$_N$’
evropejskij evropeec

Un deuxième cas de figure concerne les dérivés qui peuvent être reliés sémantiquement à plus d'un nom de base. Dans de telles situations, on parle de motivation multiple (Lopatin, 1977, pp.92-97). Pour illustrer ce cas en français, Fradin (2017) donne l'exemple d'électoral qui est lié sémantiquement à la fois à électeur et à élection[1]. Le même phénomène est assez courant dans la dérivation russe. Les règles de l'analyse dérivationnelle, telles que développées par Vinokur (1959), prévoient une distinction entre des relations formelles et sémantiques plus proches et plus éloignées (directes et indirectes) entre les mots formés sur la même base. Certaines études affirment que les motivations multiples incluent des cas où une des motivations est directe et l'autre est indirecte, par exemple, le mot вредительский (vreditel'skij) est motivé indirectement par вредить (vredit') ‘détruire’ et motivé directement par вредитель (vreditel') ‘destructeur’.

Dans le cas des adjectifs dénominaux russes, les adjectifs avec le suffixe -*n*-, motivés formellement par des noms de personnes, dénotent une relation non pas à cette personne, mais à son activité, à sa profession (158a) ou à ses traits de caractère (158b).

(158) a. инженер ‘ingénieur’ → инженерный
inžener inženernyj

b. лицемер ‘hypocrite’ → лицемерный
licemer licemernyj

Sémantiquement, ces adjectifs peuvent être motivés selon deux noms de base : un nom qui désigne une profession, un type d'activité et un nom qui désigne une science, un courant (159).

[1]L'existence d'une seule forme adjectivale pour deux noms a été analysée comme le résultat d'un principe d'économie par Roché (2008).

(159) a. ВЕТЕРИНАР ‘vétérinaire’ / ВЕТЕРИНАРИЯ ‘médecine vétérinaire’
VETERINAR VETERINARIJA
→ ВЕТЕРИНАРНЫЙ
VETERINARNYJ

b. КОНКУРЕНТ ‘concurrent’ / КОНКУРЕНЦИЯ ‘concurrence’
KONKURENT KONKURENCIJA
→ КОНКУРЕНТНЫЙ
KONKURENTNYJ

Finalement, comme les adjectifs en *-n-*, les adjectifs en *-sk-* peuvent être motivés selon deux noms : un nom qui désigne un humain et un nom qui désigne une profession ou un courant politique ou idéologique (160a) ; un ethnomyne et un toponyme (160b).

(160) a. ТОЛСТОЙ ‘Tolstoï’ / ТОЛСТОВСТВО ‘tolstoïsme’
TOLSTOJ TOLSTOVSTVO
→ ТОЛСТОВСКИЙ
TOLSTOVSKIJ

b. УЗБЕК ‘ouzbek’ / УЗБЕКИСТАН ‘Ouzbékistan’
UZBEK UZBEKISTAN
→ УЗБЕКСКИЙ
UZBEKSKIJ

L’une des solutions pour résoudre le problème de la motivation multiple consiste à conserver tous les candidats pour l’analyse. C’est la stratégie adoptée, par exemple, par Bonami et Thuilier (2019). Dans leur étude sur la concurrence des suffixes verbaux *-iser* et *-ifier*, ils considèrent que ni les propriétés formelles ni les propriétés sémantiques ne permettent d’aboutir à une heuristique satisfaisante pour déterminer la base d’un verbe donné : nom ou adjectif (161).

(161) CENTRE / CENTRAL → CENTRALISER

Ainsi, les deux candidats potentiels sont conservés pour l’étude de la concurrence. Cependant, cela peut entraîner une multiplication des entrées, ainsi qu’une multiplication des annotations des propriétés linguistiques de ces entrées. Pour résoudre ce problème, Bonami et Thuilier (2019) utilisent uniquement le radical des verbes pour les annotations phonologiques et incluent des informations sur les familles morphologiques de ces verbes ainsi que des informations sur les classes d’adjectifs à part.

Une autre solution consisterait à ne conserver qu’un seul lexème de base, ce qui soulève la question du choix d’un candidat. Ainsi, Strnadová (2014, pp.117-119) se base sur la fréquence du patron qui relie le singulier du nom au radical de dérivation et qui dépend de la similarité phonologique. La paire qui correspond au patron le plus fréquent au sein d’une suffixation est alors conservée. Par exemple, le couple

PSYCHOPATHE → PSYCHOPATHIQUE correspond à un patron plus fréquent que le couple PSYCHOPATHIE → PSYCHOPATHIQUE.

Nous adopterons le principe du choix des bases formulé par Lopatin (1977, p.90) : parmi plusieurs lexèmes de la même famille morphologique qui sont morphologiquement plus simples, celui qui est le plus proche formellement du dérivé est considéré comme sa base.

5.1.1.2 Base nominale et base verbale

Jusqu'à présent, notre attention s'est portée sur les situations où les deux bases potentielles sont de la même catégorie lexicale. Cependant, il existe des scénarios où un adjectif peut être formé à partir d'une base nominale ou d'une base verbale. C'est le cas des adjectifs qui finissent par des séquences *-tel'sk(ij)* (162a) et *-tel'n(yj)* (162b).

(162) a. УЧИТЬ 'enseigner' / УЧИТЕЛЬ 'enseignant'
UČIT' UČITEL'
→ УЧИТЕЛЬСКИЙ
UČITEL'SKIJ
ГРАБИТЬ 'voler' / ГРАБИТЕЛЬ 'voleur'
GRABIT' GRABITEL'
→ ГРАБИТЕЛЬСКИЙ
GRABITEL'SKIJ
УГНЕТАТЬ 'opprimer' / УГНЕТАТЕЛЬ 'oppresseur'
UGNETAT' UGNETATEL'
→ УГНЕТАТЕЛЬСКИЙ
UGNETATEL'SKIJ

b. НАБЛЮДАТЬ 'observer' / НАБЛЮДАТЕЛЬ 'observateur'
NABLJUDAT' NABLJUDATEL'
→ НАБЛЮДАТЕЛЬНЫЙ
NABLJUDATEL'NYJ
МСТИТЬ 'se venger' / МСТИТЕЛЬ 'vengeur'
MSTIT' MSTITEL'
→ МСТИТЕЛЬНЫЙ
MSTITEL'NYJ
ХРАНИТЬ 'garder' / ХРАНИТЕЛЬ 'gardien'
XRANIT' XRANITEL'
→ ХРАНИТЕЛЬНЫЙ
XRANITEL'NYJ

Les adjectifs en (162a) et (162b) sont sémantiquement liés à une base verbale. Toutefois, selon Vinogradov (1952, p.177), dans la plupart des cas, ces adjectifs sont directement motivés par des noms qui sont eux-mêmes construits avec le suffixe *-tel'* à

partir d'une base verbale et qui désignent des agents[2]. Cependant, il y a des cas où le nom d'agent est difficilement identifiable (163).

(163) a. НАПЛЕВАТЬ 's'en ficher' / ?НАПЛЕВАТЕЛЬ
NAPLEVAT' NAPLEVATEL'
→ НАПЛЕВАТЕЛЬСКИЙ
NAPLEVATEL'SKIJ

b. СТАРАТЬСЯ 'faire des efforts' / ?СТАРАТЕЛЬ
STARAT'SJA STARATEL'
→ СТАРАТЕЛЬСКИЙ
STARATEL'SKIJ

c. РУГАТЬ 'insulter' / ?РУГАТЕЛЬ
RUGAT' RUGATEL'
→ РУГАТЕЛЬНЫЙ
RUGATEL'NYJ

d. УСПОКОИТЬ 'calmer' / ?УСПОКОИТЕЛЬ
USPOKOIT' USPOKOITEL'
→ УСПОКОИТЕЛЬНЫЙ
USPOKOITEL'NYJ

e. ЧИХАТЬ 'éternuer' / ?ЧИХАТЕЛЬ
ČIXAT' ČIXATEL'
→ ЧИХАТЕЛЬНЫЙ
ČIXATEL'NYJ

Zemskaja (2011, pp.228-230) analyse les noms d'agent en (163) comme des mots potentiels : lorsqu'un locuteur utilise ces mots, il ne s'agit pas des lexèmes connus, mais d'une pure réalisation des possibilités morphologiques. Selon Zemskaja, ces noms d'agent peuvent être formés librement à partir des verbes et apparaissent en fonction des besoins. Uluxanov (2005, p.201) en parlant des adjectifs en *-tel'n(yj)* remarque que ces derniers sont principalement motivés par les verbes ; il note les noms d'action en *-eni(e)* comme les bases secondaires (164).

(164) a. ОСКОРБИТЬ 'insulter' / ОСКОРБЛЕНИЕ 'insulte'
OSKORBIT' OSKORBLENIE
→ ОСКОРБИТЕЛЬНЫЙ
OSKORBITEL'NYJ

[2]Ces noms peuvent être animés : ИСПОЛНИТЬ (ISPOLNIT') 'exécuter' → ИСПОЛНИТЕЛЬ (ISPOLNITEL') 'exécuteur' ; ou inanimés : ВЫКЛЮЧИТЬ (VYKLJUČIT') 'éteindre' → ВЫКЛЮЧАТЕЛЬ (VYKLJUČATEL') 'interrupteur' ; ou les deux : РАСПРЕДЕЛИТЬ (RASPREDELIT') 'distribuer' → РАСПРЕДЕЛИТЕЛЬ (RASPREDELITEL') 'distributeur' (Uluxanov, 1977, pp.95,99).

b. оправдать ‘justifier’ / оправдание ‘justification’
OPRAVDAT’ OPRAVDANIE
→ оправдательный
OPRAVDATEL’NYJ

c. желать ‘souhaiter’ / желание ‘souhait’
ŽELAT’ ŽELANIE
→ желательный
ŽELATEL’NYJ

d. общаться ‘communiquer’ / общение ‘communication’
OBŠČAT’SJA OBŠČENIE
→ общительный
OBŠČITEL’NYJ

Graudina *et al.* (2001, pp. 414-415) proposent une étude approfondie des liens sémantiques entre les deux bases potentielles et l’adjectif dérivé, s’appuyant sur des entrées de dictionnaires. Ainsi cet auteur compte plus de 600 adjectifs dont la sémantique contient un sens verbal, comme, par exemple, гадательный (GADATEL’NYJ) ‘deviner’, доказательный (DOKAZATEL’NYJ) ‘prouver’, желательный (ŽELATEL’NYJ) ‘souhaiter’, мигательный (MIGATEL’NYJ) ‘clignoter’, плавательный (PLAVATEL’NYJ) ‘nager’, страдательный (STRADATEL’NYJ) ‘souffrir’. La tendance à la double motivation n’est constatée que lorsque les noms d’agent en *-tel’* sont attestés, et désignent des personnes ayant des activités socialement saillantes, telles que воспитатель (VOSPITATEL’) ‘éducateur’, завоеватель (ZAVOEVATEL’) ‘conquérant’, зритель (ZRITEL’) ‘spectateur’, избиратель (IZBIRATEL’) ‘électeur’, исполнитель (ISPOLNITEL’) ‘exécuteur’. Le nombre d’adjectifs ayant une base nominale (réelle ou potentielle), selon Graudina *et al.* (2001, pp. 414-415), serait d’environ 4% (30 unités) du nombre total de formations similaires. De manière générale, des doublets dérivationnels sont présents pour ces noms (165).

(165) a. воспитатель ‘éducateur’
VOSPITATEL’
→ воспитательный / воспитательский
VOSPITATEL’NYJ VOSPITATEL’SKIJ

b. завоеватель ‘conquérant’
ZAVOEVATEL’
→ завоевательный / завоевательский
ZAVOEVATEL’NYJ ZAVOEVATEL’SKIJ

c. зритель ‘spectateur’
ZRITEL’
→ зрительный / зрительский
ZRITEL’NYJ ZRITEL’SKIJ

d. ИЗБИРАТЕЛЬ ‘électeur’
IZBIRATEL’
→ ИЗБИРАТЕЛЬНЫЙ / ИЗБИРАТЕЛЬСКИЙ
IZBIRATEL’NYJ IZBIRATEL’SKIJ

e. ИСПОЛНИТЕЛЬ ‘exécuteur’
ISPOLNITEL’
→ ИСПОЛНИТЕЛЬНЫЙ / ИСПОЛНИТЕЛЬСКИЙ
ISPOLNITEL’NYJ ISPOLNITEL’SKIJ

En considérant les caractéristiques dérivationnelles des adjectifs en *-tel’n(yj)* et *-tel’sk(ij)*, telles que leur motivation potentielle par un nom et un verbe, l’absence de nom d’agent attesté pour certains adjectifs, et étant donné que le sujet d’étude se concentre sur les adjectifs dénominaux, les adjectifs en *-tel’n(yj)* et *-tel’sk(ij)* seront exclus de l’analyse.

Le dernier cas de figure avec une double motivation – nominale et verbale – concerne les adjectifs finissant par la séquence *-očn(yj)* (166).

(166) ПОЛИРОВАТЬ ‘polir’ / ПОЛИРОВКА ‘polissage’
POLIROVAT’ POLIROVKA
→ ПОЛИРОВОЧНЫЙ
POLIROVOČNYJ

Lopatin (1977, p.89) remarque que, si les deux bases potentielles sont possibles, les deux sont sémantiquement liées à la même action (puisqu’il s’agit d’un verbe et d’un nom d’action). Toutefois, du point de vue formel, c’est le nom qui est plus étroitement lié à l’adjectif dérivé, comme c’est le cas pour les exemples en *-tel’sk(ij)* et *-tel’n(yj)* discutés précédemment. Cependant, contrairement à *-tel’sk(ij)* et *-tel’n(yj)*, pour lesquels certains noms de base restent des noms potentiels, tous les noms dérivés de *-očn(yj)* n’ont pas de base verbale, selon Švedova (1980, p.194) (167)[3].

(167) *ЖЕРЕБЬЕВАТЬ / ЖЕРЕБЬЁВКА ‘tirage au sort’
ŽEREB’EVAT’ ŽEREB’ËVKA
→ ЖЕРЕБЬЁВОЧНЫЙ
ŽEREB’ËVOČNYJ

De ce fait, nous allons retenir les adjectifs finissant par la séquence *-očn(yj)* dans notre étude des adjectifs dénominaux et nous allons considérer les noms en *-ovk(a)* comme leurs bases.

[3]Švedova (1980, p.194) précise toutefois que la majorité des noms d’action sont dérivés des verbes.

5.1.2 Homonymie, polysémie et variation

5.1.2.1 Homonymie et polysémie

Jusqu'à présent, nous avons associé les adjectifs à un seul nom de base. Cependant, un grand nombre de noms russes sont polysémiques (168a) ; des homonymes sont également présents (168b-168c).

(168) a. ЯЗЫК$_1$ 'langue (organe)'
JAZYK
ЯЗЫК$_2$ 'langue (linguistique)'
JAZYK

b. МИР$_1$ 'monde'
MIR
МИР$_2$ 'paix'
MIR

c. РАК$_1$ 'écrevisse'
RAK
РАК$_2$ 'cancer'
RAK

Lors de la formation des mots à partir de bases polysémiques, les relations de dérivation sont établies en fonction du sens spécifique activé (Fradin, 2014 ; Fradin, 2016). Ainsi, chaque sens donne naissance à une nouvelle série de dérivés (Arutjunova, 1961, pp.36-40 ; Uluxanov, 1977, pp.7-9). Afin de poursuivre les analyses et la modélisation, il est nécessaire de résoudre les ambiguïtés et la polysémie des lexèmes isolés.

Traditionnellement, les homonymes sont considérés par de nombreux chercheurs comme des lexèmes présentant une identité formelle (phonétique ou graphique), mais ayant des significations distinctes (Reformatskij, 1967, p.48). De plus, on considère deux lexèmes comme homonymes lorsqu'aucun procédé de dérivation sémantique, métaphore ou métonymie, n'entre en jeu. En revanche, la polysémie est une propriété linguistique qui permet d'exprimer de manière économique un grand nombre de significations à l'aide d'un nombre limité d'unités. Il est généralement admis que ces différentes significations sont liées (Vinogradov, 1977, pp.288–294) et réalisées dans leur contexte (Raxilina *et al.*, 2006 ; Toldova *et al.*, 2008 ; Šemanaeva *et al.*, 2007)[4].

En ce qui concerne les sources lexicographiques, il est courant de distinguer l'homonymie et la polysémie au niveau des entrées. Les homonymes se voient attribuer des entrées distinctes dans le dictionnaire, tandis que la polysémie est représentée dans une seule entrée, où les différents sens du mot sont énumérés.

[4]Cependant, il n'existe pas de consensus sur la définition de la polysémie, certains chercheurs remettant en question l'existence même de ce phénomène au profit de l'homonymie. Comme le soutient Ščerba (1958), il y a toujours autant de mots distincts que le lexème donné a de significations dans différents contextes.

Les adjectifs peu fréquents, notamment les hapax, présentent un avantage dans la mesure où ils permettent d'identifier facilement le sens du nom de base, car leur usage est restreint à un nombre très limité de contextes qui peuvent être vérifiés manuellement (169).

(169) a. КЛАДЕНЕЦ$_1$ 'Kladenec (épée ; artefact)' → КЛАДЕНЕЧНЫЙ
KLADENEC KLADENEČNYJ
КЛАДЕНЕЦ$_2$ 'Kladenec (esprit ; idéalité)' ↛ КЛАДЕНЕЧНЫЙ
KLADENEC KLADENEČNYJ
Пусть ***кладенечные*** *изломы Врагов, как молния, разят*
Pust' ***kladenečnye*** *izlomy Vragov, kak molnija, razjat*
'Que les ondulations **de l'épée** Foudroient les ennemis' (N.A. Kljuev. "Je suis un dévoué du peuple...", 1918 ; RusCorpora poétique)

b. БРЮНН 'Brünn, ancien nom de Brno (toponyme)' → БРЮННСКИЙ
BRJUNN BRJUNNSKIJ
БРЮННЫ 'Brünn (race)' ↛ БРЮННСКИЙ
BRJUNNY BRJUNNSKIJ
Путешествие начинается от стен ***Брюннского*** *монастыря*
Putešestvie načinaetsja ot sten ***Brjunnskogo*** *monastyrja*
'Le voyage commence aux murs du monastère **de Brünn**' (I. Guberman. "Le vrai voyage" // "Chimie et vie", 1968 ; RusCorpora général)

En ce qui concerne les adjectifs de haute fréquence, la situation est cependant plus complexe : en raison du grand nombre de contextes dans lesquels ils apparaissent, il n'est pas possible d'effectuer une vérification manuelle pour déterminer tous les sens possibles du nom correspondant. Une alternative est d'utiliser des données lexicographiques pour identifier les différents sens du nom de base. Cependant, certaines entrées peuvent inclure une dizaine de sens pour un seul mot, qui peuvent être plus ou moins similaires. Si nous cherchons à résoudre les ambiguïtés des noms polysémiques, il n'est pas possible d'adopter une granularité sémantique très fine sans augmenter considérablement la taille de la base de données par la multiplication des entrées.

Différentes approches sont proposées pour analyser les nombreux sens des mots polysémiques. Apresjan (1974, p.116) propose de réduire graduellement les significations lexicales complexes à des significations plus simples, en identifiant le composant sémantique commun à ces significations. Flaux et Van de Velde (2000, pp.8-10), à leur tour, considèrent que chaque nom a une classe d'appartenance initiale, qui peut changer sous certaines conditions. Le critère pour déterminer le sens initial d'un mot est basé sur la possibilité de déduire d'autres sens à partir de celui-ci, en utilisant des règles présentant une généralité suffisante.

D'autres approches, comme celle présentée par Haas *et al.* (2022) et dont nous nous inspirons pour la présente étude, consistent à classer les mots en fonction de classes générales plutôt que de distinctions de sens très fines ou la recherche d'un sens initial. Cette méthode intermédiaire permet de classifier les différentes significations d'un nom

en fonction des classes sémantiques suivantes : propre/commun, animé/non animé, concret/abstrait[5]. Par conséquent, lorsqu'un mot peut faire référence à la fois à des objets concrets et à des entités abstraites (170a-170b), ou encore à la fois à des entités inanimées et à des êtres animés (170c), il est doté de deux étiquettes sémantiques. Inversement, si les sens du mots polysémique ne varient pas selon les classes sémantiques mentionnées plus haut, une seule étiquette est attribuée (170d-170f).

(170) a. КВАРТАЛ 'quartier ; trimestre' : concret, abstrait
KVARTAL

b. СВОД 'voûte ; compilation' : concret, abstrait
SVOD

c. МОДЕЛЬ 'modèle ; mannequin' : humain, concret
MODEL'

d. ЛИСТ 'feuille d'arbre ; feuille de papier' : concret
LIST

e. ИКРА 'caviar ; mollet' : concret
IKRA

f. БЛЮДО 'plat ; plateau' : concret
BLJUDO

Au total, 58 noms de base sont dotés d'une double annotation sémantique.

5.1.2.2 Norme orthographique

La dernière considération lors de l'attribution d'une base à un adjectif est la conformité à la norme orthographique. Certains adjectifs dérivés de bases nominales d'origine étrangère ne sont pas encore complètement assimilés par la langue russe, ce qui se manifeste par des variations graphiques. Nous avons vérifié l'existence des noms de base dans des dictionnaires et sur Wikipedia. Si les deux variantes orthographiques sont référencées, les deux noms de base sont retenus, avec les adjectifs correspondants orthographiés différemment en conséquence (171a). En revanche, si une seule variante orthographique est répertoriée dans les sources lexicographiques, le deuxième est considéré comme une faute d'orthographe (171b).

(171) a. РИЭЛТЕР ∼ РИЭЛТОР 'agent immobilier'
RIÈLTER RIÈLTOR
ОФШОР ∼ ОФФШОР 'paradis fiscal'
OFŠOR OFFŠOR

b. ДЕМПИНГ ∼ *ДЭМПИНГ 'dumping'
DEMPING DÈMPING
ФЕЙК ∼ *ФЭЙК 'fake'
FEJK FÈJK

[5]Les classes sémantiques seront traitées plus en détail dans la section 5.2.4.

Finalement, étant donné que RusCorpora couvre la période allant de 1850 jusqu'à nos jours, il se peut que des toponymes aient changé de nom depuis leur utilisation antérieure (172). Dans ces cas, les variantes ont été conservées.

(172) Ущельное 'Uščel'noe (Ukraine)', avant 1945 – Дерекой
Uščel'noe Derekoj
Избаскан / Избоскан 'Izboskan (Ouzbékistan)', auparavant –
Izbaskan Izboskan
Избаскент
Izbaskent
Марьют / Мариут 'Mariout (Égypte)', auparavant – Марея
Mar'jut Mariut Mareja

5.2 Propriétés des noms de base

Une fois que nous avons identifié les noms de base correspondant à chaque adjectif, nous pouvons procéder à l'analyse des différentes propriétés linguistiques de ces noms de base. Nous pourrons alors commencer à déterminer s'il existe des préférences suffixales et si certaines propriétés peuvent être utilisées pour prédire le comportement dérivationnel des noms. Les propriétés linguistiques seront annotées sur plusieurs niveaux, à savoir le plan phonologique, morphologique, sémantique et étymologique. Le récapitulatif des propriétés des noms de base, y compris les informations sur les fréquences, est présenté dans la Structure de données (Annexe A1). Dans cette partie, nous utiliserons des statistiques descriptives. Cependant, étant donné que les proportions ne sont pas équilibrées (par exemple, il y a deux fois plus d'adjectifs avec le suffixe -*n*- que d'adjectifs avec les suffixes -*n*- et -*Ov*- dans les données de haute fréquence), nous introduirons des tests statistiques pour évaluer les relations entre les propriétés des noms de base et les suffixes. Pour déterminer si les associations entre eux sont statistiquement significatives, nous utiliserons le test du χ^2 (khi carré). Pour quantifier la force de ces associations, nous utiliserons le coefficient ϕ de Cramer (Cramer's V). Finalement, pour identifier les propriétés individuelles qui contribuent le plus à la statistique du χ^2, nous analyserons les résidus.

5.2.1 Mesures statistiques

Pour présenter les mesures statistiques utilisées dans cette section, nous nous baserons sur des données fictives relatives à la distribution de certaines propriétés des noms de base (`Prop1` et `Prop2`) ainsi que de certains suffixes (`Suff1`, `Suff2`, `Suff3`), présentées dans le tableau 5.1.

Le test du χ^2 d'indépendance (ou test du khi carré) est un test statistique qui permet de vérifier s'il existe une relation statistiquement significative entre deux variables catégorielles. Il se base sur la différence entre les valeurs observées dans le tableau de contingence et les valeurs attendues ; plus précisément sur la somme des

Prop	Suff1	Suff2	Suff3	Total
Prop1	40	65	55	160
Prop2	70	50	60	180
Total	110	115	115	340

Tableau 5.1: Distribution de données fictives

écarts carrés entre les valeurs observées et les valeurs attendues, divisés par les valeurs attendues.

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(O_{ij} - A_{ij})^2}{A_{ij}}$$

où :

χ = la valeur khi carré
l = nombre de lignes,
c = nombre de colonnes,
O_{ij} = la fréquence observée pour la cellule (i,j)
A_{ij} = la fréquence attendue pour la cellule (i,j).

Le test d'indépendance repose sur l'hypothèse nulle selon laquelle il n'existe aucune association entre les variables, c'est-à-dire entre les lignes l et les colonnes c dans le tableau croisé. Les fréquences attendues correspondent aux fréquences qui seraient observées si les variables étaient indépendantes, c'est-à-dire si l'hypothèse nulle était vraie et s'il n'y avait aucune différence dans la distribution des propriétés pour chaque suffixe (Levshina, 2015, pp.210-215 ; Gries, 2016, pp.189-200). Les fréquences attendues sont calculés de la manière suivante :

$$A_{ij} = \frac{T_i T_j}{N}$$

où :

A_{ij} = la fréquence attendue pour la cellule (i,j),
T_i = le total des observations de la ligne i,
T_j = le total des observations de la colonne j,
N = le total de toutes les observations dans le tableau.

Il est à souligner que pour pouvoir appliquer le test du χ^2, toutes les fréquences attendues doivent être supérieures à 5[6]. Les fréquences attendues pour les données fictives sont exposées dans le tableau 5.2.

Après avoir calculé les fréquences attendues, nous pouvons calculer le χ^2 (tableau 5.3).

[6]Dans le cas où certaines fréquences attendues sont inférieures à 5, il est possible de recourir au test exact de Fisher pour évaluer la significativité des liens entre les variables.

Prop	Suff1	Suff2	Suff3	Total
Prop1	51.76	54.12	54.12	160
Prop2	58.24	60.88	60.88	180
Total	110	115	115	340

Tableau 5.2: Distribution des valeurs attendues pour les données fictives

Prop	Suff1	Suff2	Suff3	Total
Prop1	2.67	2.19	0.02	
Prop2	2.37	1.94	0.01	
Total	5.04	4.13	0.03	$\chi =$ **9.20**

Tableau 5.3: χ^2 pour les données fictives

En plus de la valeur de χ^2, le résultat du test dépend également de la taille du tableau. Plus le nombre de cellules est important, plus la probabilité de trouver des divergences augmente. Cette considération est prise en compte dans la notion de degrés de liberté, qui correspond au nombre de valeurs pouvant varier librement. Pour calculer le nombre de degrés de liberté d'un tableau de contingence, il est nécessaire de soustraire 1 du nombre de lignes et du nombre de colonnes, puis de multiplier entre eux les résultats obtenus :

$$dl = (l-1)(c-1)$$

Par conséquent, les degrés de liberté pour les données fictives sont : (2-1)*(3-1)=2. En suivant la convention de Levshina (2015, p.222) dans la notation des degrés de liberté entre parenthèses, nous présenterons systématiquement les résultats de χ^2 la manière suivante : $\chi^2(2) = 9.20$. $p<0.01$.

Finalement, la valeur p correspond à la probabilité d'obtenir la statistique de test et des résultats encore plus extrêmes par pur hasard, en supposant que l'hypothèse nulle est vraie. La valeur critique de p est généralement fixée à 0.05. La valeur p que nous avons obtenue est inférieure à cette valeur critique, nous pouvons donc rejeter l'hypothèse nulle et conclure qu'il existe un lien statistiquement significatif entre les propriétés et les suffixes dans les données fictives[7].

Cependant, le test de χ^2 ne suffit pas à décrire complètement les données. Avec l'augmentation de la taille de l'échantillon (comme c'est le cas pour nos données réelles), le hasard diminue, ce qui facilite la détection des différences entre les groupes (c'est-à-dire que la puissance statistique augmente). Ainsi, pour les grandes tailles d'échantillon, même de petits effets peuvent devenir significatifs. Nous avons déterminé que le lien entre les deux variables dans les données fictives est significatif, mais la valeur de χ^2 ne permet pas d'évaluer l'intensité de ce lien. Pour la quantifier, et

[7]Nous allons effectuer les calculs statistiques à l'aide de la librairie `SciPy` en python (Virtanen *et al.*, 2020).

notamment lorsqu'il s'agit des tableaux de plus de deux lignes et de plus de deux colonnes, comme le tableau 5.1 (Gries, 2013, p.185) nous allons utiliser le coefficient ϕ, également appelé V de Cramer :

$$\phi = \sqrt{\frac{\chi^2}{N(min(l,c)-1)}}$$

χ^2 = le chi carré,
N = le total de toutes les fréquences observées dans le tableau,
min(l,c) = le minimum entre le nombre de lignes et le nombre de colonnes.

La mesure ϕ varie entre 0 et 1, où les valeurs proches de 0 indiquent une association très faible, tandis que les valeurs proches de 1 indiquent une association très forte.

Pour les données fictives, le coefficient ϕ est ainsi de 0.16. Cette mesure indique que l'association entre les propriétés et les suffixes fictifs est faible mais significative, comme l'a montré le test de χ^2.

En résumé, le test du khi carré et le coefficient V de Cramer peuvent être utilisés pour établir s'il existe une association significative entre deux variables catégorielles et pour quantifier cette association. Cependant, ils ne permettent pas d'identifier les cases spécifiques du tableau croisé (dans notre exemple, 6 cases) qui contribuent le plus à l'association entre les variables. Pour répondre à cette question, il est nécessaire d'analyser les résidus.

Les résidus correspondent à la différence entre les fréquences observées et les fréquences attendues, divisée par la racine carrée des fréquences attendues :

$$r = \frac{(O_{ij} - A_{ij})}{\sqrt{A_{ij}}}$$

où :

O_{ij} = la fréquence observée pour la cellule (i,j)
A_{ij} = la fréquence attendue pour la cellule (i,j).

Les résidus peuvent prendre des valeurs positives ou négatives. Plus ces valeurs s'éloignent de zéro, plus l'effet est important. Les valeurs des résidus pour les données fictives sont présentées dans le tableau 5.4.

Prop	Suff1	Suff2	Suff3
Prop1	-1.64	1.48	1.12
Prop2	1.54	-1.39	-0.11

Tableau 5.4: Les résidus pour les données fictives

Les résultats indiquent que la non-préférence du `Suff1` pour la `Prop1` (-1.64) contribue le plus à l'association entre les variables, suivie de près par la préférence du même suffixe pour la `Prop2` (1.54).

Dans la section suivante, nous exposerons les résultats issus de l'application de ces analyses statistiques aux noms de base des adjectifs de haute et basse fréquence.

5.2.2 Phonologie

5.2.2.1 Structure syllabique

Pour calculer le nombre de syllabes dans le nom de base, nous allons utiliser la forme qui correspond au nominatif singulier (et donc la forme de citation des noms)[8].

Dans la section 1.2, nous avons noté que, contrairement aux consonnes, les phonèmes vocaliques en russe peuvent toujours former le noyau d'une syllabe. Pour automatiser le calcul des syllabes dans un mot donné, on peut ainsi se baser sur le nombre de voyelles qu'il contient. Les résultats sont présentés dans le tableau 5.5.

	H Freq			B Freq		
SyllN	-*Ov*-	-*n*-	-*sk*-	-*Ov*-	-*n*-	-*sk*-
1	212	61	48	50	12	21
2	222	559	295	98	151	235
3	85	390	274	75	185	198
4	10	132	116	21	79	65
5+	0	50	40	5	12	30

Tableau 5.5: Distribution du nombre de syllabes ; haute et basse fréquence

On observe que la majorité des adjectifs sont construits à partir de noms comportant deux ou trois syllabes. On constate aussi que si les noms monosyllabiques favorisent le suffixe -*Ov*-, les noms de deux et trois syllabes ont une préférence pour -*n*- et -*sk*-.

La corrélation entre la structure syllabique et le suffixe adjectival est hautement significative sur le plan statistique (`H Freq` : $\chi^2(8) = 518.65$, $p<0.001$; `B Freq` : $\chi^2(8) = 114.33$, $p<0.001$). L'association est forte dans les données de haute fréquence ($\phi = 0.32$) ; dans les données de basse fréquence elle est plutôt faible ($\phi = 0.21$).

La distribution des résidus est présentée sur la figure 5.1.

Dans les données de haute et de basse fréquence, la préférence des noms monosyllabiques pour le suffixe -*Ov*- est le facteur qui contribue le plus à la statistique du χ^2, bien que l'effet soit plus faible dans les données de basse fréquence. La non-préférence des noms monosyllabiques pour les suffixes -*n*- et -*sk*-, ainsi que la non-préférence des noms comportant trois syllabes ou plus pour le suffixe -*Ov*- contribuent également à la statistique, mais dans la moindre mesure.

[8]Dans la section 1.3, nous avons examiné les différentes allomorphies thématiques. Bien que la forme du nominatif singulier ne prenne pas en compte ces variations, nous l'utiliserons pour l'annotation phonologique, notamment pour déterminer la longueur du nom de base en syllabes et la position de l'accent tonique (elle sera examinée plus bas dans cette section). Les informations morphologiques et morphophonologiques, quand à elles, seront traitées séparément dans la section 5.2.3.

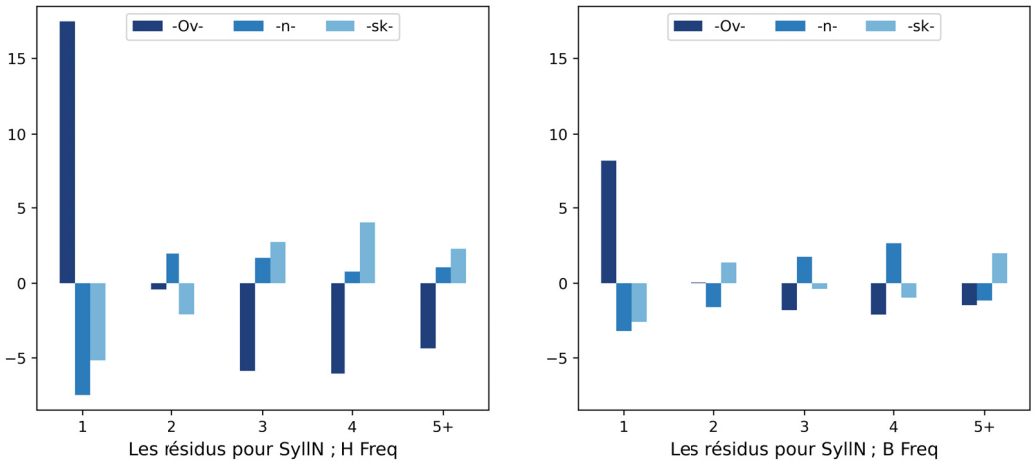

Figure 5.1: Résidus pour le nombre de syllabes ; haute et basse fréquence

5.2.2.2 Position de l'accent

Nous allons adopter deux conventions pour annoter la position de l'accent. Tout d'abord, nous utiliserons la forme de citation des noms de base (NOM.SG.) pour déterminer quelle syllabe porte l'accent tonique, à partir de la fin du mot. Deuxièmement, nous tiendrons compte de tout le paradigme nominal et annoterons le type accentuel selon la typologie de Zaliznjak (2003) (cf. la section 1.2.3).

Comme nous l'avons souligné dans la section 1.2.3, la plupart des noms russes ont un seul accent tonique. Les mots avec deux accents sont moins fréquents et sont généralement des mots composés récemment apparus en russe et dont la première partie correspond à un nom abrégé (173a). Les composés plus anciens sont caractérisés par un seul accent tonique (173b). Il convient de rappeler que, selon Švedova (1980, pp.90-92), en cas de double accentuation, l'accent primaire est généralement positionné plutôt vers la fin du mot, tandis que l'accent secondaire est placé vers le début (cf. la discussion dans la section 1.2.3). Nous allons suivre ces instructions pour l'annotation des noms de base concernés.

(173) a. СОЦРЕАЛИЗМ /ˌsot͡srʲeaˈlʲizm/ /socr'eal'ízm/
SOCREALIZM
< *социальный реализм* 'réalisme socialiste'
social'nyj realizm
ПАРТБИЛЕТ /ˌpartbʲiˈlʲet/ /partb'il'ét/
PARTBILET
< *партийный билет* 'carte du parti'
partijnyj bilet

b. ВОДОПАД /vodoˈpad/ /vodopád/ 'chute d'eau'
VODOPAD

СТИХОТВОРЕНИЕ /stʲixotvo'rʲenʲije/ /st'ixotvor'én'ije/ ‘poésie’
STIXOTVORENIE

Le tableau 5.6 présente la répartition de la position de l’accent tonique selon les suffixes utilisés.

	H Freq			B Freq		
AccSyllN	*-Ov-*	*-n-*	*-sk-*	*-Ov-*	*-n-*	*-sk-*
aad+	9	128	115	14	42	78
ad	138	525	258	87	219	203
d	382	539	400	148	178	268

Tableau 5.6: Distribution de la position de l’accent ; haute et basse fréquence

La plupart des noms employés pour la formation des adjectifs présentent une accentuation soit sur l’avant-dernière syllabe, soit sur la dernière syllabe. Les données de haute et de basse fréquence montrent une tendance pour le suffixe *-Ov-*, dont l’utilisation diminue lorsque l’accent est éloigné de la fin des noms.

Tout comme pour la structure syllabique, la corrélation entre la position de l’accent tonique et le choix des suffixes est significative (`H Freq` : $\chi^2(4) = 141.41$, $p<0.001$; `B Freq` : $\chi^2(4) = 37.40$, $p<0.001$), mais faible dans les deux corpus (`H Freq` : $\phi = 0.17$; `B Freq` : $\phi = 0.12$).

La figure 5.2 présente la distribution des résidus.

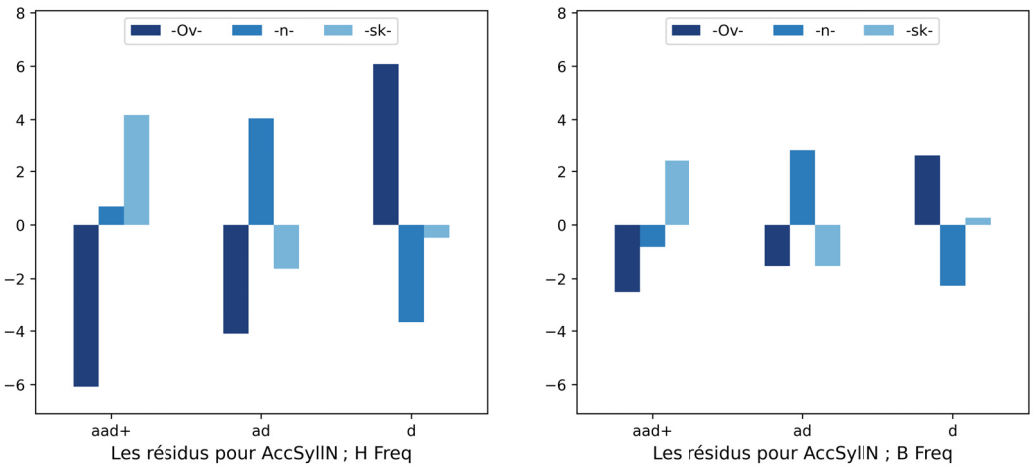

Figure 5.2: Résidus pour la position de l’accent ; haute et basse fréquence

Les résidus montrent que les effets sont les mêmes pour les données de haute et de basse fréquence (ces effets sont cependant moins forts dans les données de basse fréquence). La préférence des noms dont la dernière syllabe est accentuée pour le suffixe *-Ov-* et la non-préférence de ce même suffixe pour tous les autres noms contribuent

le plus à la corrélation. La préférence des noms dont l'accent est sur l'avant-dernière syllabe pour le *-n-* et des noms avec l'accent sur l'antépénultième syllabe pour le suffixe *-sk-* contribue aux statistiques, quant à elle, de manière moins importante.

Contrairement à la structure syllabique, les résidus relatifs à la position de l'accent mettent en évidence une nette distinction entre les trois suffixes, en fonction de leurs préférences respectives pour les trois types d'accentuation.

En ce qui concerne la typologie de Zaliznjak (2003), pour rappel, il existe six types accentuels (de `a` à `f`) définis en fonction des modèles de position de l'accent dans l'ensemble des paradigmes nominaux (dans les différents cas et nombres), prenant en compte la position de l'accent sur le thème ou sur le suffixe de flexion. Le type accentuel `0` concerne les noms indéclinables (cf. la discussion et les exemples dans la section 1.2.3). Étant donné le faible nombre de noms dans les classes de `c` à `f`, ainsi que leurs distributions similaires, ces données ont été regroupées. Le tableau 5.7 synthétise les résultats obtenus.

	H Freq			B Freq		
AccZal	*-Ov-*	*-n-*	*-sk-*	*-Ov-*	*-n-*	*-sk-*
0	1	6	16	4	13	39
a	358	1046	717	208	399	501
b	117	59	29	32	23	8
c-f	53	81	11	5	4	1

Tableau 5.7: Distribution des types accentuels ; haute et basse fréquence

Les noms appartenant au type `a` (avec un accent sur le thème dans tous les mots formes) sont les plus nombreux à former des adjectifs dans les deux corpus. Le suffixe *-Ov-* semble se combiner majoritairement avec les noms du type `b` (avec un accent qui est toujours sur le suffixe flexionnel, sauf si celui-ci est absent), tandis que le suffixe *-sk-* est dominant avec les noms indéclinables[9].

La corrélation entre les types accentuels et les suffixes est de nouveau significative (`H Freq`: $\chi^2(6) = 243.87$, p<0.0001 ; `B Freq` : $\chi^2(6)$: 67.24, p<0.0001), mais, comme pour la position de l'accent, cette corrélation est faible (`H Freq` : $\phi = 0.22$; `B Freq` : $\phi = 0.16$).

Les résidus présentés sur la figure 5.3 montrent la contribution de chaque type accentuel à la corrélation.

Dans les données de haute fréquence, la préférence des noms de type `b` pour le suffixe *-Ov-* a l'effet le plus fort, l'effet des autres variables (la défavorisation des noms des types `b-f` pour *-sk-* et la non-préférence des noms de type `a` pour *-Ov-*) est beaucoup plus faible. Dans les données de basse fréquence, les effets sont moins marqués, et c'est

[9]Il est à noter que ces noms indéclinables correspondent en majorité aux noms propres toponymes, par exemple, Амальфи (Amal'fi) 'Amalfi (Italie)', Киото (Kioto) 'Kyoto (Japon)', Колорадо (Kolorado) 'Colorado (États-Unis)', Сомали (Somali) 'Somalie', Царицино (Caricyno) 'Tsaritsyno (Russie)', Таити (Taiti) 'Tahiti (Polynésie française)'

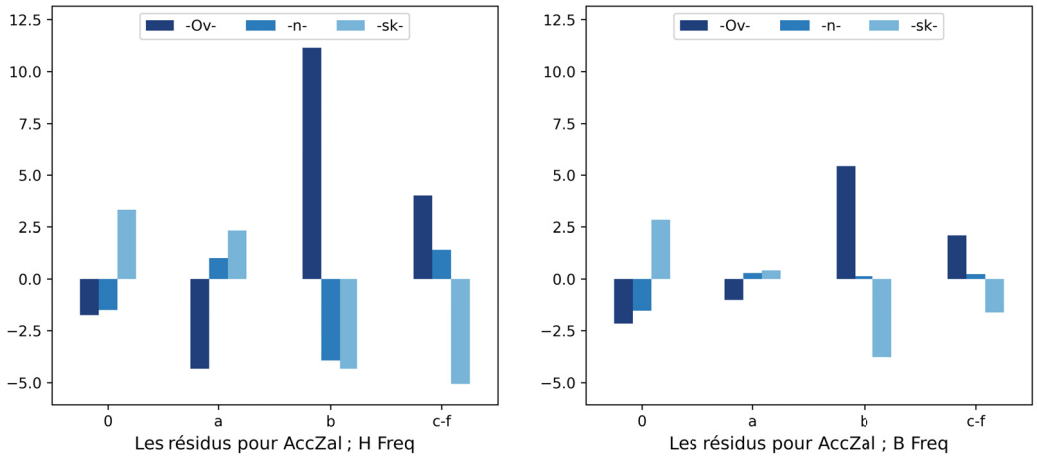

Figure 5.3: Résidus pour les types accentuels ; haute et basse fréquence

surtout la préférence des noms des types `b-f` pour -*Ov*- qui contribuent à la corrélation.

5.2.2.3 Derniers phonèmes des radicaux

L'annotation du dernier phonème ne peut plus reposer sur la forme de citation ou les mots formes. Il est donc nécessaire de recourir aux thèmes pour effectuer cette tâche. Dans la section 1.4.1, nous avons étudié les espaces thématiques, en présentant les thèmes et les radicaux et en précisant qu'il peut y avoir plusieurs thèmes pour un seul lexème. Il se pose alors la question de savoir quel thème choisir pour l'annotation du dernier phonème. La présence potentielle de plusieurs thèmes complique la tâche. De plus, il n'existe pas de hiérarchie entre les thèmes, ce qui rend difficile de privilégier le thème qui correspondrait au mieux à la forme de citation (nominatif singulier). Nous opterons ainsi pour l'annotation des propriétés du dernier phonème des radicaux des adjectifs, puisque chaque adjectif a un radical unique, la question méthodologique du choix ne se pose plus. Ce choix, cependant, ne tient pas compte de différentes allomorphies thématiques, discutées dans les sections 1.4.3 et 1.4.2. Ces allomorphies éventuelles feront objet d'annotations à part, que nous décrirons dans la partie 5.2.3.

Dans la section 1.2.1, nous avons adopté le point de vue de l'école phonologique de Moscou, qui identifie 5 phonèmes vocaliques et 32 phonèmes consonantiques en russe. Cependant, l'analyse de la distribution de 37 variables dans les données de haute et basse fréquence peut être fastidieuse et chronophage. Nous regrouperons alors les consonnes en fonction de leur point d'articulation et les voyelles comme classe à part (Garde, 1998, p.17) :

- Consonnes labiales :
  - labiales occlusives (bilabiales) /p, p', b, b', m, m'/,

  - labiales fricatives (labio-dentaires) /f, f', v, v'/ ;

- Consonnes dentales /t, t', d, d', c, s, s', z, z', n, n', r, r', l, l'/ ;
- Consonnes palatales /č, š, ž, j/ ;
- Consonnes vélaires /k, g, x/ ;
- Voyelles /a, e, o, u, i/.

Le tableau 5.8 présente certaines tendances concernant le dernier phonème des radicaux. Tout d'abord, il n'y a qu'un seul radical se terminant par une voyelle (`Vow`) dans les données de haute fréquence (il s'agit de la dérivation Бордо 'Bordeaux' → бордоский), contrairement aux données de basse fréquence où ce type de radicaux est plus fréquent avec les suffixes -*sk*- et -*Ov*-. De plus, on observe que ce sont les thèmes finissant par les dentales qui sont les plus nombreux parmi les noms qui dérivent les adjectifs dénominaux.

	H Freq			B Freq		
DPhoRad	-*Ov*-	-*n*-	-*sk*-	-*Ov*-	-*n*-	-*sk*-
Vow	0	0	1	3	0	25
cAlv	51	331	107	23	159	38
cDent	289	752	527	122	237	361
cLab	34	106	67	17	40	41
cVel	155	3	71	84	3	84

Tableau 5.8: Distribution du dernier phonème des radicaux ; haute et basse fréquence

Les associations entre les derniers phonèmes et les suffixes sont significatives dans les deux ensembles de données (`H Freq` : $\chi^2(8) = 431.39$, $p<0.001$, `B Freq` : $\chi^2(8) = 293.19$, $p<0.001$), cette association est plutôt forte dans les données de basse fréquence (`H Freq` : $\phi = 0.29$; `B Freq` : $\phi = 0.34$).

L'analyse des résidus (figure 5.4) révèle que ce sont principalement les consonnes vélaires et, dans une moindre mesure, les consonnes alvéolaires qui contribuent à cette association.

Si les vélaires ont une nette préférence pour -*Ov*-, les alvéolaires favorisent le suffixe -*n*-. De manières générale, les vélaires et les alvéolaires permettent de distinguer -*n*- et -*Ov*-, les autres consonnes n'ont pas de préférences très marquées.

5.2.3 Morphologie

5.2.3.1 Genres

Les noms russes se répartissent en trois genres : masculin, féminin et neutre. Toutefois, comme mentionné dans la section 1.3.1, il existe également des noms dits

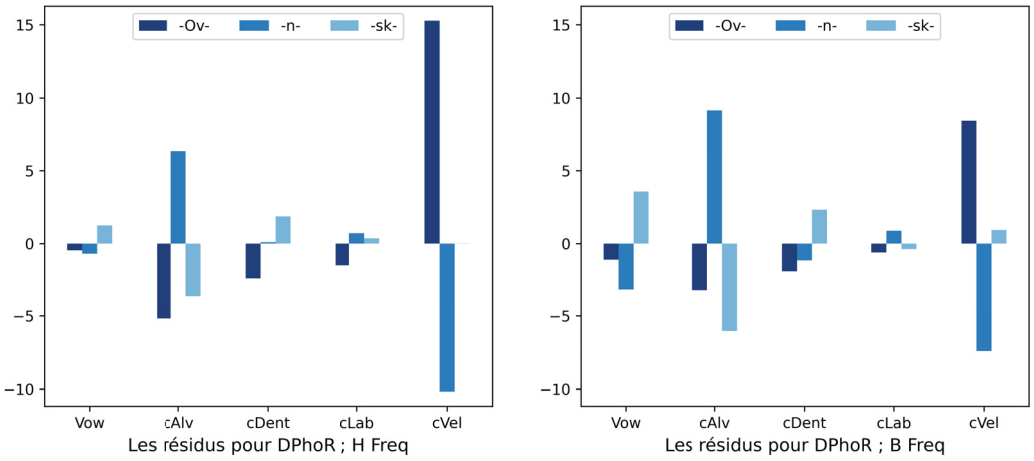

Figure 5.4: Résidus pour les derniers phonèmes des radicaux ; haute et basse fréquence

du genre commun, qui peuvent être masculins ou féminins en fonction de leur référent. Bien que ces noms soient peu nombreux dans le corpus, pour des raisons de commodité dans les analyses statistiques, ils ont été regroupés avec les noms féminins (en raison de leur similitude de distribution).

Le tableau 5.9 montre la distribution des genres et des suffixes. Les noms masculins et féminins sont les plus nombreux à former les adjectifs dénominaux.

	H Freq			B Freq		
Genre	*-Ov-*	*-n-*	*-sk-*	*-Ov-*	*-n-*	*-sk-*
f	83	502	176	62	188	63
m	433	613	579	175	232	473
n	13	77	18	12	19	13

Tableau 5.9: Distribution des genres ; haute et basse fréquence

La corrélation est à nouveau significative (`H Freq` : $\chi^2(4) = 198.17$, p<0.001 ; `B Freq` : $\chi^2(4) = 136.87$, p<0.001), mais faible (`H Freq` : $\phi = 0.20$; `B Freq` : $\phi = 0.24$).

L'examen des résidus permet d'approfondir l'analyse : en effet, la préférence des noms féminins pour le suffixe *-n-* contribue le plus à la corrélation. Les données de la figure 5.5 montrent que la défavorisation des noms féminins pour suffixe *-Ov-* et des noms masculins pour le suffixe *-n-* est également importante dans le corpus de haute fréquence. Dans les données de basse fréquence, c'est la préférence des noms féminins pour *-n-* et la non-préférence de ces mêmes noms pour *-sk-* qui contribuent le plus à la corrélation.

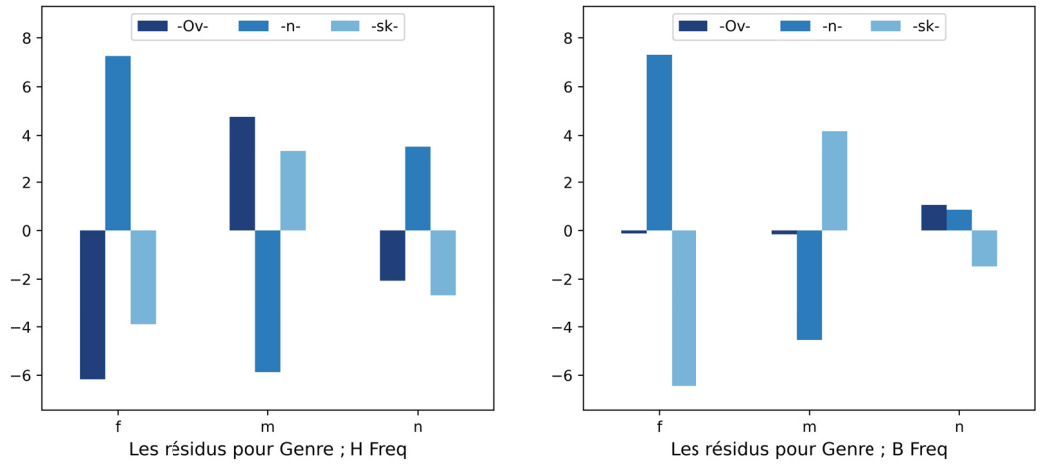

Figure 5.5: Résidus pour le genre ; haute et basse fréquence

5.2.3.2 Classes flexionnelles

Dans cette partie, nous allons examiner les classes flexionnelles en fonction de deux classifications : la classification canonique qui distingue trois classes de flexion, principalement en fonction du genre des noms, et la classification proposée par Zaliznjak (2003, pp.25-76) qui distingue neuf classes en se basant sur les dernières lettres du thème nominal (et non sur les phonèmes). Afin d'assurer une distinction claire, nous adopterons les chiffres romains `I`, `II` et `III` pour représenter les classes flexionnelles canoniques, tandis que nous nous servirons des chiffres arabes de `0` à `9` pour désigner les classes flexionnelles selon la typologie de Zaliznjak.

Bien que la première classification soit basée sur la distinction des noms par genre, les classes flexionnelles `I`, `II` et `III` ne correspondent pas exactement aux genres masculin, féminin et neutre que nous avions annotés plus haut dans cette section : pour rappel, la classe flexionnelle `I` regroupe les noms masculins dépourvus de flexion au nominatif singulier et les noms neutres avec les suffixes flexionnels *-o*, *-e* *-ë* au nominatif singulier ; la classe flexionnelle `II` se compose de noms masculins et féminins avec les suffixes flexionnels *-a* et *-ja*, ainsi que de noms du genre commun ; la classe flexionnelle `III` est constituée de noms féminins sans suffixe flexionnel au nominatif singulier et dont le thème se termine par une consonne molle ou par une chuintante (voir la discussion et les exemples dans la section 1.3.1).

Le tableau 5.10 suggère que les noms de la classe `I` sont majoritaires à servir de base pour la formation des adjectifs.

La corrélation entre les classes flexionnelles et les suffixes est significative (`H Freq` : $\chi^2(4)=157.02$, $p<0.001$; `B Freq` : $\chi^2(4)=133.14$; $p<0.001$) ; cette corrélation est faible (`H Freq` : $\phi = 0.20$; `B Freq` : $\phi = 0.24$).

L'analyse des résidus (figure 5.6) montre que la classe `II` contribue le plus à la

	H Freq			B Freq		
ClFlex	*-Ov-*	*-n-*	*-sk-*	*-Ov-*	*-n-*	*-sk-*
I	445	688	592	187	249	485
II	67	415	166	54	173	51
III	17	89	15	8	17	13

Tableau 5.10: Distribution des classes flexionnelles ; haute et basse fréquence

statistique avec une préférence pour le suffixe *-n-* et une défavorisation du suffixe *-Ov-* dans les données de haute fréquence. En ce qui concerne les données de basse fréquence, c'est le suffixe *-sk-* qui est défavorisé par la classe `II`. Dans les deux ensembles de données, on observe également que la classe `I` défavorise le suffixe *-n-*.

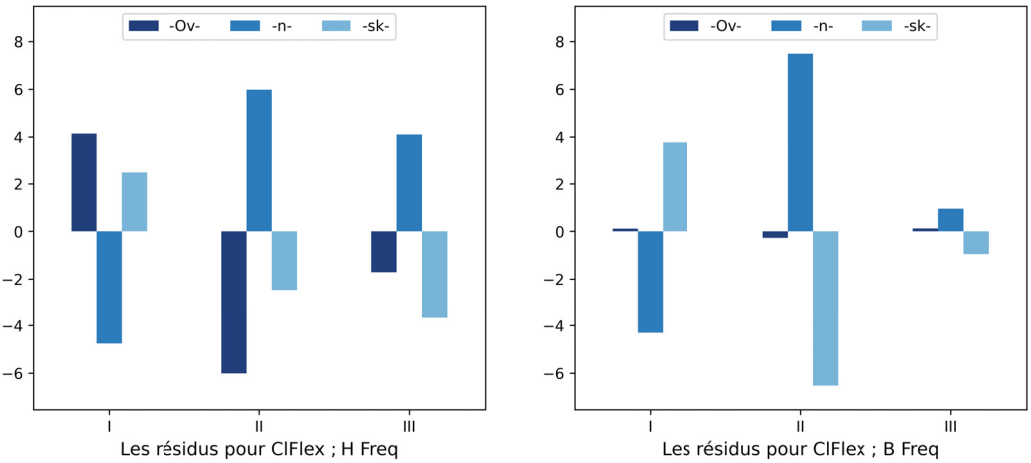

Figure 5.6: Résidus pour les classes flexionnelles ; haute et basse fréquence

La classification de Zaliznjak, basée sur les graphèmes, diffère de celle que nous avons adoptée pour annoter les propriétés du dernier phonème des thèmes dans la section 5.2.2, qui se base sur le point d'articulation des consonnes et qui inclut les voyelles dans une classe distincte. Zaliznjak distingue les thèmes se terminant par des consonnes dures, des consonnes molles, gutturales, chuintantes et la consonne affriquée <*ц*>, regroupe les thèmes se terminant par <*ŭ*> ou par une voyelle autre que <*u*>, ceux se terminant par <*u*>, les noms de la troisième déclinaison, ainsi que les noms indéclinables qui font partie de classes distinctes (cf. également la discussion et les exemples dans la section 1.3.1).

Dans la section 5.2.2, nous avons regroupé certains types accentuels de Zaliznjak, car certains d'entre eux sont sous-représentés et ont des distributions similaires. En revanche, les classes flexionnelles semblent différer davantage dans leurs distributions et présentent donc moins de similitudes (tableau 5.11). De plus, bien que les classes `1`

et 3 soient les plus nombreuses, il n'est pas possible de considérer que les autres classes sont sous-représentées[10].

	H Freq			B Freq		
ClFlexZal	*-Ov-*	*-n-*	*-sk-*	*-Ov-*	*-n-*	*-sk-*
0	1	6	16	4	13	39
1	247	682	433	117	242	334
2	34	53	52	14	5	34
3	156	210	96	84	110	82
4	30	43	14	11	18	19
5	29	29	18	2	3	7
6	6	32	38	5	10	15
7	9	50	91	6	21	7
8	17	87	15	6	17	12

Tableau 5.11: Distribution des classes flexionnelles (Zaliznjak) ; haute et basse fréquence

Il faut noter que la corrélation entre les classes flexionnelles selon la typologie de Zaliznjak et les suffixes est significative (`H Freq` : $\chi^2(16) = 219.37$, p<0.001 ; `B Freq` : $\chi^2(16) = 81.95$, p<0.001), mais faible (`H Freq` : $\phi = 0.18$; `B Freq` : $\phi = 0.23$).

Les résidus (figure 5.7) indiquent que la préférence de la classe flexionnelle 7 (thèmes se terminant par <*u*>) pour le suffixe *-sk-* et celle de la classe flexionnelle 3 (gutturale finale) pour *-Ov-* contribuent le plus à la corrélation dans les données de haute fréquence.

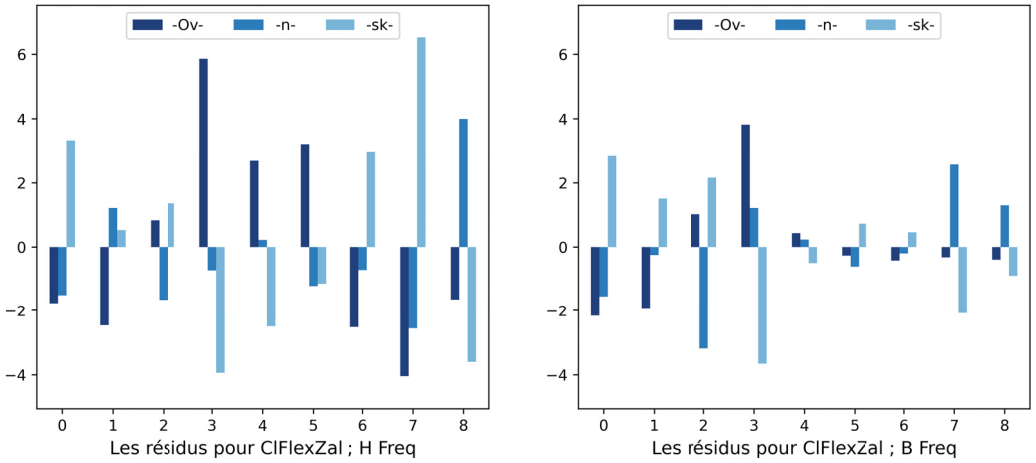

Figure 5.7: Résidus pour les classes flexionnelles (Zaliznjak) ; haute et basse fréquence

[10]Nous nous basons surtout sur les fréquences attendues pour ces classes, et toutes ces fréquences sont supérieures à 5.

On remarque que la classe `8` (la classe flexionnelle canonique `III`) favorise aussi le suffixe *-n-* et que la classe `0` (noms indéclinables) préfère *-sk-*, tandis que le suffixe *-sk-* est moins privilégié par les classes `3` et `8`, la classe `7` tend à éviter *-Ov-*. Dans les données de basse fréquence, on observe également une préférence des classes `0` et `3` pour les suffixes *-sk-* et *-Ov-*, respectivement.

5.2.3.3 Allomorphies thématiques

Nous allons conclure la discussion sur les propriétés morphologiques en analysant les distributions de différentes allomorphies thématiques (cf. la section 1.4). Il s'agit des alternances d'une voyelle avec une voyelle ∅ (1.4.2) et de la mouillure et palatalisation des consonnes, et des transformations linéaires, comme ajout, suppression, remplacement et interférences des segments finaux des thèmes (cf.section 1.4.3). Ces propriétés ont été annotées en fonction de leur présence ou absence dans l'espace thématique.

Les résultats (tableau 5.12) montrent que la plupart des noms ne présentent pas d'alternance voyelle/∅ dans leurs espaces thématiques, que ce soit dans les données de haute ou basse fréquence. La corrélation est statistiquement significative (`H Freq` : $\chi^2(2) = 58.35$, p<0.001 ; `B Freq` : $\chi^2(2) = 98.61$, p<0.001), cette association est plus forte dans les données de basse fréquence (`H Freq` : $\phi = 0.15$; `B Freq` : $\phi = 0.28$).

	H Freq			B Freq		
AllomV	*-Ov-*	*-n-*	*-sk-*	*-Ov-*	*-n-*	*-sk-*
0	436	1034	737	213	347	541
1	93	158	36	36	92	8

Tableau 5.12: Distribution des allomorphies vocaliques ; haute et basse fréquence

Les résidus sont présentés sur la figure 5.8.

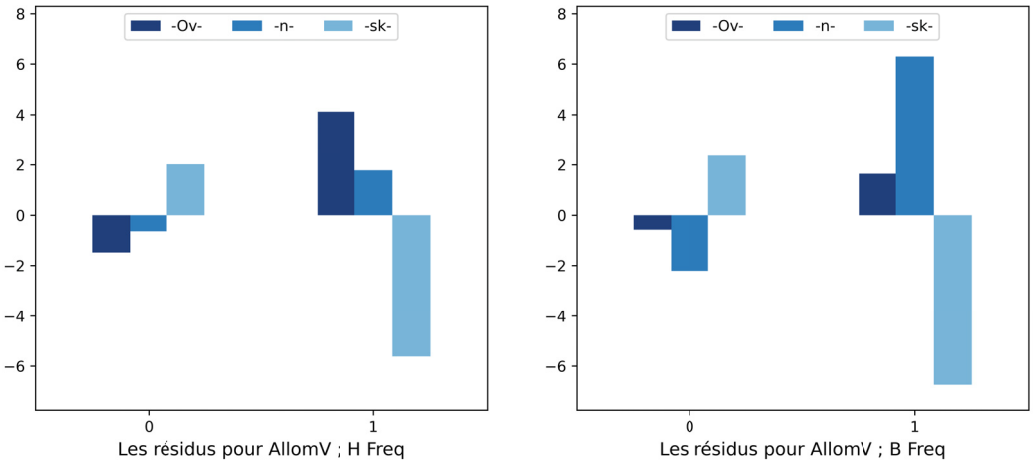

Figure 5.8: Résidus pour les allomorphies vocaliques ; haute et basse fréquence

Il semble que la défavorisation des noms ayant une allomorphie voyelle/∅ pour le suffixe -*sk*- qui contribue le plus à la statistique. Cette tendance est observée dans les deux corpus. Dans les données de haute fréquence, les noms avec ce type d'allomorphie favorisent le suffixe -*Ov*- plutôt que -*n*-. Dans les données de basse fréquence, c'est plutôt le suffixe -*n*- qui est privilégié. En revanche, lorsque les noms ne présentent pas d'allomorphie, c'est le suffixe -*sk*- qui est préféré dans les deux sous-corpus.

Le tableau 5.13 montre que la majorité des noms de base ne présentent pas d'allomorphies consonantiques, tout comme les noms avec les allomorphies vocaliques observés plus haut.

	H Freq			B Freq		
AllomC	-*Ov*-	-*n*-	-*sk*-	-*Ov*-	-*n*-	-*sk*-
0	524	845	668	246	313	522
mouil	3	129	94	3	33	24
palat	2	218	11	0	93	3

Tableau 5.13: Distribution des allomorphies consonantiques ; haute et basse fréquence

La corrélation est significative (`H Freq` : $\chi^2(4) = 294.28$, p<0.001 ; `B Freq` : $\chi^2(2) = 192.02$, p<0.001) et plus forte dans les données de basse fréquence (`H Freq` : $\phi = 0.24$; `B Freq` : $\phi = 0.28$).

Les résidus (figure 5.9) prouvent que ce sont les noms présentant l'allomorphie consonne non palatale / consonne palatale (et leur préférence pour le suffixe -*n*-) qui contribuent le plus à la corrélation.

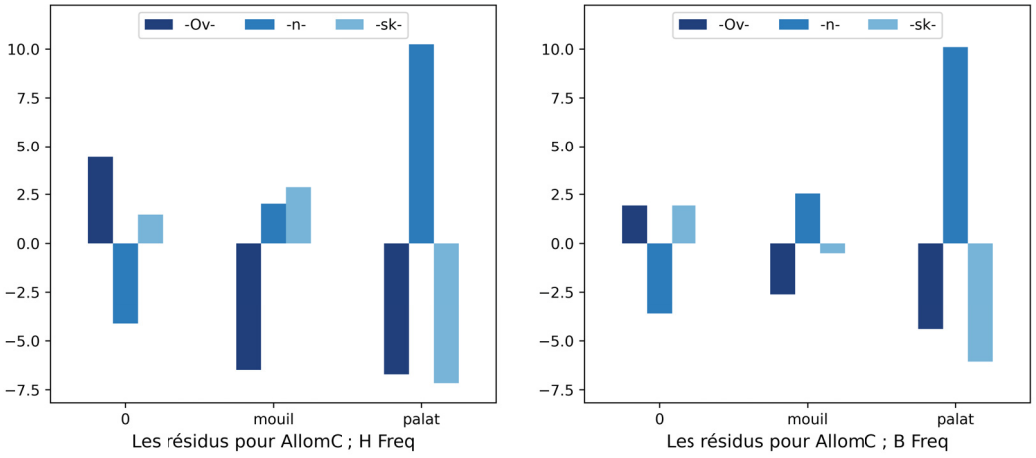

Figure 5.9: Résidus pour les allomorphies consonantiques ; haute et basse fréquence

De plus, on observe simultanément une non-préférence de ces noms pour les suffixes -*Ov*- et -*sk*-. Les noms présentant une alternance consonne dure / consonne molle défavorisent également le suffixe -*Ov*-. Ce suffixe est privilégié lorsque les thèmes ne

présentent pas d'allomorphies consonantiques. De manière générale, le suffixe -*n*- se démarque davantage des suffixes -*sk*- et -*Ov*- dans la direction des résidus pour les noms sans allomorphies et pour les noms ayant un thème palatalisé.

Le tableau 5.14 présente la répartition des allomorphies qui concernent les segments finaux des thèmes. La tendance pour les noms à ne pas avoir ce type d'allomorphie est encore plus marquée que pour les allomorphes vocaliques et consonantiques. La suppression de matériel phonologique et l'interférence avec le suffixe se produisent notamment dans le cas des adjectifs dérivés en -*sk*-. Cependant, les données qui contiennent les allomorphies segmentales sont majoritairement sous-représentées, avec des fréquences attendues inférieures à 5 dans la plupart des cas, ce qui ne permet pas de tirer des conclusions pertinentes avec les métriques statistiques choisies. En raison de la faible fréquence de ce type d'allomorphie, nous exclurons cette propriété de la modélisation de la concurrence (qui sera présentée dans le chapitre 7).

	H Freq			B Freq		
Segm	-*n*-	-*sk*-	-*Ov*-	-*n*-	-*sk*-	-*Ov*-
0	527	1125	621	240	396	516
ajout	1	5	2	2	14	1
inf	0	1	35	3	0	0
rempl	0	4	3	0	2	1
supp	1	57	112	4	27	31

Tableau 5.14: Distribution des allomorphies segmentales finales des thèmes ; haute et basse fréquence

5.2.4 Sémantique

Il n'y a pas d'unanimité entre les linguistes en ce qui concerne la classification des mots sur le plan sémantique. Selon Švedova (1980, p.458), il est possible de classer les noms en différentes catégories, parmi lesquelles figurent les noms propres et les noms communs, les noms collectifs, les noms massifs, les noms concrets et abstraits, ainsi que les noms animés et inanimés. Flaux et Van de Velde (2000, pp.2-39), à leur tour, distinguent les noms propres et communs ; dans la catégorie des noms communs, des sous-catégories sont distinguées, telles que concret/abstrait, extensif/intensif, dénombrable/indénombrable, animé/inanimé, humain/non humain, naturel/fabriqué. Il est généralement admis que, quelle que soit la méthode de classification utilisée, les catégories ne sont pas toujours clairement délimitées et se chevauchent. Néanmoins, les mots de chaque catégorie partagent des caractéristiques morphologiques et syntaxiques communes.

Pour cette étude, nous allons utiliser quatre catégories sémantiques afin d'annoter les caractéristiques des noms de base : les noms propres, les noms humains/animés, les noms concrets et les noms abstraits. Nous allons aborder ces catégories dans les

sections suivantes et nous ferons le point sur les catégories supplémentaires énumérées par Švedova (1980) et Flaux et Van de Velde (2000). Les tests et les exemples pour chaque catégorie sont détaillée dans le guide d'annotation sémantique (Annexe A2).

5.2.4.1 Noms propres

Les noms propres représentent une classe à part entière. Ne faisant pas partie ni des noms concrets, ni des noms abstraits, ils peuvent être à la fois les uns et les autres, selon Raxilina *et al.* (2009) : МГУ (MGU) 'Université d'État de Moscou' comme bâtiment, concret ; Кинотавр (Kinotavr) 'Kinotavr (festival)' comme évènement, abstrait. Švedova (1980, pp.458-459) distingue dans la catégorie des noms propres les noms propres au sens strict du terme (les noms géographiques et astronomiques ainsi que les noms de personnes et d'animaux) et les appellations. Ces derniers concernent les noms communs qui s'emploient en tant que noms propres (174).

(174) *журнал "Здоровье"* 'revue "Zdorov'e"'
žurnal "Zdorov'e"
∼ *здоровье* 'santé'
zdorov'e
завод "Серп и Молот" 'usine "Serp i Molot"'
zavod "Serp i Molot"
∼ *серп и молот* 'faucille et marteau'
serp i molot

Selon Flaux et Van de Velde (2000, pp.2-4), les noms propres présenteraient une plus grande proximité sémantique avec les pronoms déictiques, le démonstratif et les pronoms de première et deuxième personne que les noms communs. Les noms propres ne sont pas porteurs d'un concept ; de ce fait, les noms propres sont caractérisées par Kleiber (1981, p.385) comme prédicats de dénomination. A la différence des noms communs, les noms propres désignent une entité en même temps qu'ils la dénomment. La sémantique des noms propres détermine leur comportement morphologique et syntaxique : ces mots n'ont généralement pas de formes du pluriel[11], et ils sont caractérisés par une absence de la détermination.

D'un point de vue de linguistique de corpus, les noms propres forment, selon (Ljaševskaja et Šarov, 2009, pp.1-21), un cluster moins stable que les noms communs (concrets et abstraits), car leur fréquence dépend fortement des textes sélectionnés et de la période considérée. En l'occurrence, le corpus peut être représentatif de ce que lit actuellement la population russe, ainsi que des noms propres qui apparaissent dans ces textes. Le référencement des noms propres dans des corpus, ensemble avec les noms communs, peut rendre les sources lexicographiques rapidement obsolète. En outre, les

[11]Les formes au pluriel sont toutefois normales pour désigner différentes personnes ayant le même prénom (*в одном классе несколько Светлан* (*v odnom klasse neskol'ko Svetlan*) 'dans une classe, il y a plusieurs Svetlana') ou le même nom de famille (*династия сталеваров Кузнецовых* (*dinastija stalevarov Kuznecovyx*) 'la dynastie des forgerons Kuznetsov') (Švedova, 1980, p.459).

noms de personnages fictifs peuvent être très répétitifs dans une œuvre, ce qui peut fausser leur fréquence totale.

5.2.4.2 Noms humains/animés

Tous les noms peuvent être classés en animés ou inanimés. En russe, comme en français, le couple animé/inanimé est souvent lié au couple humain/non humain : les noms animés font référence aux êtres vivants, tels que les humains et les animaux (175a) ; et les noms inanimés font référence à tous les autres objets et phénomènes (175b) (Švedova, 1980, p.460 ; Flaux et Van de Velde, 2000, p.35).

(175) a. ЧЕЛОВЕК ‘homme’
ČELOVEK
СЫН ‘fils’
SYN
БЕЛКА ‘écureuil’
BELKA
НАСЕКОМОЕ ‘insecte’
NASEKOMOE

b. СТЕНА ‘mur’
STENA
ИНСТИТУТ ‘institut’
INSTITUT
ДОБРОТА ‘bonté’
DOBROTA
ДВИЖЕНИЕ ‘mouvement’
DVIŽENIE

Flaux et Van de Velde (2000, pp.35-37) listent quelques propriétés morpho-syntaxiques des noms animés typiques pour le français, tel que le choix du pronom interrogatif (*qui* pour les noms animés et *que* pour les noms inanimés), les constructions syntaxiques de la possession inaliénable pour les noms animés (avec le verbe *avoir* qui peuvent s’inverser en *être à*) et les constructions instrumentales qui supposent aussi un sujet animé. En ce qui concerne le russe, les noms animés, en règle générale, se répartissent entre le masculin et le féminin, tandis que les noms inanimés sont répartis entre les trois genres morphologiques : masculin, féminin et neutre. De plus, les paradigmes des noms animés et inanimés au pluriel sont distincts. Les noms animés ont la même forme à l’accusatif et le génitif au singulier pour le masculin et au pluriel pour le masculin et le féminin (176a) ; les noms inanimés ont une forme de l’accusatif qui correspond à la forme du nominatif (176b) (Švedova, 1980, pp.460-463).

(176) a.

/нет/	*братьев*	*и*	*сестёр*
/net/	*brat’-ev*	*i*	*sestër*
avoir$_{\text{PRS.NEG}}$	frère-M.GEN.PL	CONJ	sœur$_{\text{F.GEN.PL}}$

‘[il n’y a pas de] frères ni sœurs’
[я *увиде-л]* *брать-ев* *и* *сестёр*
[ja *uvide-l]* *brat’-ev* *i* *sestër*
1.NOM.SG voir-PST.M.SG frère-M.ACC.PL CONJ $\text{sœur}_{\text{F.ACC.PL}}$
‘[j’ai vu] des frères et des sœurs’

b. *книги-и* *[леж-ат* *на* *стол-е]*
knig-i *[lež-at* *na* *stol-e]*
livre-F.NOM.PL poser-PRS.3.PL PREP table-M.LOC.SG
‘des livres [sont sur la table]’
[я *купи-л]* *книг-и,*
[ja *kupi-l]* *knig-i*
1.NOM.SG acheter-PST.M.SG livre-F.ACC.PL
‘[j’ai acheté] des livres’

De plus, comme en français, les noms animés et inanimés diffèrent pour le choix du pronom interrogatif (КТО (KTO) ‘qui’ pour les noms animés ; ЧТО (ČTO) ‘que’ pour les noms inanimés).

5.2.4.3 Noms concrets et abstraits

Il existe différents critères pour déterminer les noms abstraits. Les paramètres linguistiques peuvent inclure des irrégularités formelles (l’absence du pluriel), ainsi que l’existence de synonymes et antonymes. Les critères extra-linguistiques se fondent sur la représentation du concept dans le monde réel : les mots dénotant des objets ou entités matériels occupant une portion définie d’espace et manifestant une forme, perceptibles par les sens, sont considérés comme concrets (177a). Les mots abstraits font référence à une idée ou concept intangible ou à des notions qui ne peuvent pas être perçues d’une autre manière (177b) (Zolotarëva, 2003 ; Schmid, 2012 ; Haas *et al.*, 2022).

(177) a. ДЕРЕВО ‘arbre’
DEREVO
МАШИНА ‘voiture’
MAŠINA
СОБАКА ‘chien’
SOBAKA

b. ЛЮБОВЬ ‘amour’
LJUBOV’
СВОБОДА ‘liberté’
SVOBODA
СЧАСТЬЕ ‘bonheur’
SČAST’E

La tradition selon laquelle l'opposition entre concret et abstrait correspond à l'opposition entre sensible et non-sensible peut être remise en question. Tous les noms abstraits dénotent des entités non-sensibles, mais certains noms concrets peuvent également dénoter des idées ou des concepts qui n'existent que dans l'esprit. En d'autres termes, certains noms concrets peuvent également dénoter des entités non-sensibles. Flaux et Van de Velde (2000, pp.29-39, 56) incluent dans cette catégorie les noms PHRASE, SONATE, POÈME et d'autres semblables[12], ainsi que les noms d'entités imaginaires telles que ANGE ou DRAGON[13], et les entités physiques non visibles, telles que AIR. Les noms concrets et abstraits n'ont pas les mêmes propriétés logiques et morphologiques. Les noms abstraits sont caractérisés par leur nature prédicative, ils sont logiquement issus des prédicats ; de ce fait, les noms abstraits sont massivement dérivés de verbes ou d'adjectifs.

Un autre point de vue sur la classification concret/abstrait est présenté dans Švedova (1980, pp.459-460) qui se base sur la propriété dénombrable/indénombrable. Ainsi, les noms concrets font référence à des entités, des personnes et tous les phénomènes de la réalité qui peuvent être représentés séparément et soumis à calcul (178a). Tous les noms concrets, à l'exception des noms qui n'ont qu'une forme *pluralia tantum*, ont des formes au singulier et au pluriel. Les noms abstraits, en revanche, désignent des concepts abstraits, des propriétés, des qualités, des actions et des états (178b).

(178) a. КАРАНДАШ 'crayon'
KARANDAŠ
КОЛЬЦО 'bague'
KOL'CO
ИНЖЕНЕР 'ingénieur'
INŽENER
БИТВА 'bataille'
BITVA
ВОЙНА 'guerre'
VOJNA

b. СМЕХ 'rire'
SMEX
РАБСТВО 'esclavage'
RABSTVO
ДОБРОТА 'bonté'
DOBROTA
БЛИЗОСТЬ 'proximité'
BLIZOST'

[12]Ces noms ne dénotent pas les réalités, mais les *idéalités concrètes dénombrables* selon la terminologie de Flaux et Van de Velde (2000) ou les *objets cognitifs* selon la terminologie de Haas *et al.* (2022) qui, en absence d'exécution, ne sont pas accessibles aux sens.

[13]Ces entités concrètes rentrent dans la catégorie 'animé' en russe.

ЛОВКОСТЬ ‘dextérité’
LOVKOST’

Outre l’opposition entre les noms concrets et abstraits, d’autres classifications distinguent les noms collectifs, ainsi que les massifs et comptables.

Les noms collectifs désignent des individus constitués d’une pluralité interne d’entités isolables de même type perçues comme un tout (179a) (Flaux et Van de Velde, 2000 ; Benninger, 2001). Les noms appelés communément *собирательные существительные* en russe (179b) ont des caractéristiques grammaticales qui ne sont pas applicables aux noms collectifs : ils ne s’emploient pas au pluriel (179c), ne peuvent pas être déterminés par un quantifieur (179d), ne s’emploient pas dans un syntagme N + $N_{GEN.}$ pour désigner une grande quantité (179e) (Beliakov, 2014, pp.97-98).

(179) a. СТАЯ ‘meute’
STAJA
АРМИЯ ‘armée’
ARMIJA
ЛЕС ‘forêt’
LES
СТАДО ‘troupeau’
STADO

b. ЮНОШЕСТВО ‘jeunesse’
JUNOŠESTVO
АРИСТОКРАТИЯ ‘aristocratie’
ARISTOKRATIJA
БЕДНОТА ‘les pauvres’
BEDNOTA

c. **молодёж-и* ‘jeunesse-F.NOM.PL’
molodëž-i

d. **две молодёжи* ‘deux jeunesses’
dve molodëži

e. **множество молодёжи* ‘ensemble de jeunesse’
množestvo molodëži

Les noms massifs, ou indénombrables, sont à leur tour liés à la continuité, elle-même liée à l’homogénéité : une substance homogène ne présente pas de divisions préconstituées[14]. Ces noms sont aussi utilisés soit uniquement au singulier, soit uniquement au pluriel et font référence aux denrées alimentaires (180a), matériaux (180b), tissus (180c), fossiles, métaux (180d), éléments chimiques, médicaments (180e), cultures agricoles (180f) et autres masses homogènes.

[14]De ce fait, ils sont opposés aux noms dénombrables (Flaux et Van de Velde, 2000, p.33), ce qui suppose la discontinuité et l’existence d’entités distincts et insécables.

(180) a. ЖИР ‘graisse’
ŽIR
КРУПА ‘céréales’
KRUPA
МУКА ‘farine’
MUKA
САХАР ‘sucre’
SAXAR

b. ГИПС ‘plâtre’
GIPS
ЦЕМЕНТ ‘ciment’
CEMENT

c. БАРХАТ ‘velours’
BARXAT
СИТЕЦ ‘satin’
SITEC

d. ЖЕЛЕЗО ‘fer’
ŽELEZO
ИЗУМРУД ‘émeraude’
IZUMRUD

e. УРАН ‘uranium’
URAN
АСПИРИН ‘aspirine’
ASPIRIN

f. ОВЁС ‘avoine’
OVËS
ПШЕНИЦА ‘blé’
PŠENICA

La classification avancée par Švedova (1980) est en contradiction avec deux autres classifications : d’une part, les noms massifs, que Švedova exclut des noms concrets, sont sans doute indénombrables mais ils sont toutefois tangibles. D’autre part, certains noms comptables, tels que ВОЙНА (VOJNA) ‘guerre’, que Švedova classe parmi les noms concrets, ne dénotent ni un concept tangible ni une idéalité.

Dans la présente étude nous adopterons la classification concret/abstrait proposée par Flaux et Van de Velde (2000). De cette manière, la distinction dénombrable/indénombrable ne constitue pas une catégorie des noms séparée, mais croise en partie la distinction concret/abstrait : des noms dénombrables et des noms indénombrables se retrouvent parmi les noms concrets et parmi les noms abstraits.

La distinction entre abstrait et concret n’est cependant pas clairement définie. Certains noms ne peuvent être classés comme concrets ou abstraits qu’en fonction du contexte. Par exemple, ИНСТИТУТ (INSTITUT) ‘institut’ peut être considéré comme

un nom concret ou abstrait en fonction du contexte dans lequel il est utilisé. Si ce terme fait référence à un bâtiment spécifique, il est alors considéré comme un nom concret. En revanche, si le terme fait référence à une organisation ou à une institution créée dans un but spécifique, il est alors considéré comme un nom abstrait. Nous avons évoqué ce type de polysémie dans la section 5.1.2.

Nous avons procédé à l'annotation des propriétés sémantiques des noms de base en fonction de leurs quatre classes sémantiques : les noms propres, les noms humains/animés, les noms concrets et les noms abstraits[15]. Le tableau 5.15 présente la répartition de ces différentes classes.

	H Freq			B Freq		
ClSem	-*Ov*-	-*n*-	-*sk*-	-*Ov*-	-*n*-	-*sk*-
a	153	557	33	74	151	12
c	317	611	20	133	223	15
h	66	44	252	33	45	104
p	5	4	469	9	20	418

Tableau 5.15: Distribution des classes sémantiques ; haute et basse fréquence

Le tableau indique que les noms désignant des êtres humains (`h`) et en particulier les noms propres (`p`) ont tendance à se combiner avec -*sk*-, tandis que les noms abstraits (`a`) et concrets (`c`) ont plutôt tendance à privilégier -*n*-. Ces mêmes tendances peuvent également être observées dans les données de basse fréquence, mais à une moindre échelle.

La corrélation entre les classes sémantiques et les suffixes est significative (`H Freq` : $\chi^2(4) = 1\ 964.73$, $p<0.001$; `B Freq` : $\chi^2(4) = 845.30$, $p<0.001$), et cette association est très forte (`H Freq` : $\phi = 0.62$; `B Freq` : $\phi = 0.58$).

Les résultats de l'analyse des résidus confirment les observations (figure 5.10).

La préférence des noms propres et des noms humains pour le suffixe -*sk*- a la contribution la plus importante à la corrélation. De plus, -*sk*- est défavorisé par les noms concrets et abstraits. Dans les données de basse fréquence, il n'existe presque aucune différence dans les préférences des noms humains pour l'un des trois suffixes. De manière générale, les classes sémantiques permettent de faire une distinction nette entre le suffixe -*sk*- et les suffixes -*n*- et -*Ov*-.

5.2.5 Étymologie

Malgré le fait que l'objectif principal de cette thèse soit l'étude de la morphologie du russe en synchronie, il est impossible d'ignorer les événements historiques qui ont

[15]Le nombre total de fréquences présent dans ce tableau est supérieur à celui observé pour d'autres propriétés des noms de base. Cela est dû au fait que certains noms ont été annotés avec plusieurs catégories telles que concrets et abstraits, humains et concrets, etc. en raison de leur homonymie ou polysémie (voir la discussion dans la section 5.1.2).

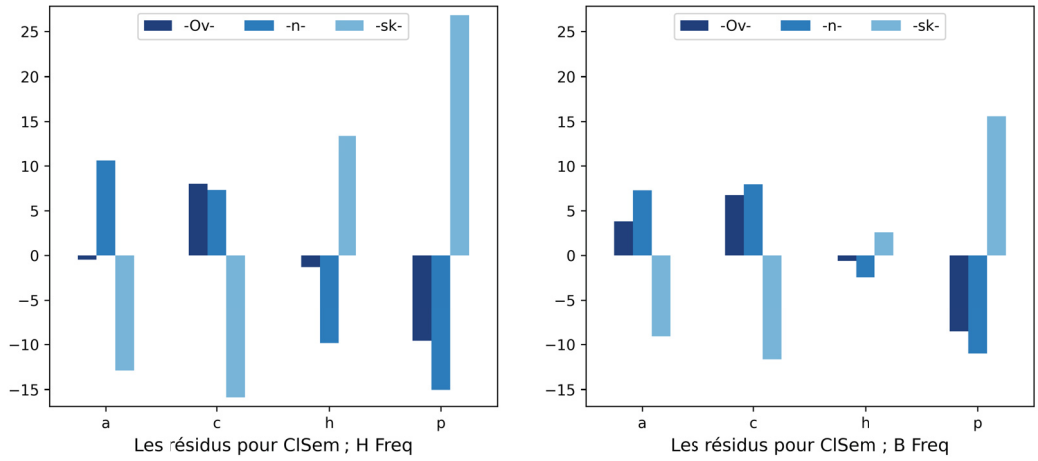

Figure 5.10: Résidus pour les classes sémantiques ; haute et basse fréquence

influencé le développement et la structure actuelle de cette langue. Notre recherche se base sur une classification binaire entre les mots d'origine slave et les emprunts. Les deux étiquettes sont conventionnelles : l'origine slave comprend les mots non seulement d'origine slave, mais aussi originaires du vieux slave, ainsi que les mots formés à partir des racines slaves ; l'emprunt désigne tout mot ayant une origine différente[16].

Tout au long de son évolution, la langue russe a intensément procédé à l'emprunt d'éléments provenant d'autres langues (Krysin, 2008 ; Corbett et Comrie, 2003 ; Breuillard et Viellard, 2015 ; Sakhno, 2015), par exemple, des éléments gréco-latines dans les domaines technique, scientifique, philosophique et politique au XVIIe siècle. Le russe littéraire moderne est né au milieu du XVIIIe siècle et a combiné à l'époque des éléments slavons et russes[17] (Garde, 1998). Depuis 300 ans, le russe s'enrichit de nombreux mots venant entre autre du français, du néerlandais de l'allemand et du polonais (XVIIIe-XIXe siècles), ainsi que, à partir de la seconde moitié du XXe siècle, de l'anglais. Aujourd'hui le russe s'enrichit notamment par les emprunts de l'anglais américain.

Il existe plusieurs approches de classification des emprunts en linguistique contemporaine, qui se basent sur différents critères tels que la langue-source, la période d'emprunt, le domaine de fonctionnement des unités lexicales et le degré d'assimilation du lexique étranger par la langue-réceptrice. Cependant, la classification basée sur le degré d'assimilation est la plus couramment utilisée. Néanmoins, dans notre étude, nous opterons pour une distinction binaire entre les mots d'origine slave et les mots d'autres origines, pour des raisons pragmatiques : afin de ne pas alourdir les annotations et de

[16]Ces étiquettes correspondent principalement à ce que Sakhno (2015) appelle mots 'russes' et mots 'occidentaux'.

[17]Le russe désignait à l'époque la langue parlée, différente de la langue écrite – le slavon : la langue écrite commune des Slaves orthodoxes, appelée communément *vieux slave*.

ne pas introduire une multitude de variables dans l'analyse.

Nous avons utilisé plusieurs dictionnaires pour déterminer l'origine des noms de base des adjectifs dérivés (Semënov, 2003 ; Šanskij, 2004 ; Fasmer, 2006 ; Krylov, 2008). Les exemples ainsi que la méthodologie de travail avec les cas complexes sont détaillés dans le guide d'annotation étymologique (Annexe A3).

Du point de vue morphologique, les éléments slavons peuvent être reconnus formellement. Du point de vue sémantique, ils se rattachent aux couches spontanées et au vocabulaire abstrait. D'après Krysin (2008, p.482), les néologismes déclinables (se terminant par une consonne, qu'ils soient concrets ou abstraits) présentent les plus grandes capacités de formation de mots. Ils peuvent facilement dériver des adjectifs de relation avec les suffixes *-n-*, *-Ov-* et *-sk-* (181).

(181) АБАЖУР 'abat-jour' → АБАЖУРНЫЙ
ABAŽUR ABAŽURNYJ
МАЙОНЕЗ 'mayonnaise' → МАЙОНЕЗНЫЙ
MAJONEZ MAJONEZNYJ
БАРТЕР 'troc' → БАРТЕРНЫЙ
BARTER BARTERNYJ
ГРУНТ 'terre' → ГРУНТОВЫЙ
GRUNT GRUNTOVYJ
КОНСАЛТИНГ 'conseil en gestion' → КОНСАЛТИНГОВЫЙ
KONSALTING KONSALTINGOVYJ
БРОКЕР 'courtier' → БРОКЕРСКИЙ
BROKER BROKERSKIJ

Le tableau 5.16 présente la distribution de l'origine étymologique des noms et les trois suffixes en question.

	H Freq			B Freq		
Source	*-Ov-*	*-n-*	*-sk-*	*-Ov-*	*-n-*	*-sk-*
autre	304	663	637	169	292	507
slave	225	529	136	80	147	42

Tableau 5.16: Distribution d'origine étymologique ; haute et basse fréquence

La corrélation de l'origine et des suffixes est significative (`H Freq` : $\chi^2(2) = 160.30$, p<0.001 ; `B Freq` : $\chi^2(2) = 115.42$, p<0.001), cette correlation est plus forte dans les données de basse fréquence (`H Freq` : $\phi = 0.25$; `B Freq` : $\phi = 0.31$).

L'analyse des résidus (figure 5.11) confirme que la non utilisation du suffixe *-sk-* avec les noms d'origine slave est le principal facteur contribuant à la corrélation, ainsi que la préférence de ce suffixe pour les noms d'origine étrangère. La distribution des suffixes *-n-* et *-Ov-* est similaire et s'oppose à la distribution du suffixe *-sk-*.

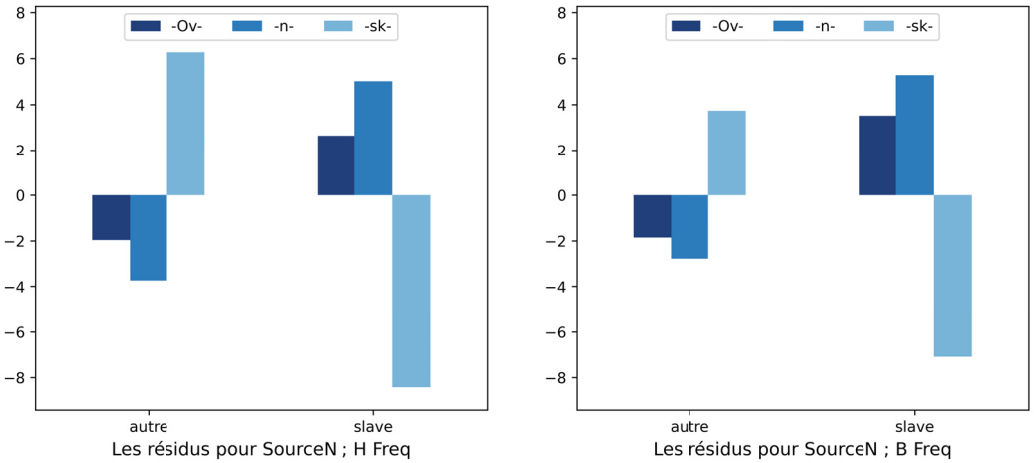

Figure 5.11: Résidus pour l'origine étymologique ; haute et basse fréquence

Conclusion

Dans ce chapitre, nous nous sommes concentrés sur l'annotation des propriétés des noms de base et leur distribution en fonction des suffixes *-n-*, *-sk-* et *-Ov-*. Cependant, nous avons rencontré plusieurs difficultés dans la détermination d'un nom de base unique, en raison du fait que les adjectifs peuvent être motivés par plusieurs lexèmes (noms mais aussi verbes), et du fait que de nombreux noms russes sont polysémiques. Nous avons donc exclu de notre analyse les adjectifs dérivés contenant la séquence *-itel'* et avons procédé, dans certains cas, à une double annotation sémantique si le sens du nom peut être caractérisé comme concret ou abstrait en fonction des contextes.

L'étude des distributions des propriétés des noms de base en relation avec les suffixes *-n-*, *-sk-* et *-Ov-* a mis en évidence le fait que toutes les corrélations sont statistiquement significatives, mais que leur ampleur diffère. Le tableau 5.17 synthétise les résultats.

Les seules propriétés des noms de base fortement corrélées aux suffixes sont leurs classes sémantiques. Les propriétés modérément corrélées sont la structure syllabique et les derniers phonèmes des radicaux dans les données de haute fréquence. Toutefois, la longueur des noms de base en syllabes est assez faible dans les données de basse fréquence. De plus, dans les données de basse fréquence l'étymologie et les allomorphies consonantiques et vocaliques sont plus fortement corrélées au choix entre les suffixes, comparé aux données de haute fréquence.

L'analyse des résidus a révélé que certaines propriétés des noms de base permettent de distinguer assez clairement les suffixes concurrents. Ainsi, la longueur des noms de base en syllabes permet d'isoler *-Ov-* des deux autres suffixes, les derniers phonèmes des radicaux contribuent aussi au démarquage de *-Ov-* ; les classes sémantiques et l'origine étymologique marquent les frontières entre *-sk-* et les deux autres suffixes ; enfin,

	H Freq					B Freq			
Prop	χ	dl	p	ϕ	Prop	χ	dl	p	ϕ
ClSem	1964.73	6	0	0.62	ClSem	845.30	6	0	0.58
SyllN	518.65	8	0	0.32	DPhoR	293.19	8	0	0.34
DPhoR	431.39	8	0	0.29	SourceN	115.42	2	0	0.31
SourceN	160.30	2	0	0.25	AllomV	98.61	2	0	0.28
AllomC	294.28	4	0	0.24	AllomC	192.02	4	0	0.28
AccZal	243.87	6	0	0.22	Genre	136.87	4	0	0.24
ClFlexZal	219.37	16	0	0.21	ClFlex	133.14	4	0	0.23
Genre	198.17	4	0	0.20	SyllN	114.33	8	0	0.21
ClFlex	157.02	4	0	0.18	ClFlexZal	81.95	16	0	0.18
AccSyllN	141.41	4	0	0.17	AccZal	67.24	6	0	0.16
AllomV	58.35	2	0	0.15	AccSyllN	37.40	4	0	0.12

Tableau 5.17: Récapitulatif des statistiques ; haute et basse fréquence

la distribution des derniers phonèmes des radicaux, des genres et des allomorphies consonantiques permet de différencier tous les trois suffixes -*n*- de -*sk*- et -*Ov*-.

Toutefois, ces conclusions ont été établies à partir d'analyses univariées, où chaque distribution de propriétés des noms de base a été examinée de manière isolée. Par conséquent, il n'est pas possible de comprendre comment ces propriétés interagissent lorsqu'elles sont combinées. Dans le chapitre 7, nous utiliserons une modélisation multivariée pour étudier ces interactions et déterminer comment chacune de ces propriétés influence la concurrence dans ces conditions.

Avant de passer à la modélisation multivariée, il reste nécessaire de résoudre deux problèmes relatifs à l'annotation des noms de base : l'annotation sémantique et l'annotation étymologique, qui ont été faites manuellement. Même si nous avons suivi un protocole qui couvre la majorité des cas, les résultats de ces annotations peuvent être potentiellement contestables. De plus, la granularité des classes prédéfinies peut être sujette à débat. Nous allons nous concentrer sur les approches qui permettent d'éviter ce genre de problèmes dans le chapitre 6.

Chapitre 6

Approches aux données catégorielles

Sommaire

Introduction

Dans le chapitre 5, nous avons présenté les propriétés linguistiques des noms ainsi que leurs annotations. Toutefois, nous avons conclu que les propriétés sémantiques et étymologiques peuvent poser des problèmes, car leur annotation dépend du nombre de classes préalablement définies et peut également être affectée par les biais des annotateurs. Dans les sections 5.2.4 et 5.2.5, nous avons procédé à des annotations manuelles, en caractérisant l'étymologie au moyen de deux valeurs (slave/non-slave) et la sémantique au moyen de quatre valeurs (propre, humain ou animé, concret, abstrait). Cette partie de l'étude présentera deux approches distinctes développées pour rendre les annotations sémantiques et étymologiques plus objectives (ou pour réduire leur subjectivité) d'un côté ; et, de l'autre côté, pour introduire davantage de variation (par

rapport à la distinction entre les variables catégorielles) en quantifiant la sémantique et l'étymologie des noms.

Dans la section 6.1 nous présenterons une méthodologie pour s'emparer complètement de l'annotation manuelle d'origine étymologique des noms et pour annoter la totalité du corpus automatiquement avec des scores étymologiques (Bobkova et Montermini, 2021). Dans la section 6.2 nous présenterons une méthodologie inspirée de Fedden *et al.* (2021) qui se base sur les annotations manuelles de la sémantique (cf. la section 5.2.4) mais qui vise à réduire leurs biais en introduisant les scores sémantiques. Ces derniers seront aussi calculés automatiquement pour chaque nom de base.

Puisqu'il s'agit des variables continues, nous utiliserons des tests statistiques appropriés : l'ANOVA (Analyse de Variance) et le test de Kruskal-Wallis. L'ANOVA est une méthode statistique qui permet de déterminer si les moyennes de plusieurs groupes sont significativement différentes les unes des autres et qui se base sur les variances intra- et intergroupes[1]. En l'occurrence, nous l'utiliserons pour comparer les scores sémantiques ou étymologiques moyens entre les trois suffixes en question. L'hypothèse nulle testée par cette méthode est que les scores moyens pour chaque suffixe sont équivalents. Comme dans le cas de test χ^2 que nous avons utilisé dans le chapitre 5, nous effectuerons les calculs d'ANOVA à l'aide de la librairie `SciPy` en python (Virtanen *et al.*, 2020) qui fournit la valeur p appropriée pour la statistique F. Si p est inférieur à 0.05, l'hypothèse nulle sera rejetée. Il est à noter que l'ANOVA peut fournir des estimations inexactes de la valeur p lorsque les données ne sont pas normalement distribuées. Ainsi, nous allons compléter les résultats obtenus avec l'ANOVA par le test de Kruskal-Wallis. Ce dernier, en tant qu'équivalent non paramétrique de l'ANOVA, n'émet aucune hypothèse quant à la normalité de la distribution[2]. Le test de Kruskal-Wallis fournit également une statistique H et la valeur p, et sera aussi implémenté en `SciPy`.

6.1 Scores étymologiques

Dans la section 5.2.5, nous avons effectué une annotation étymologique basée sur les informations provenant de dictionnaires. Cependant, comme le remarquent Aronoff et Fuhrhop (2002, p.469) sur l'exemple de l'anglais, (l'observation peut être étendue à d'autres langues), il n'est pas clair comment les connaissances étymologiques sont reflétées dans la conscience linguistique des locuteurs naïfs, qui ne sont pas obligatoirement experts en histoire de la langue. De plus, Bauer *et al.* (2015, p.583) affirment que la distinction entre le lexique natif et le lexique emprunté est faiblement présente, voire inexistante, dans la conscience linguistique des locuteurs. Néanmoins, Adams (2014) avance l'hypothèse selon laquelle les locuteurs peuvent mobiliser cette distinction en se basant sur les propriétés phonologiques des mots, telles que la

[1] Cf. Brezina (2021, pp.482-484) pour le détail des calculs.

[2] Cf. Levshina (2015, pp.178-179) ou Brezina (2021, pp.486-487) pour la discussion.

structure syllabique et la position de l'accent tonique (par exemple, en anglais, les mots monosyllabiques ou bisyllabiques avec l'accent tonique sur l'avant-dernière syllabe sont d'origine native, tandis que les mots d'origine étrangère sont généralement trisyllabiques ou plus et ont l'accent tonique en début de mot).

Dans cette section, nous partirons du principe que les locuteurs natifs ne possèdent pas nécessairement de connaissances en étymologie, et nous nous appuierons sur la structure graphique des mots[3] pour proposer une méthode de classification qui distingue entre les noms slaves et les noms empruntés. Cette méthode sera basée sur les bigrammes, c'est-à-dire les combinaisons de deux graphèmes, qui sont généralement rencontrés dans le lexique slave et le lexique étranger. Pour constituer des corpus représentatifs du lexique prototypiquement slave et prototypiquement étranger, nous utiliserons les données du Wiktionnaire.

6.1.1 Wiktionnaire

Le Wiktionnaire est un projet collaboratif en ligne visant à créer un dictionnaire multilingue libre et gratuit pour toutes les langues du monde. Cette ressource est entretenue par une communauté de bénévoles qui contribuent à la collecte, à la définition, à la traduction et à la structuration des données lexicales. Le Wiktionnaire fonctionne et s'organise de manière similaire à Wikipédia, mais son objectif est de fournir des informations sur les définitions, les synonymes, les antonymes, les étymologies et les traductions des mots plutôt que des articles encyclopédiques.

Actuellement, le Wiktionnaire russe[4] contient 1 255 179 articles sur les mots, les affixes et les expressions[5], y compris des termes techniques, de l'argot, des injures et même des variantes d'orthographe incorrectes utilisées dans la littérature pour représenter des dialectes ou des expressions populaires et orales (comme *щас* (*ščas*) < *сейчас* (*sejčas*) 'maintenant' ou *чё* (*čë*) < *что* (*čto*) 'quoi'). Lors de la constitution du Wiktionnaire, les sources principales mobilisées sont les dictionnaires ainsi que des corpus. La présence d'un mot dans une langue donnée doit être validée par au moins une de ces sources (par exemple, le mot doit figurer dans l'un ou plusieurs dictionnaires de référence ou être employé de manière régulière dans les corpus). En cas de désaccord sur certaines propriétés d'une unité lexicale, la priorité (en termes de fiabilité) est donnée aux données des corpus.

La conception de ce dictionnaire vise à fournir une description intégrale et uniforme de toutes les unités lexicales, les affixes et expressions idiomatiques et figées d'une langue donnée. Le Wiktionnaire est uniforme en raison de la standardisation des modèles de rédaction d'articles. Cette uniformité rend le dictionnaire facilement utilisable comme source d'information lexicographique et permet également une maintenance aisée et efficace. Le Wiktionnaire combine les fonctions de divers types de dictionnaires

[3]Nous utilisons la structure graphique pour des raisons de commodité, l'orthographe et la phonologie étant très similaires en russe.

[4]Accessible via le lien : https://ru.wiktionary.org/wiki/Викисловарь:Заглавная_страница.

[5]Au 09 mars 2023.

classiques, tels que les dictionnaires explicatifs, orthographiques, grammaticaux, phraséologiques et étymologiques, ainsi que les thésaurus, la présence des informations depuis chaque type de dictionnaire est reflétée dans la structure des articles. C'est le niveau étymologique qui nous intéresse davantage dans le cadre de cette étude.

Pour pouvoir établir les scores étymologiques, nous nous sommes basés sur deux ensembles de données du Wiktionnaire : le lexique général[6] et les emprunts[7] en russe.

En utilisant ces deux bases de données du Wiktionnaire, nous avons obtenu une liste de tous les lemmes et une liste de mots étiquetés comme des emprunts. Cette méthode présente l'avantage d'avoir une source lexicographique préexistante pour les mots d'origine étrangère, évitant ainsi une annotation manuelle. Cependant, l'inconvénient est que le vocabulaire slave n'est pas explicitement répertorié ; nous pouvons seulement accéder à l'ensemble du corpus. Pour obtenir une approximation du vocabulaire slave, il est nécessaire de procéder à un nettoyage du corpus tout venant. Le tableau 6.1 présente le volume de données dans les deux ensembles avant et après les filtrages[8].

Lexique	Volume
'Slave'	14 996
	(192 585 avant les filtrages)
Emprunté	15 007
	(10 962 avant le ré-équilibrage)

Tableau 6.1: Données de Wiktionnaire

En comparant le volume de données entre le lexique tout venant et le lexique emprunté, on constate une différence significative, le lexique tout venant étant presque 18 fois plus important que le lexique emprunté. Cette différence s'explique par le fait que le lexique tout venant comprend à la fois le lexique étranger et une grande quantité de noms propres ainsi que de mots dérivés, tels que des mots composés, suffixés et préfixés, ce qui n'est pas le cas pour le lexique emprunté (toutefois, certains dérivés et certains noms propres sont faiblement représentés dans le lexique emprunté). En outre, il est important de noter que tous les emprunts présents dans le corpus tout venant ne sont pas nécessairement répertoriés dans le lexique emprunté.

Le premier filtrage a exclu le lexique étranger du lexique courant, résultant en 185 098 entrées restantes dans le lexique 'slave'. Cependant, ce lexique reste très bruyant. Par exemple, dans le lexique étranger, seuls les mots операция (operacija) 'opération' et контроперация (kontroperacija) 'contre-opération' sont répertoriés, tandis que dans le lexique courant, après ce premier filtrage, on trouve également des mots tels que псевдооперация

[6]Lien : https://ru.wiktionary.org/wiki/Индекс:Русский_язык.

[7]Lien : https://ru.wiktionary.org/wiki/Категория:Заимствования_в_русском_языке.

[8]Dans ce contexte, l'étiquette `slave` est encadrée par des guillemets, car les données avant les filtrage incluent le lexique général. Ce n'est qu'après l'application de différents filtres que l'ensemble de données vise à se rapprocher autant que possible du lexique prototypiquement slave.

(PSEVDOOPERACIJA) ‘pseudo opération’, СПЕЦОПЕРАЦИЯ (SPECOPERACIJA) ‘opération spéciale’, ФИНОПЕРАЦИЯ (FINOPERACIJA) ‘opération financière’, КРИПТООПЕРАЦИЯ (KRIPTOOPERACIJA) ‘opération crypto’, etc. Le lexique étranger contient aussi des mots tels que АБСТРАКТ (ABSTRAKT) ‘abstrait$_{N}$’, АБСТРАКТНОСТЬ (ABSTRAKTNOST’) ‘abstraction’, АБСТРАКЦИОНИЗМ (ABSTRAKCIONIZM) ‘abstractionnisme’, АБСТРАКЦИОНИСТ (ABSTRAKCIONIST) ‘abstractionniste$_{M}$’ et АБСТРАКЦИОНИСТКА (ABSTRAKCIONISTKA) ‘abstractionniste$_{F}$’, tandis que le lexique courant, après le filtrage, contient tout de même АБСТРАКТИВИЗМ (ABSTRAKTIVIZM) ‘abstractivisme’.

Le deuxième filtrage a été nécessaire afin d’enlever ce type de problème et de supprimer les mots du lexique tout venant qui contiennent partiellement des mots du lexique étranger. Ce deuxième filtrage a résulté en 70 722 mots. Le volume implique la possibilité d’une pollution supplémentaire dans le corpus.

Le troisième filtrage a éliminé tous les mots qui ont des préfixes ou des suffixes d’origine étrangère (*-acija*, *anti-*, *audio-*, *avto-*, *astro-*, *dis-*, *endo-*, *evro-*, *-gen*, *gipo-*, *giper-*, *-ing*, *-izm*, *kvazi-*, *mega-*, *mikro-*, *mono-*, *-morf*, *nano-*, *-oid*, *psevdo-*, *retro-*, *tele-*). Ce dernier processus de filtrage automatique a abouti à un total de 23 600 mots. Les mots étrangers qui ont été supprimés du lexique tout venant ont été transférés dans le lexique des emprunts pour rééquilibrer les corpus. Ainsi, le volume du lexique étranger a atteint 15 007 entrées.

La dernière étape a été la révision manuelle du lexique tout venant pour éliminer tout bruit supplémentaire. Ce processus a permis d’obtenir un corpus final de 14 996 mots qui représente au mieux l’ensemble du vocabulaire slave. En outre, les tailles des deux corpus, le lexique slave et le lexique étranger, sont équilibrées.

6.1.2 Bigrammes

Nous avons constitué deux ensembles de données distincts : le lexique slave et le lexique emprunté. Nous avons ensuite procédé à l’analyse des bigrammes de chaque mot, en incluant le symbole # qui indique une frontière de mot. Des exemples des bigrammes sont présentés dans le tableau 6.2.

Mot	Bigramme
обман	`#о об бм ма ан н#`
obman	
‘mensonge’	
скорпион	`#с ск ко ор рп пи ио он н#`
skorpion	
‘scorpion’	

Tableau 6.2: Exemple de division de mots en bigrammes

L’idée sous-jacente à l’utilisation des bigrammes consiste à examiner les

combinaisons de deux lettres qui apparaissent dans le lexique d'origine slave et dans le lexique emprunté, ainsi que leurs fréquences respectives. En prenant en compte la fréquence des bigrammes dans les ensembles de données slaves et étrangères, nous pouvons placer chaque bigramme sur une ligne continue pour évaluer dans quelle mesure il se rapproche d'une combinaison typiquement étrangère ou slave, ou s'il se situe quelque part entre les deux. De plus, en travaillant avec les bigrammes à partir des deux ensembles de données extraits de Wiktionnaire, nous pouvons éviter tout biais de codage manuel.

Les mots d'origine slave présentent un nombre inférieur de types de bigrammes (874) comparé aux mots empruntés (888). Le travail avec ces deux ensembles de données a abouti à l'obtention d'un total de 961 bigrammes uniques (tableau 6.3).

Lexique	Types	Tokens
Slave	874	126 222
Emprunté	888	129 551
Total	961	

Tableau 6.3: Bigrammes (types et tokens) dans le lexique slave et le lexique emprunté

Le processus de calcul des scores étymologiques des bigrammes est résumé dans le tableau 6.4. Nous avons compté le nombre d'occurrences de chaque bigramme unique dans les deux ensembles de données (`sl_count`, `fo_count`), puis nous avons calculé le pourcentage d'occurrence de chaque bigramme par rapport au nombre total d'occurrences dans chaque ensemble (`sl_prop`, `fo_prop`). En utilisant les pourcentages, nous avons pu attribuer un poids plus important aux bigrammes qui ont une fréquence élevée (`ка`, `ия`), tandis que les bigrammes avec une fréquence plus faible (`ап` ou `ут`) ont eu un score proche de zéro et donc un impact moins significatif sur le score final. Enfin, pour combiner les résultats des deux ensembles en un seul score (`sl-fo`), nous avons choisi de soustraire les scores : si un bigramme apparaît plus fréquemment dans le lexique slave que dans le lexique emprunté, son score sera positif (`ка`, `ап`), et vice versa (`ут`, `ия`).

Bigramme	sl_count	fo_count	sl_prop	fo_prop	sl–fo
ка	3 039	1 520	2.407663	1.173283	1.234380
ап	313	311	0.247976	0.240060	0.007916
ут	286	306	0.226585	0.236200	-0.009616
ия	63	1 552	0.049912	1.197984	-1.148072

Tableau 6.4: Méthode de calcul des scores étymologiques de bigrammes

L'échantillon de bigrammes compte 961 observations, avec une valeur minimale de -1.1481 et une valeur maximale de 1.8185. La moyenne étant de 0 avec un écart-type de 0.1939. La médiane (0.0008) et la moyenne sont presque identiques. Cette distribution nous permet de déterminer le seuil étymologique : les bigrammes ayant des valeurs

positives sont considérés comme représentatifs du lexique slave, tandis que ceux ayant des valeurs négatives sont caractéristiques du lexique étranger. Le tableau 6.5 présente dix bigrammes – cinq slaves et cinq étrangers – avec les scores les plus élevés, pouvant ainsi être considérés comme prototypiquement slaves ou étrangers.

	Bigramme	count_sl	count_fo	sl–fo
Slaves	е#	2 652	366	1.818546
	а#	4 340	2 189	1.748704
	ка	3 039	1 520	1.234380
	#п	2 407	1 115	1.046293
	по	1 738	508	0.984815
Étrangers	ия	63	1 552	-1.148072
	#а	250	1 524	-0.978307
	я#	498	1 707	-0.923085
	т#	495	1 560	-0.811993
	ци	52	1 069	-0.783960

Tableau 6.5: Les bigrammes les plus prototypiquement slaves et les plus prototypiquement étrangers

6.1.3 Scores étymologiques

La dernière étape consiste à projeter les scores de chaque bigramme obtenus dans la section 6.1.2 sur les noms de base dans les données de haute et de basse fréquence de RuDénom. Étant donné que les bigrammes ont des valeurs positives et négatives et sont centrés autour de zéro, pour calculer le score étymologique de chaque nom, nous avons additionné les valeurs de chaque bigramme qui compose ce nom. En conséquence, les noms avec un score étymologique positif tendent à être d'origine slave, tandis que les noms avec un score étymologique négatif tendent à être d'origine étrangère.

Le tableau 6.6 présente la distribution des scores étymologiques obtenus pour les noms de base des adjectifs de haute et de basse fréquence, la figure 6.1 systématise ces résultats.

	H Freq			B Freq		
Stats	-*Ov*-	-*n*-	-*sk*-	-*Ov*-	-*n*-	-*sk*-
moyenne	-0.31	0.05	-1.12	-0.20	0.01	-1.06
écart type	1.48	1.92	1.61	1.69	2.20	1.35
min	-4.25	-5.20	-6.63	-4.34	-5.00	-5.29
max	5.05	5.53	4.37	4.43	5.31	3.51

Tableau 6.6: Distribution des scores étymologiques ; haute et basse fréquence

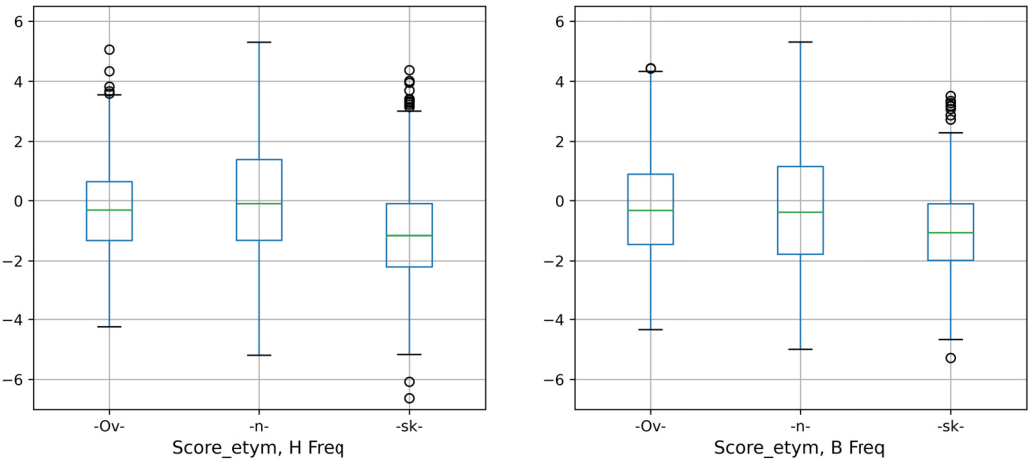

Figure 6.1: Scores étymologiques ; haute et basse fréquence

Les tendances observées dans les données de haute et de basse fréquence sont similaires. Près de 75% des noms qui se combinent avec le suffixe -*sk*- ont un score étymologique négatif, ce qui indique que ce suffixe a une préférence pour les noms d'origine étrangère. En revanche, les distributions des scores pour les suffixes -*n*- et -*Ov*- sont globalement centrées autour de 0 (avec une légère tendance pour la moyenne du suffixe -*Ov*- à être poussée vers les scores négatifs), ce qui implique qu'il y a moins de préférences au niveau de l'étymologie des noms de base pour ces deux suffixes. Ces observations sont généralement en accord avec l'analyse effectuée dans la section 5.2.5, où l'étymologie a été considérée comme une propriété binaire. Selon cette méthodologie d'annotation, le suffixe -*sk*- préférait également les noms d'origine étrangère. Les distributions pour les suffixes -*n*- et -*Ov*- étaient similaires, cependant, selon la distribution des résidus, il y avait une légère préférence pour les noms d'origine slave.

Les moyennes des scores étymologiques entre les trois suffixes présentent des différences significatives (à la fois avec les tests paramétriques et non paramétriques) qui ne peuvent pas être attribuées au simple hasard (`H Freq` : F=113.02, p<0.001, H=197.51, p<0.001 ; `B Freq` : F=49.47, p<0.001, H=78.46, p<0.001).

Nous avons aussi comparé les scores étymologiques à l'annotation manuelle de l'étymologie. Le tableau 6.7 présente le taux d'accord entre ces annotations[9]. En majorité (entre 68% et 77%), les scores obtenus par calcul automatique correspondent aux annotations manuelles.

Les tableaux 6.8 et 6.9 présentent les lexèmes qui peuvent être considérés comme les plus prototypiquement slaves et les plus prototypiquement étrangers dans les deux

[9]La stricte équivalence à zéro n'est pas présente car aucun score étymologique de zéro n'est observé dans les bases de données.

	H Freq	B Freq
`Score_etym` > 0 & `Source` == `slav`	68.46%	71.82%
`Score_etym` < 0 & `Source` == `autre`	77.44%	76.60%

Tableau 6.7: Couverture des scores étymologiques

corpus de haute et de basse fréquence (`Score_etym`). Les résultats sont cohérents à l'annotation manuelle (`Source`) : tous les lexèmes ayant un score négatif ont été étiquetés comme étrangers, et vice versa.

	Lexème	Source	Score_étym
Slaves	*поясница* 'bas du dos' *pojasnica*	slave	5.303177
	воскресенье 'dimanche' *voskresen'e*	slave	5.137607
	печёнка 'foie' *pečënka*	slave	5.073001
	пенька 'chanvre' *pen'ka*	slave	5.053308
	плёнка 'pellicule' *plënka*	slave	5.051385
Étrangers	*Ингерманландия* 'Ingrie (Russie)' *Ingermanlandija*	autre	-6.628645
	артиллерия 'artillerie' *artillerija*	autre	-6.079748
	аудитория 'public' *auditorija*	autre	-5.198070
	Александрия 'Alexandrie' *Aleksandrija*	autre	-5.169374
	кондитер 'pâtissier' *konditer*	autre	-5.077568

Tableau 6.8: Les lexèmes les plus prototypiquement slaves et les plus prototypiquement étrangers ; haute fréquence

Parmi les noms qui sont considérés comme les plus représentatifs du lexique d'origine slave, on trouve quelques bigrammes qui sont également considérés comme prototypiques, tels que `а#` et `ка` qui sont très nombreux, `#п` qui est aussi souvent présent. On observe également `во` et `нь` (comme dans ВОСКРЕСЕНЬЕ (VOSKRESEN'E) 'dimanche'), qui font partie des dix bigrammes les plus prototypiquement slaves. Pour ce qui est des noms considérés comme les plus prototypiquement étrangers, on observe les bigrammes les plus étrangers `ия`, `я#` dans la plupart des cas, ainsi que `#а` et `т#`. Tous ces bigrammes correspondent, à leur tour, aux bigrammes typiquement étrangers.

	Lexème	Source	Score_étym
Slaves	*скакалка* ‘corde à sauter’ *skakalka*	slave	5.307946
	здравоохранение ‘système de santé’ *zdravoohranenie*	slave	5.105303
	правнучка ‘arrière petite fille’ *pravnučka*	slave	4.934499
	сводка ‘communiqué’ *svodka*	slave	4.777330
	выправка ‘alignement’ *vypravka*	slave	4.727759
Étrangers	*жандармерия* ‘gendarmerie’ *žandarmerija*	autre	-5.287943
	фарингит ‘pharyngite’ *faringit*	autre	-4.998637
	ландмилиция ‘police militaire’ *landmilicija*	autre	-4.671550
	Фредериксберг ‘Frederiksberg’ *Frederiksberg*	autre	-4.568788
	истерия ‘hystérie’ *isterija*	autre	-4.568710

Tableau 6.9: Les lexèmes les plus prototypiquement slaves et les plus prototypiquement étrangers ; basse fréquence

Nous avons également comparé certains cas problématiques pour l’annotation manuelle[10] avec les scores étymologiques calculés à base de bigrammes. Le tableau 6.10 en présente un échantillon.

Les noms classés dans la catégorie (i) proviennent des langues non-slaves parlées sur le territoire de la Fédération de Russie. Les noms de la catégorie (ii) sont des exemples de mots étrangers empruntés très tôt dans l’histoire de la langue (comme en vieux slave, par exemple). Les exemples de la catégorie (iii) représentent des mots dont l’origine est inconnue ou discutable. La colonne `Source` (qui représente l’annotation manuelle) montre que les noms des catégories (i) et (ii) ont été annotés comme étant étrangers, tandis que les noms de la catégorie (iii) ont été annotés comme étant slaves, conformément au guide d’annotation étymologique. La colonne `Score_etym` présente les résultats d’attribution de scores étymologiques à ces noms.

Les scores étymologiques des noms référencés en (i) et (ii) sont en général cohérents avec l’annotation manuelle, à l’exception d’un score positif pour le mot ТАЙГА (TAJGA) ‘taïga’. En revanche, il y a plus de variation dans les scores des noms en (iii). Dans le cas de ДУРМАН (DURMAN) ‘datura’, tous les bigrammes qui composent ce mot ont

[10]Certains de ces cas sont référencés dans le guide d’annotation étymologique (Annexe A3).

	Lexème	Source	Score_etym
(i)	*Зея* ‘Zeïa (Russie)’ *Zeja*	autre	-0.346272
	Майкоп ‘Maïkop (Russie)’ *Majkop*	autre	-1.696462
	Стерлитамак ‘Sterlitamak (Russie)’ *Sterlitamak*	autre	-3.148773
	таз ‘bassine’ *taz*	autre	-0.511984
	тайга ‘taïga’ *tajga*	autre	0.925702
(ii)	*сапфир* ‘saphir’ *sapfir*	autre	-0.976759
	сахар ‘sucre’ *saxar*	autre	-1.060928
	тын ‘palissade’ *tyn*	autre	-0.545583
(iii)	*дурман* ‘datura’ *durman*	slav	-1.752995
	козырь ‘atout’ *kozyr’*	slav	-1.049414
	мочевина ‘urée’ *močevina*	slav	1.788217
	сапсан ‘faucon pèlerin’ *sapsan*	slav	-0.721113
	сапог ‘botte’ *sapog*	slav	0.888582
	мусор ‘ordure’ *musor*	slav	-1.168993
	галька ‘galet’ *gal’ka*	slav	1.609973
	Гатчина ‘Gatchina (Russie)’ *Gatčina*	slav	1.038211

Tableau 6.10: Échantillon des scores étymologiques pour les cas ambigus

un score étymologique négatif, à l’exception de ду qui présente un score positif, bien que faible (0.0415). Dans le cas de САПСАН (SAPSAN) ‘faucon pèlerin’, quatre sur sept bigrammes ont un score étymologique négatif (ан, н#, са, пс), ce qui rend le score final négatif. Dans le cas de САПОГ (SAPOG) ‘botte’, le nombre de bigrammes positifs et négatifs est équivalent (по, #с, ап et са, ог, г# respectivement), mais les scores positifs sont plus élevés.

6.2 Scores sémantiques

Dans la section 5.2.5 nous avons utilisé des données de dictionnaires pour l'annotation étymologique manuelle. Dans ce cas, le seul point subjectif consistait en le fait de déterminer la frontière entre les noms d'origine slave et étrangère. En revanche, l'annotation sémantique manuelle est encore plus sujette à des biais, car elle implique davantage de catégories et que leurs frontières dépendent du parti pris théorique et de la méthodologie choisie (voir la section 5.2.4). Cependant, contrairement à la connaissance étymologique, un locuteur ordinaire d'une langue donnée peut avoir une meilleure compréhension de la distinction entre les concepts propres/communs, animé/non-animé, concret/abstrait (bien que cette dernière dichotomie soit plus complexe que les trois autres).

Dans cette section, nous chercherons à minimiser les biais présents dans l'annotation sémantique manuelle en calculant les scores sémantiques pour chacune des quatre classes (propre, humain/animé, concret, abstrait). A cet effet, nous utiliserons les méthodes de la sémantique distributionnelle et les modèles vectoriels préentrainés de RusVectōrēs disponibles pour la langue russe.

6.2.1 Analyse distributionnelle

L'hypothèse distributionnelle, émise par Harris, 1954 ; Firth, 1957 et Miller et Charles, 1991, entre autres, soutient que la proximité sémantique entre les mots se reflète dans la proximité de leur distribution. Cette hypothèse stipule que des mots qui se retrouvent dans des contextes similaires ont tendance à avoir des significations similaires. Ce principe a été concrétisé dans des modèles de sémantique distributionnelle où les mots sont représentés sous forme de vecteurs contextuels (Sahlgren, 2008 ; Lenci, 2018 ; Boleda, 2020).

Dans les premiers modèles, des modèles de comptage, le vecteur d'un mot a été estimé directement en comptant ses co-occurrences avec d'autres mots dans un corpus. Chaque dimension du vecteur représentait alors le degré d'association du mot avec chaque contexte présent dans le corpus (Fabre et Lenci, 2015 ; Bonami et Guzmán Naranjo, 2023).

Plus récemment, des modèles prédictifs ont été développés à l'aide d'outils basés sur des réseaux de neurones, tels que Word2Vec (Mikolov *et al.*, 2013) et fastText (Bojanowski *et al.*, 2017). Ces modèles sont entraînés pour prédire les mots susceptibles d'apparaître dans un contexte donné, en utilisant l'apprentissage automatique non supervisé. Ils ont reçu une attention particulière en raison de leur performance et efficacité, ainsi que de la facilité d'utilisation des réseaux de neurones artificiels entraînés sur de grands corpus pour apprendre des vecteurs distributionnels. De ce fait, les modèles vectoriels sont largement utilisés en linguistique informatique et en traitement automatique de langues et permettent d'analyser divers phénomènes linguistiques.

Cependant, cette méthode n'est pas sans défauts. Tout d'abord, les vecteurs ne représentent pas les lexèmes, mais plutôt les formes fléchies. De plus, les vecteurs

sémantiques peuvent contenir beaucoup de bruit et être indisponibles pour les éléments peu fréquents et les hapax. Bien que ces vecteurs puissent produire des résultats similaires aux intuitions sémantiques précises des locuteurs, l'interprétation des dimensions des vecteurs peut être difficile et nécessiter des analyses qualitatives pour étudier les phénomènes linguistiques. Malgré ces inconvénients, l'utilisation de vecteurs sémantiques peut fournir des informations précieuses pour la représentation sémantique lexicale. Toutefois, dans les situations où l'interprétabilité est plus importante que la précision, les modèles de comptage restent privilégiés (Boleda, 2020 ; Varvara *et al.*, 2021 ; Bonami et Guzmán Naranjo, 2023).

L'objectif de notre étude ne consiste pas à interpréter les valeurs numériques spécifiques de chaque dimension pour chaque vecteur. Ce qui nous intéresse, c'est que ces modèles permettent une analyse quantitative de la sémantique des mots. Ces modèles offrent plusieurs fonctionnalités, notamment la possibilité d'effectuer des opérations mathématiques sur les vecteurs (addition, soustraction), de trouver des mots qui sont sémantiquement proches du mot de requête, de calculer une similarité sémantique exacte entre des couples de mots, de visualiser les vecteurs de mots et leurs relations géométriques, d'obtenir le vecteur brut (tableau de valeurs réelles) pour le mot de requête, etc.

La figure (6.2) donne une représentation de vecteurs fictifs en deux dimensions associés aux termes *homme* (0.4, 0.3), *femme* (0.8, 0.1), *roi* (0.2, 0.7) et *reine* (0.9, 0.5).

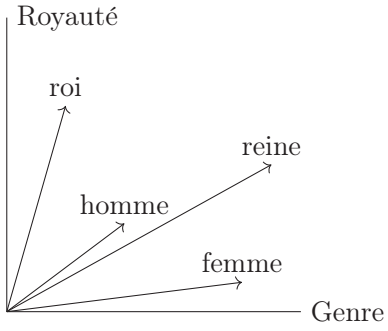

Figure 6.2: Espace vectoriel fictif

Dans le cas où seuls les trois premiers vecteurs sont disponibles, il est possible néanmoins de déterminer un vecteur candidat pour le terme *reine* grâce à des opérations mathématiques. Pour ce faire, il convient de soustraire le vecteur associé à *homme* de celui associé à *roi* et d'y ajouter le vecteur associé à *femme*. Le calcul est ainsi le suivant : $\vec{reine} = \vec{roi} - \vec{homme} + \vec{femme}$. Le vecteur prédit doit être proche dans l'espace vectoriel du vecteur réel correspondant à *reine*. Cette similitude peut être expliquée par le fait que les relations sémantiques entre *roi* et *reine* sont similaires à celles entre *homme* et *femme*.

En outre, les modèles de sémantique distributionnelle permettent de mesurer la

proximité sémantique entre deux mots représentés par leurs vecteurs en utilisant la distance cosinus entre ces vecteurs. La mesure de similarité sémantique résultante peut prendre des valeurs allant de -1 (vecteurs opposés) à 1 (vecteurs équivalents), 0 indiquant des vecteurs orthogonaux.

$$\text{distance\_cosinus}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

où :

$A\cdot$ B = le produit scalaire des vecteurs A et B,
$||A||$ = la norme (magnitude) de A (la racine carrée de la somme des carrés).

Dans l'exemple présenté dans la figure 6.2, la mesure de similarité sémantique entre les vecteurs de *roi* et *homme* est de 0.87, tandis que celle entre *roi* et *reine* est de 0.74, et celle entre *roi* et *femme* est de 0.30. Ces distances cosinus suggèrent que le mot *roi* est plus proche sémantiquement du mot *homme* que des autres mots ; il est aussi proche du mot *reine*, mais dans une moindre mesure. En revanche, il est assez éloigné du mot *femme* dans cet espace vectoriel fictif.

Pour effectuer des calculs sur les vecteurs qui correspondent aux représentations réelles des mots, il est nécessaire d'utiliser des modèles de sémantique distributionnelle entraînés sur de grands corpus de données. Cependant, l'apprentissage de tels modèles peut exiger des ressources de calcul considérables. Par conséquent, il est important de disposer de modèles pré-entraînés et prêts à télécharger pour faciliter l'analyse sémantique des mots. De tels modèles pré-entraînés pour la langue russe sont mis à disposition gratuitement sur la plateforme RusVectōrēs[11] (Kutuzov et Kuzmenko, 2017).

6.2.2 RusVectōrēs

Les modèles proposés par RusVectōrēs sont basés sur l'utilisation de différents corpus de textes, tels que le Corpus National Russe, la version russe de Wikipédia, des flux d'actualités en provenance de nombreux sites d'actualités en russe, Araneum Russicum Maximum (un corpus web de textes en russe recueillis en 2016), Taïga (un corpus ouvert et structuré de la langue russe annoté morphologiquement et syntaxiquement), GeoWAC (un échantillon de documents en russe à partir du dépôt CommonCrawl), ainsi que 9 millions de pages Web en russe sélectionnées au hasard (Kutuzov et Kuzmenko, 2017). Ces corpus sont utilisés pour entraîner les modèles distributionnels et ainsi fournir des vecteurs sémantiques pour un grand nombre de mots russes.

Le service RusVectōrēs offre une interface graphique permettant d'observer les relations géométriques entre les mots et de réaliser divers calculs sur les relations sémantiques entre les mots de la langue russe. De plus, il est possible de télécharger

[11] Accessible au https://rusvectores.org/ru/.

des modèles de distribution sémantique pré-entraînés. Nous avons choisi cette dernière option.

Pour garantir la précision des calculs, il est préférable que les modèles soient entraînés sur un corpus pré-traité : un corpus muni d'étiquettes morpho-syntaxiques qui prennent en compte les homonymes appartenant à des catégories lexicales différentes. Dans l'optique du paradigme du Web sémantique, chaque mot dans chaque modèle de RusVectōrēs possède un identifiant unique indiquant explicitement son lemme et sa partie du discours, ainsi que quelques autres propriétés morphologiques et sémantiques. Par exemple, le mot Москва (Moskva) 'Moscou' a une représentation `москва_PROPN` ; лес (les) 'forêt' correspond à `лес_NOUN`. On observe que les modèles de RusVectōrēs différencient les noms propres des noms communs. Toutefois, ils ne font pas de distinction entre les noms animés/humains et les noms non-animés/non-humains, ni entre les mots concrets et abstraits.

Pour la suite de cette étude nous avons utilisé le modèle `ruwikiruscorpora_upos_cbow_300_10_2021`[12].

Le modèle a été entraîné sur l'intégralité de RusCorpora, ainsi que sur la version de novembre 2021 du corpus Wikipédia. La taille totale du corpus final utilisé pour l'entraînement est de 1.2 milliard de mots, comprenant 249 333 lemmes.

En ce qui concerne l'architecture de ce modèle, il est construit en utilisant la méthode *Continuous Bag-of-Words* (*CBOW*). Il existe deux options d'architecture possibles pour les modèles de représentation de mots, à savoir *CBOW* et *Skip-gram*. L'architecture *CBOW* prédit le mot cible à partir du contexte, tandis que l'architecture *Skip-gram* prédit le contexte à partir du mot cible. En général, l'architecture *CBOW* est plus rapide et convient mieux aux mots fréquents, tandis que l'architecture *Skip-gram* est plus adaptée pour les mots rares et permet de capturer des nuances de sens plus précises. Le modèle *CBOW* le plus récent de RusVectōrēs est `ruwikiruscorpora_upos_cbow_300_10_2021`. Ce modèle est plus complet que le modèle *Skip-gram* le plus récent, `ruwikiruscorpora_upos_skipgram_300_2_2019` (comprenant 788 millions de mots et 248 978 lemmes). En conséquence, le modèle *CBOW* a une meilleure couverture, contenant plus de 65% des données présentes dans le corpus RuDénom, tandis que le modèle *Skip-gram* n'en contient que 55%. Le taux de couverture est la raison principale pour poursuivre notre étude avec le modèle *CBOW*.

Ce modèle utilise un seuil de fréquence de 5 pour sélectionner les mots à inclure dans le corpus d'entraînement. Le seuil de fréquence est le nombre minimum d'occurrences de chaque mot dans le corpus. Ainsi, les mots qui n'apparaissent qu'une seule fois dans le corpus (les hapax) ne sont pas inclus dans les représentations vectorielles de ce modèle. En général, les modèles plus robustes sont associés à des mots qui ont une fréquence élevée dans le corpus. Par conséquent, les modèles vectoriels ne sont pas appropriés pour l'analyse des mots à faible fréquence.

Le modèle dont il est question contient 300 vecteurs au total, une taille typique pour les modèles de sémantique distributionnelle (la taille des vecteurs se situe généralement

[12]Disponible ici : `https://rusvectores.org/ru/models/`, dernier accès : le 22/10/2022.

entre 100 et 300.). L'augmentation du nombre de vecteurs peut améliorer la précision du modèle.

Dans ce modèle, la taille de la fenêtre est de 10. La fenêtre est une zone qui encadre le mot cible et qui permet de définir le contexte de chaque occurrence. Une fenêtre plus grande peut potentiellement améliorer la précision du modèle, mais elle peut également introduire du bruit et nuire à sa performance.

Il convient de noter que le corpus a été annoté automatiquement avec des étiquettes morpho-syntaxiques de type Universal Tags.

6.2.3 Scores sémantiques

Dans le cadre de la classification traditionnelle adoptée dans la section 5.2.4, un nom peut être catégorisé comme propre, humain/animé, concret ou abstrait. Cependant, comme nous l'avons souligné dans la même section, dans certains cas il est impossible d'attribuer une seule annotation à un même nom sans tenir compte du contexte, en raison de l'homonymie ou de la polysémie.

Les scores sémantiques que nous allons présenter dans cette section sont inspirés de l'approche de Fedden *et al.* (2021), qui étudient les genres en allemand. Selon cette approche, le genre d'un nom dépend en partie de sa similarité sémantique par rapport à d'autres noms neutres, masculins et féminins. Nous estimons que cette idée peut également s'appliquer à la classe sémantique d'un nom, en prenant en compte la proximité du nom en question avec des noms des quatre classes sémantiques. Nous souhaitons introduire une perspective plus graduelle dans l'annotation classique, qui attribuerait à chaque nom un score de proximité par rapport aux noms propres, un scores de proximité par rapport aux noms humains/animés, et ainsi de suite. Toutefois, nous nous attendons à ce que les noms propres soient plus proches d'autres noms propres, les noms humains soient plus proches d'autres noms humains, etc. De plus, nous cherchons à différencier la double annotation de certains noms homonymiques ou polysémiques en utilisant les scores pour déterminer si ces noms sont plus concrets ou plus abstraits en fonction des contextes d'utilisation.

Comme indiqué dans la section 6.2.1, la sémantique distributionnelle utilise des vecteurs pour représenter les mots dans un espace multidimensionnel basé sur leur contexte. Le tableau 6.11 présente un exemple du modèle RusVectōrēs, dans lequel les vecteurs ont une dimension de 300.

Comme indiqué précédemment, les modèles de RusVectōrēs font une distinction entre les noms propres et les noms communs, mais ils ne sont pas en mesure de rendre compte de la polysémie des mots. Par exemple, le mot аудитория (auditorija) 'salle de cours/auditoire' peut désigner une entité concrète ou abstraite en fonction du contexte ; par contre, ce mot est représenté par un seul vecteur *аудитория_NOUN*.

Nous avons extrait 8 465 vecteurs bruts qui correspondent aux noms de base répertoriés dans RuDenom et qui sont également présents dans le modèle RusVectōrēs. Il convient de rappeler que le taux de couverture de ces vecteurs est supérieur à 65%

Lemme	1	2	...	299	300
москва_PROPN *moskva* ‘Moscou’	-2.103401	-3.102879	...	1.757076	0.597245
француз_NOUN *francuz* ‘français$_N$’	1.264681	0.811717	...	2.314334	2.678855
аудитория_NOUN *auditorija* ‘salle de cours/ auditoire’	-0.362413	0.419851	...	3.915248	0.436309
пегас_PROPN *pegas* ‘Pégase’	0.497675	-1.489553	...	-0.947289	1.912709

Tableau 6.11: Échantillon des vecteurs du modèle RusVectōrēs

(par rapport à un total de 12 592 entrées dans RuDénom)[13]. Nous avons ensuite utilisé ces vecteurs pour construire une matrice de distances cosinus, dont un exemple est présenté dans le tableau 6.12.

	год *god* ‘année’	*время* *vremja* ‘temps’	*район* *rajon* ‘quartier’	*город* *gorod* ‘ville’	*часть* *čast’* ‘partie’
год	1	0.205027	-0.031615	0.041173	-0.048046
время	0.205027	1	-0.070623	0.084779	0.197424
район	-0.031615	-0.070623	1	0.450161	0.325029
город	0.041173	0.084779	0.450161	1	0.228692
часть	-0.048046	0.197424	0.325029	0.228692	1

Tableau 6.12: Échantillon de la matrice de distances cosinus

Dans ce tableau, nous pouvons observer que le terme le plus proche de ВРЕМЯ (VREMJA) ‘temps’ est ГОД (GOD) ‘année’ avec une distance de 0.205027, suivi de près par ЧАСТЬ (ČAST’) ‘partie’ avec une distance de 0.197424. Pour sa part, le mot РАЙОН (RAJON) ‘quartier’ est le plus proche du mot ГОРОД (GOROD) ‘ville’ avec une distance de 0.450161, mais il présente également une similarité assez élevée avec le mot ЧАСТЬ (ČAST’) ‘partie’ avec une distance de 0.325029. Ces scores suggèrent que le mot ЧАСТЬ (ČAST’) ‘partie’ peut avoir une interprétation spatiale ou temporelle.

Pour obtenir les scores sémantiques, nous avons procédé en calculant la distance

[13]Nous avons déjà souligné dans la section 4.2 que les données utilisées pour l’étude de la concurrence (tableau 4.7) diffèrent légèrement des données réelles (tableau 4.6), en anticipant le fait que les modèles RusVectōrēs ne couvrent pas tous les noms que nous avons répertoriés pour les adjectifs dénominaux.

moyenne entre chaque mot et ses cinq voisins les plus proches (avec les distances cosinus les plus élevées) parmi les noms propres, les noms humains ou animés, les noms concrets et les noms abstraits. Nous avons ainsi obtenu quatre valeurs numériques pour les scores sémantiques correspondant à chaque mot. Le tableau 6.13 présente un exemple des voisins les plus proches pour le mot ГОРОД 'ville' ainsi que les scores sémantiques moyens correspondants pour chaque classe.

Il n'est pas surprenant que le mot russe ГОРОД (GOROD) 'ville' soit étroitement lié aux toponymes désignant les villes russes, avec une distance moyenne de 0.4154640. Cependant, étant lui-même annoté comme concret, ce mot est le plus proche des noms concrets. Ces derniers désignent principalement des divisions géographiques ou des entités territoriales, avec une distance moyenne de 0.4964450. En revanche, ГОРОД (GOROD) 'ville' présente la plus faible similarité avec des noms humains ou animés (distance moyenne de 0.2391714).

La figure 6.3 présente la distribution des scores sémantiques obtenues pour les noms de base des adjectifs de haute et de basse fréquence.

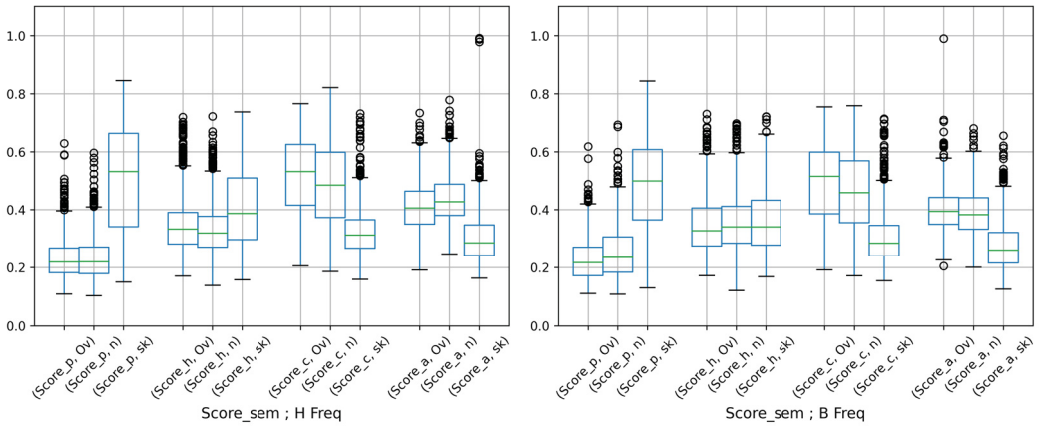

Figure 6.3: Scores sémantiques ; haute et basse fréquence

En général, les scores présentent des différences plus marquées pour le suffixe -*sk*- que pour les suffixes -*n*- et -*Ov*-. On remarque que les noms suffixés en -*sk*- sont fortement associés au score élevé `Score_p`, tandis que les noms avec un score `Score_a` élevé préfèrent les suffixes -*n*- ou -*Ov*-. Les tendances sont similaires entre les données de haute et de basse fréquence, sauf pour les noms humains/animés, où les différences sont plus marquées dans les données de haute fréquence que dans les données de basse fréquence : les scores `Score_h` pour les trois suffixes sont très proches. Ces observations sont similaires à celles que nous avons faites dans la section 5.2.4 concernant l'annotation des classes sémantiques catégorielle. Nous avons également constaté une forte corrélation entre les noms propres et le suffixe -*sk*-. Les résidus ont aussi montré que, dans les données de basse fréquence, les noms humains n'ont pas de préférence nette pour l'un des trois suffixes.

Classe	Lemme	Distance	Moyenne
p	*Тула* ‘Toula (Russie)’	0.444706	
	Tula		
	Белгород ‘Belgorod (Russie)’	0.414320	
	Belgorod		
	Дербент ‘Derbent (Russie)’	0.413031	0.4154640
	Derbent		
	Астрахань ‘Astrakhan (Russie)’	0.404934	
	Astraxan’		
	Рязань ‘Riazan (Russie)’	0.400329	
	Rjazan’		
h	*феодал* ‘féodal’	0.248236	
	feodal		
	араб ‘arabe$_{\text{N}}$’	0.242708	
	arab		
	скитник ‘gens du voyage’	0.238078	0.2391714
	skitnik		
	чухонец ‘Chukhna (peuple)’	0.237090	
	čuxonec		
	мельник ‘meunier’	0.229745	
	mel’nik		
c	*столица* ‘capitale’	0.602304	
	stolica		
	местечко ‘endroit’	0.533641	
	mestečko		
	село ‘village’	0.458518	0.4964450
	selo		
	район ‘région’	0.450161	
	rajon		
	север ‘nord’	0.437601	
	sever		
a	*община* ‘communauté’	0.321299	
	obščina		
	километр ‘kilomètre’	0.309593	
	kilometr		
	верста ‘verste (mesure)’	0.299408	0.2986140
	versta		
	объезд ‘détour’	0.289635	
	ob"ezd		
	стройка ‘chantier’	0.273135	
	strojka		

Tableau 6.13: Les voisins les plus proches de *город* ‘ville’

Les distributions observées dans les trois groupes présentent des différences significatives, comme en témoigne les résultats des tests paramétriques et non paramétriques (les statistiques F et H systématiquement supérieures à 100 dans les données de haute et de basse fréquence), et les valeurs p, qui sont systématiquement inférieures à 0.001. Cependant, dans les données de basse fréquence, on observe l'absence de différence significative entre les distributions de `Score_h` (F=3.37, p=0.03, H=3.06, p=0.22).

Nous avons procédé à une comparaison entre les scores sémantiques et l'annotation manuelle, les résultats sont présentés dans le tableau 6.14. On observe que le taux d'accord entre les deux annotations est globalement très élevé, se situant entre 77% et 97%. Cependant, on constate que les données de haute fréquence sont plus cohérentes que les données de basse fréquence, à l'exception des noms concrets, pour lesquels les taux d'accord sont quasiment identiques.

	H Freq	B Freq
max(Scores) = `Score_p` & `ClSem` = `p`	97.31%	93.07%
max(Scores) = `Score_h` & `ClSem` = `h`	92.83%	86.41%
max(Scores) = `Score_c` & `ClSem` = `c`	93.00%	93.71%
max(Scores) = `Score_a` & `ClSem` = `a`	81.13%	77.67%

Tableau 6.14: Couverture des scores sémantiques

Finalement, le tableau 6.15 présente un échantillon de noms problématiques pour l'annotation sémantique. La première partie regroupe des mots pouvant avoir des sens concrets ou abstraits selon le contexte. Le deuxième bloc contient des noms pouvant désigner des entités concrètes ou abstraites ainsi que des êtres animés. Le troisième bloc rassemble des noms concrets et abstraits, à la différence de la première partie, ce sont des noms déverbaux. Enfin, la quatrième partie regroupe des homonymes et les mots polysémiques. Les scores les plus élevés sont marqués en gras.

Comme on peut l'observer, la plupart de ces mots sont plus proches de noms abstraits. En revanche, институт (institut) 'institut', квартал (kvartal) 'quartier/trimestre', свод (svod) 'résumé/voûte', вытяжка (vytjažka) 'hotte/extraction' et смазка (smazka) 'graissage/lubrifiant' sont considérés comme des noms concrets. Les noms qui peuvent désigner à la fois des êtres animées et des entités concrètes/abstrais ne sont pas nécessairement plus proches des noms humains/animés. Au contraire, le mot штат (štat) 'état/effectif' se rapproche davantage des noms propres, ce qui peut être expliqué par la distribution conjointe de Соединённые штаты Америки (Soedinënnye štaty Ameriki) 'les États Unis [d'Amérique]'. En ce qui concerne les homonymes, seul граф (graf) 'compte/graphique' se rapproche de son sens 'compte', tandis que les autres homonymes se rapprochent plutôt de noms abstraits.

Base	Anim_p	Anim_h	Anim_c	Anim_a
аудитория ‘salle de cours/auditoire’ *auditorija*	0.845813	1.732999	1.756627	**2.185530**
блок ‘bloc/blocage’ *blok*	0.667620	1.439287	2.494616	**2.518757**
деталь ‘détail’ *detal’*	0.807151	1.457728	2.402061	**2.710020**
институт ‘institut’ *institut*	1.067260	1.922039	**3.009680**	2.925718
квартал ‘quartier/trimestre’ *kvartal*	1.673866	1.076700	**2.510101**	1.814992
класс ‘salle de cours/classe’ *klass*	0.566674	1.290439	1.698867	**2.016153**
репортаж ‘reportage’ *reportaž*	0.861449	1.807523	1.926078	**2.050283**
свод ‘code/voûte’ *svod*	1.079015	1.188704	**2.704322**	1.827372
область ‘région/domaine’ *oblast’*	1.385872	0.976949	1.839012	**2.010242**
схема ‘schéma’ *sxema*	0.807850	1.732471	2.636087	**2.968305**
охрана ‘sécurité/vigiles’ *oxrana*	1.000052	1.497832	1.599417	**2.147250**
модель ‘modèle/mannequin’ *model’*	1.113870	1.816941	2.454072	**2.654722**
штат ‘état/personnel’ *štat*	**2.787478**	1.331878	1.349829	1.399858
вытяжка ‘hotte/extraction’ *vytjažka*	0.669117	1.452017	**2.542383**	2.534699
остановка ‘arrêt’ *ostanovka*	1.267195	1.545303	2.127770	**2.148110**
смазка ‘graissage/lubrifiant’ *smazka*	0.622349	1.321646	**2.731760**	2.413598
граф ‘compte/graphique’ *graf*	2.128341	**3.117517**	1.499513	1.660943
зефир ‘guimauve/zéphyr’ *zefir*	1.196840	2.107144	2.304338	**2.353183**
мир ‘monde/paix’ *mir*	1.556206	1.419006	1.566739	**1.688171**
рак ‘écrevisse/cancer’ *rak*	1.180937	1.515575	1.948508	**2.497593**
язык ‘langue’ *jazyk*	1.168719	1.462037	2.301304	**2.614100**

Tableau 6.15: Les voisins les plus proches pour les cas ambigus

Conclusion

Dans cette section nous avons présenté deux méthodes élaborées pour quantifier la sémantique et l'étymologie des noms.

Les deux méthodes présentent des avantages et des inconvénients. Le calcul des scores étymologiques permet de se dispenser des annotations manuelles et permet de calculer ce score pour n'importe quel mot. Ainsi, même si les bigrammes ont été calculés à partir des données de Wiktionnaire, il est possible de transposer l'analyse à des mots qui ne s'y sont pas référencés. L'inconvénient est que seuls les mots du lexique étranger sont explicitement référencés dans les sources utilisées, et non pas ceux du lexique slave. Cela implique la nécessité de mettre en place des traitements supplémentaires pour approximer le lexique slave, qui, finalement, peut contenir du bruit.

En revanche, le calcul des scores sémantiques présenté dans ce chapitre ne dispense pas les chercheurs des annotations manuelles. De plus, comme les représentations vectorielles sont construites pour le lexique avec un seuil de fréquence donné, les mots de faible fréquence sont pas inclus. Seuls les noms de base qui apparaissent à la fois dans notre corpus et dans le modèle choisi peuvent être analysés. Ainsi, cette méthode conduit à l'exclusion des noms pour lesquels il n'existe pas de représentations vectorielles.

Dans les deux cas, nous avons constaté un taux élevé de concordance entre les annotations manuelles et automatiques. Toutefois, certaines différences peuvent refléter soit la structure phonologique d'un mot d'origine slave qui ne lui est pas typique (et vice-versa), soit une distribution d'un terme concret qui se rapproche de la distribution des noms propres.

Partie III

Étude de cas

Chapitre 7

Modélisation de la concurrence

Sommaire

Introduction

Dans les chapitres 5 et 6, nous avons exploré comment divers facteurs, tels que les aspects phonologiques, morphologiques, sémantiques et étymologiques, peuvent influencer le choix d'un suffixe spécifique. Bien qu'une analyse univariée ait été réalisée dans ces chapitres, il est nécessaire de déterminer comment l'importance de chaque prédicteur individuel varie lorsque ceux-ci interagissent simultanément. Ainsi, il convient d'utiliser des méthodes adéquates pour une analyse multivariée.

La section 7.1 présentera un éventail de modèles disponibles pour ce type d'analyse, en exposant les avantages et les inconvénients de chacun, ainsi que les métriques permettant de les évaluer. La section 7.2 se concentrera sur les modèles sélectionnés et leur application aux données relatives aux adjectifs de haute et basse fréquence. Dans un premier temps, nous évaluerons à quel point il est possible de faire des prédictions correctes des suffixes en fonction des propriétés des noms de base et nous réaliserons

une analyse multivariée en classant ces propriétés en fonction de leur importance prédictive. Ensuite, nous évaluerons quel est le modèle optimal pour décrire les données de haute et basse fréquence. Un modèle optimal se caractérise par sa simplicité et son adéquation. La simplicité d'un modèle réside dans le nombre réduit de prédicteurs, tandis que son adéquation se mesure par sa performance en termes de prédictions. Enfin, nous analyserons les modèles optimaux sur la base des interactions entre les prédicteurs, et accorderons une attention particulière aux erreurs éventuelles.

7.1 Méthodologie

Une méthode statistique couramment utilisée pour modéliser la relation entre différentes variables est la régression logistique. Récemment, de nouvelles méthodes basées sur les arbres, provenant du domaine de l'apprentissage automatique, ont émergé en tant qu'alternatives. Selon Baayen *et al.* (2013), la régression logistique et les arbres de décision offrent généralement des analyses convergentes avec des avantages complémentaires. Dans la section 7.1.1, nous présenterons les avantages et les inconvénients de la régression logistique et des arbres de décision, ainsi que de deux autres méthodes basées sur les arbres (forêts aléatoires et arbres boostés). Dans la section 7.1.2, nous présenterons également les métriques utilisées pour évaluer les performances de ces modèles.

Avant de continuer, il convient de définir la terminologie. Les propriétés phonologiques, morphologiques, sémantiques et étymologiques des noms qui influencent le choix d'un suffixe adjectival sont les variables explicatives ou indépendantes, également appelées prédicteurs (souvent désignées par la notation X), tandis que le suffixe lui-même représente la variable dépendante ou la réponse (notée Y). Cette variable dépendante est catégorielle car elle correspond aux trois suffixes -*n*-, -*sk*- et -*Ov*-. La plupart des variables indépendantes sont également catégorielles, et leurs distributions ont été présentées dans le chapitre 5. De plus, des variables non catégorielles ont été introduites pour la sémantique et l'étymologie, comme discuté dans le chapitre 6. Pour modéliser la concurrence entre les trois suffixes en question, il est nécessaire de choisir un classifieur, c'est-à-dire un modèle qui, en tenant compte des variables indépendantes, tente de prédire la variable dépendante. Comme mentionné précédemment, nous limiterons la discussion à quatre modèles : la régression logistique, les arbres de décision, les forêts aléatoires et les arbres boostés.

7.1.1 Choix du modèle

7.1.1.1 Méthodes statistiques

Comme le remarquent Tagliamonte et Baayen (2012), la pertinence des méthodes statistiques n'est plus discutable pour les études linguistiques, la question se pose surtout sur le choix des méthodes appropriées. L'outil standard est le modèle de regression logistique, mais il existe également des outils plus récents et novateurs

comme les algorithmes basés sur les arbres qui permettent une analyse de données non paramétrique (sans hypothèse sur la distribution réelle à partir de laquelle un échantillon a été prélevé pour une étude).

La régression logistique est une méthode statistique couramment utilisée lorsque la variable dépendante est binaire[1]. Il s'agit d'un modèle de régression appartenant à une famille de techniques appelées modélisation linéaire généralisée (Baayen, 2008 ; Hilpert et Blasi, 2021). L'objectif de la régression logistique est d'expliquer la variable dépendante en fonction de différents prédicteurs ; en d'autres termes, de trouver une relation entre les variables indépendantes et la probabilité d'un résultat particulier. Si la variable dépendante est binaire (par exemple, si l'on étudie seulement la concurrence entre -*n*- et -*sk*-), la probabilité d'un résultat particulier peut correspondre à la probabilité de choisir -*n*- (succès), étant donné les propriétés des noms de base (-*sk*- est alors considéré comme l'échec).

La régression logistique est basée sur différents calculs statistiques.

Le rapport de chances correspond au ratio entre la probabilité de succès (-*n*-) et la probabilité d'échec (-*sk*-), étant donné une variable prédictrice X :

$$\frac{P(Y = n|X)}{1 - P(Y = n|X)}$$

Le rapport de chances est exprimé en termes de 'n fois plus de chances' (ou 'n fois moins de chances'), par exemple, 'étant donné X, le suffixe -*n*- a 5 fois plus de chances d'être choisi que -*sk*-'. Il peut donc prendre des valeurs dans l'intervalle [0, ∞], ce qui rend les résultats quelque peu difficiles à interpréter lorsqu'il s'agit d'une comparaison entre différentes valeurs. Grace à la fonction *logit*, le rapport de chances est alors transformé en logarithme pour faciliter les calculs et l'interprétation :

$$\log\left(\frac{P(Y = n|X)}{1 - P(Y = n|X)}\right)$$

La valeur du logarithme est négative dans l'intervalle [-∞,0] lorsque le nombre d'observations pour -*sk*- est supérieur au nombre d'observations pour -*n*-. Elle est nulle lorsque les comptes sont égaux. Elle est positive dans l'intervalle [0,∞] lorsque les comptes pour -*n*- dépassent les comptes pour -*sk*- (Gries, 2013, pp.319-328). Si les signes positif ou négatif sont plus simples à interpréter comparé au rapport de chances, ces valeurs doivent être ajustées davantage pour correspondre à la probabilité d'un événement : un nombre réel compris entre 0 et 1. 0 correspondrait ainsi à ce que l'utilisation de -*n*- est complètement improbable ; 1, à son tour, indiquerait qu'il est complètement certain que -*n*- sera utilisé. Cela est obtenu grâce à la fonction logistique inverse, qui transforme les valeurs de logarithme en probabilités :

[1]La régression logistique peut également servir à prédire plus de deux catégories de la variable dépendante, ce qui correspond précisément à l'objectif de ce chapitre : modéliser la concurrence entre trois suffixes. Pour des raisons de clarté, nous allons présenter les calculs associés à la régression logistique binaire dans cette section.

$$P(Y = n|X) = \frac{e^{\beta X+\alpha}}{1 + e^{\beta X+\alpha}}$$

où :

$\beta X + \alpha$ = la combinaison pondérée des prédicteurs[2]

Grâce à des calculs statistiques précis, la régression logistique permettrait ainsi de voir quelles variables prédictrices sont les plus fortement corrélées avec l'un des suffixes adjectivaux, la force de cette corrélation ainsi que les préférences particulières. Un exemple de synthèse des résultats de régression logistique est présenté dans la figure 7.1[3]. Dans cet exemple, le but du modèle est de prédire si un locuteur utilisera *will* ou *be going to* sur la base de quatre variables indépendantes : l'âge (`periodlate`), le genre (`gendermal`), la sémantique du verbe (`semanticsnonag`) et la formalité du texte (`formalityfor` et `formalityinf`). Il est à noter que les autres valeurs d'âge, genre, sémantique du verbe et la formalité du texte sont choisis comme niveau de référence (intercept).

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.04637    0.12511   0.371    0.711
periodlate    -0.19773    0.08760  -2.257    0.024 *
gendermal      0.14426    0.09711   1.485    0.137
semanticsnonag 0.36329    0.08715   4.168 3.07e-05 ***
formalityfor   1.41616    0.16108   8.791  < 2e-16 ***
formalityinf  -0.91277    0.09573  -9.535  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 7.1: Exemple d'un tableau des coefficients de la régression logistique

Tout d'abord, nous pouvons interpréter la significativité des variables indépendantes (à l'aide de la valeur p, mais aussi plus explicitement avec les étoiles qui accompagnent ces valeurs) et la préférence de chaque prédicteur pour une des deux réponses à (l'aide des signes positifs ou négatifs des coefficients). Notamment, la période est significative (`*`) et présente un signe négatif, ce qui signifie que l'utilisation de *be going to* est plus fréquente lorsqu'il s'agit de `periodlate`, comparé à `periodearly` qui est dans l'intercept. Le coefficient pour le genre n'est pas significatif (valeur p élevée, absence d'étoiles), ce qui indique que les hommes et les femmes sont indiscernables dans leur utilisation de la marque de futur. Le coefficient pour la sémantique du verbe montre que les verbes non-agentifs biaisent significativement (`***`) les locuteurs vers *will*. La formalité est représentée par deux lignes. Les textes formels montrent un biais vers

[2]Cette combinaison pondérée des prédicteurs correspond à l'équation de la droite de régression. Nous reviendrons sur cette équation dans le chapitre 8.

[3]L'exemple est tiré de Hilpert et Blasi (2021, p.514) ; nous nous concentrons uniquement sur le tableau des coefficients.

will (valeur positive), tandis que les textes informels montrent un biais vers *be going to* (valeur négative).

L'interprétation ultérieure concerne les chiffres eux-mêmes derrière chaque coefficient. Par exemple, dans les contextes avec des verbes non-agentifs, le coefficient 0.36 correspond au logarithme du rapport de chances. La fonction exponentielle permet de convertir ces valeurs en rapports de chances, qui peuvent être interprétés plus facilement. Dans le cas des verbes non-agentifs, la valeur 0.36, transformée par la fonction exponentielle, donne un rapport de chances de 1.433. Cela signifie que les chances pour les verbes non-agentifs d'être utilisés avec *will* sont 43.3% plus élevées que les chances pour les verbes agentifs.

Nous pouvons ainsi constater que l'utilisation de la régression logistique requiert des compétences statistiques pour interpréter le résultat du modèle. De plus, comme il s'agit d'une méthode paramétrique, la possibilité même de son utilisation repose sur la satisfaction de plusieurs hypothèses (Hilpert et Blasi, 2021, p.515), dont les principales concernent l'indépendance des prédicteurs : chaque variable prédictrice doit être indépendante des autres variables contenues dans le même jeu de données. Autrement dit, il ne doit y avoir de possibilité de prédire les valeurs d'une variable prédictrice à partir d'une ou de plusieurs autres variables. La régression logistique est ainsi limitée lorsqu'elle est appliquée à des situations complexes, où les données incluent les variables colinéaires (Levshina, 2021, p.611).

Contrairement aux données expérimentales bien équilibrées, les données de corpus sont souvent caractérisées par des prédicteurs collinéaires ; l'hypothèse d'indépendance des prédicteurs n'est pas toujours satisfaite, ce qui entraîne des coefficients de régression instables (Gries, 2019, p.617). Cette instabilité peut rendre complexe l'évaluation de la direction et de la force d'un effet[4]. De plus, la régression logistique a tendance à sous-estimer ou surestimer l'importance des variables qui ont une faible fréquence (Guzmán Naranjo et Bonami, 2021).

En somme, comme le soulignent Baayen *et al.* (2013), l'emploi de la régression logistique implique une vérification préalable des hypothèses pour la mise en place de ce modèle paramétrique et une expertise solide en statistiques pour l'interprétation des résultats.

Dans ce contexte, les méthodes d'apprentissage automatique deviennent particulièrement intéressantes pour résoudre ces problématiques. Ces méthodes sont non-paramétriques et sont ainsi susceptibles de fournir des solutions alternatives pour traiter efficacement les défis posés par les données de corpus et les prédicteurs corrélés.

7.1.1.2 Méthodes basées sur les arbres

Une alternative particulièrement populaire qui apparaît de plus en plus dans les études est la famille de méthodes basées sur les arbres, comprenant en particulier les

[4]Dans l'analyse de la concurrence entre les suffixes *-n-* et *-sk-*, nous démontrons que les phonèmes finaux des radicaux sont fortement corrélés à la présence d'une consonne palatalisée dans l'espace thématique des noms (Bobkova, 2022b).

arbres de classification (ou de décision), les forêts aléatoires (Breiman, 2001) ou les arbres boostés (Friedman, 2001).

Les méthodes basées sur les arbres fonctionnent en divisant à plusieurs reprises des ensembles de données en deux parties de manière à ce que la division conduise à une plus forte augmentation de précision de classification. Cette famille d'approches peut être considérée comme une alternative non paramétrique à la régression, puisqu'elle est potentiellement beaucoup moins affectée par la colinéarité des données qui rendent les modèles de régression si difficiles (Gries, 2019, p.618).

Un exemple d'arbre de décision est présenté dans la figure 7.2[5]. L'objectif de la classification est de prédire si une proposition principale précédera une proposition subordonnée (`mc-sc`) ou si elle la suivra (`sc-mc`). `Conj` fait référence à la conjonction qui introduit la proposition subordonnée (*weil/because*, *bevor/before*, *als/when* ou *nachdem/after*) ; `SubOrdType` représente le type de subordonnée (causal ou temporel) ; `LenghtDiff` correspond à la différence de longueur entre la proposition principale et la proposition subordonnée (en mots).

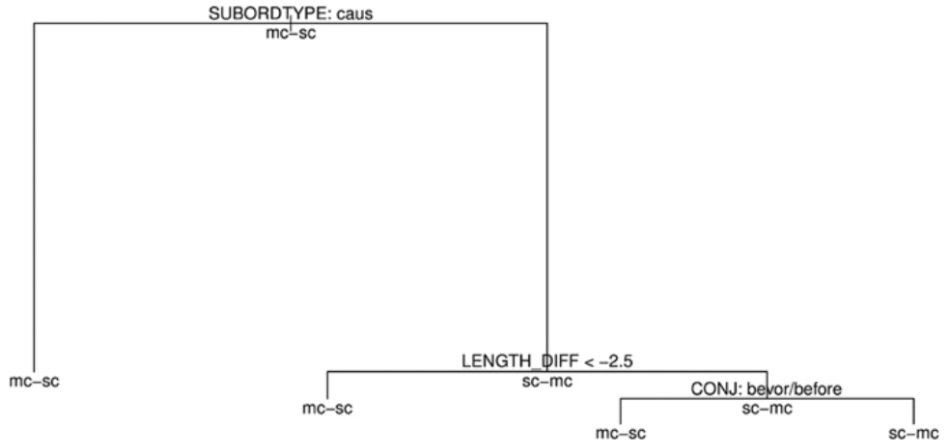

Figure 7.2: Exemple d'un arbre de décision

Cet arbre peut être lu de manière suivante : en partant du haut, si le type de proposition subordonnée est causal, aller à gauche et prédire `mc-sc` ; si le type de proposition subordonnée est temporel, alors aller à droite et vérifier la différence de longueur entre la proposition principale et la proposition subordonnée : si cette différence est moins de -2.5, aller à gauche et prédire également `mc-sc`, sinon aller à droite et vérifier si la conjonction est *bevor/before*. Si c'est le cas, prédire à nouveau `mc-sc`, si ce n'est pas le cas, prédire `sc-mc`. Le même arbre peut être résumé de manière plus concise : toujours prédire `mc-sc`, sauf si la différence de longueur entre la proposition principale et la proposition subordonnée est supérieure à -2.5 et que les conjonctions sont *als/when* ou *nachdem/after*.

[5]La figure et l'interprétation qui suit sont tirées de Gries (2001, p.619).

Dans le contexte des arbres de décision, la classification est généralement plus facile à interpréter que le tableau de synthèse et les coefficients rapportés pour les modèles de régression grâce à l'intuitivité que présentent les arbres (Baayen *et al.*, 2013). Néanmoins, Gries (2019, pp.620-622) avertit que, lorsqu'il s'agit d'un grand arbre, le synthétiser en texte peut s'avérer complexe et peu intuitif, en raison de plusieurs facteurs. Premièrement, chaque nœud suivant de l'arbre implique une condition 'si' supplémentaire. Cela rend les résumés plus difficiles à appréhender avec l'augmentation de la profondeur de l'arbre. Ainsi, l'intuitivité des arbres peut être remise en question. Deuxièmement, la division binaire inhérente à ces méthodes peut compliquer l'interprétation des effets des prédicteurs non catégoriels (la figure 7.2 n'en contient qu'un seul). Contrairement à un modèle de régression qui renvoie une pente significative pour un prédicteur non catégoriel, un arbre de décision peut utiliser plusieurs divisions binaires pour représenter cette pente, ce qui nécessite parfois de combiner plusieurs divisions à différents endroits de l'arbre.

Les arbres de décision sont puissants car ils sont capables de modéliser des relations qui sont irrégulières et nuancées. Cependant, cette caractéristique les rend sujets au surentraînement, qui se produit lorsqu'un modèle est trop complexe et s'adapte trop étroitement aux données d'entraînement. Cela peut entraîner de mauvaises performances du modèle lorsqu'il est évalué sur de nouvelles données qu'il n'a jamais vues auparavant (Plonsky *et al.*, 2015, p.599). De plus, les arbres de décision ne sont pas toujours particulièrement stables ou robustes : Gries (2019, p.622) souligne que même de petites variations dans les valeurs des prédicteurs peuvent entraîner de grands changements dans les prédictions.

Pour éviter le surentraînement des arbres de décision, il est possible de recourir à des forêts aléatoires (Breiman, 2001). Les forêts aléatoires sont une extension des méthodes basées sur les arbres qui sont de plus en plus utilisées dans les études linguistiques actuelles. Il s'agit d'une méthode d'apprentissage automatique qui combine plusieurs arbres de décision pour réaliser une prédiction finale (cf. la figure 7.3[6]).

Il s'agit d'un algorithme non paramétrique, ce qui signifie qu'il ne fait aucune hypothèse concernant la distribution des données ou la relation entre les caractéristiques et la variable cible.

Les forêts aléatoires entraînent de nombreux arbres sur des sous-ensembles des données choisies au hasard. Les prédictions finales sont obtenues en collectant les informations des arbres individuels, avec les résultats de leurs apprentissages respectifs. Les forêts aléatoires offrent généralement des prédictions plus précises que les arbres de classification standards (Baayen *et al.*, 2013, p.265).

Baayen *et al.* (2013, pp.254-266) citent plusieurs avantages des forêts aléatoires, en plus de leurs performances élevées : elles offrent un moyen d'évaluer l'importance relative des différents prédicteurs dans le modèle ; cette méthode est capable de détecter des interactions complexes qui échappent aux modèles logistiques ; de plus, les forêts (tout comme les arbres de décision simples) sont bien adaptés aux données avec des

[6]Illustration tirée de Khan *et al.* (2021).

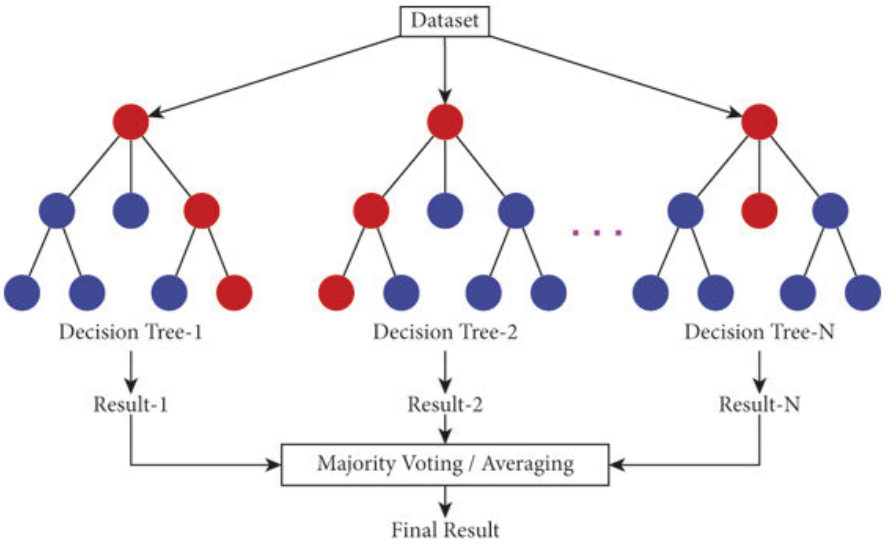

Figure 7.3: Illustration d'une forêt aléatoire

variables indépendantes qui contiennent peu de niveaux.

Cependant, lorsque le nombre de niveaux des prédicteurs devient important, les forêts deviennent très complexes, ce qui entraîne un temps de calcul important. De plus, le modèle peut devenir également computationnellement intense avec de grands ensembles de données et/ou un grand nombre d'arbres de décision (Baayen *et al.*, 2013, pp.254-266). Un autre problème que présentent les forêts est la multitude d'hyperparamètres à régler, par exemple, le nombre d'arbres dans la forêt, le nombre maximal de prédicteurs à prendre en compte à chaque division, la profondeur de chaque arbre, le nombre minimal de données autorisées après la division, etc. Trouver les hyperparamètres optimaux peut être chronophage et nécessiter des expérimentations minutieuses.

Néanmoins, l'inconvénient principal des forêts aléatoires pour les études linguistiques est que, comparées aux modèles de régression et aux arbres de décision simples, les forêts sont difficiles à interpréter. Ainsi, il n'y a pas d'arbre unique qui puisse être représentatif des données utilisées. Cependant, comme le suggèrent Baayen *et al.* (2013, pp.254-266), il est possible d'approximer le fonctionnement d'une forêt aléatoire développée en utilisant un arbre de décision simple entraîné sur le même ensemble de données.

La dernière méthode présentée ici est le boosting, une technique d'apprentissage regroupant plusieurs algorithmes de classifieurs binaires (par exemples, les arbres) pour en améliorer les performances. Le boosting apprend progressivement en combinant, étape par étape, plusieurs arbres (généralement plus faibles et moins performants) afin

de créer un prédicteur plus robuste et précis (cf. la figure 7.4[7]).

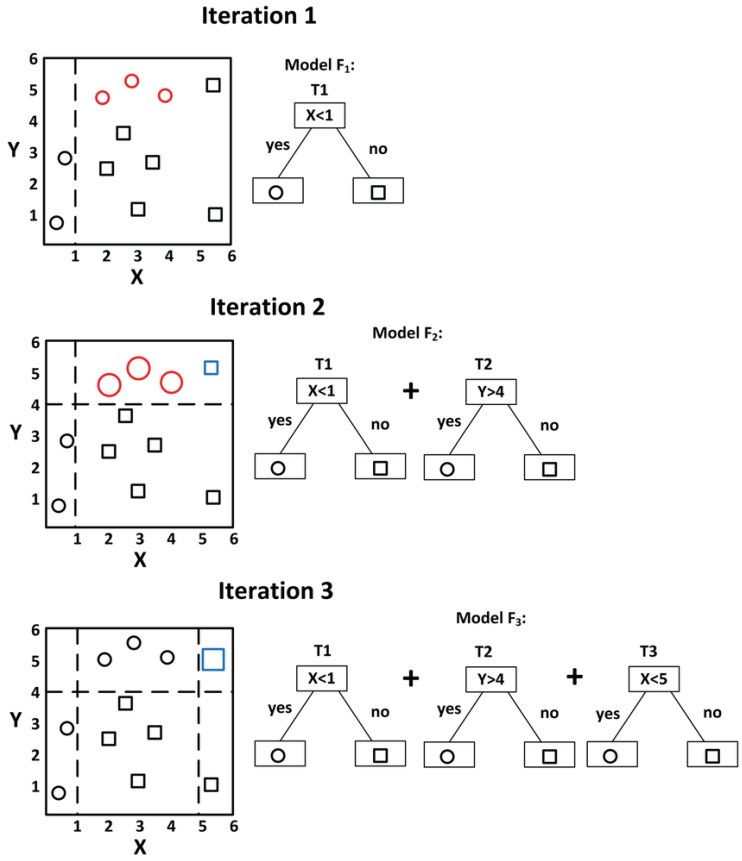

Figure 7.4: Illustration des arbres boostés

Cette approche ressemble aux forêts aléatoires, mais construit une forêt d'arbres de manière séquentielle : chaque nouvel arbre de décision dépend des arbres précédents, tient compte de leurs résultats de classification et cherche à corriger leurs erreurs (James *et al.*, 2013, pp.345-348).

Les approches telles que les arbres boostés apprenant progressivement ont tendance à être performantes et efficaces pour les ensembles de données volumineux et complexes, grâce à l'utilisation de diverses méthodes d'optimisation. Contrairement à la régression logistique, elles ne nécessitent pas l'indépendance et la non-corrélation des prédicteurs, sont plus efficaces dans la gestion de données catégorielles avec de nombreux niveaux et ne sont pas impactées si certains prédicteurs sont faiblement représentés (Guzmán Naranjo et Bonami, 2021). Elles sont aussi moins sujettes au surentraînement que les arbres de décision simples et plus rapides en temps de calcul que les forêts aléatoires.

[7]Illustration tirée de Zhang *et al.* (2018).

Toutefois, comme les forêts aléatoires, les arbres boostés possèdent plusieurs hyperparamètres qui doivent être réglés, tels que le nombre d'arbres, la profondeur de chaque arbre, les paramètres de régularisation, etc. La recherche des hyperparamètres optimaux peut prendre du temps et nécessiter une expérimentation exhaustive. De plus, tout comme les forêts aléatoires, les arbres boostées ne sont pas facilement interprétables. Leur fonctionnement peut néanmoins être approximé, de la même manière que pour les forêts aléatoires, en entraînant un arbre de décision simple sur l'ensemble de données.

Compte tenu des avantages et des inconvénients des quatre méthodes examinées dans cette section, nous retiendrons deux approches : les arbres boostés et les arbres de décision simples. Pour les arbres boostés, nous utiliserons l'implémentation `XGBoost` (Chen et Guestrin, 2016) en Python[8]. En ce qui concerne les arbres de décision simples, nous emploierons le package `rpart` (Therneau *et al.*, 2015) en R.

Les arbres boostés serviront pour évaluer dans quelle mesure nous pouvons prédire le suffixe adjectival étant donné les propriétés des noms de base, pour mettre en place une analyse multivariée et évaluer l'importance de chaque prédicteur pour le choix du suffixe adjectival. Ils nous permettront aussi d'identifier les modèles optimaux pour les données de haute et de basse fréquence, c'est-à-dire les plus performants avec un nombre minimal de prédicteurs et d'analyser les erreurs. Les arbres de décision simples, quant à eux, faciliteront la compréhension du modèle optimal et mettront en lumière l'interaction entre les prédicteurs.

Avant de procéder à la modélisation, il convient d'aborder les métriques, en particulier celles utilisées pour évaluer la performance globale des modèles optimaux, pour détecter les erreurs et pour estimer le pouvoir explicatif des propriétés des noms.

7.1.2 Évaluation et interprétation des modèles

L'évaluation des modèles d'apprentissage automatique constitue une étape cruciale pour examiner leur performance et leur capacité à généraliser face à de nouvelles données. Cette évaluation requiert généralement la séparation des données en deux ensembles distincts : un ensemble d'entraînement, destiné à l'ajustement des paramètres afin d'atteindre les meilleures performances possibles, et un ensemble de test, permettant de mesurer la capacité du modèle à s'étendre à de nouvelles données.

7.1.2.1 Validation croisée

La validation consiste à employer un ensemble de données de test distinct de celui utilisé lors de l'apprentissage du modèle. Après avoir entraîné le modèle sur l'ensemble

[8]Les paramètres de modèles ont été optimisés avec la recherche en grille sur un échantillon de données. Cette technique permet d'énumérer toutes les combinaisons possibles d'hyperparamètres à tester, puis à évaluer la performance de chaque combinaison sur un ensemble de données d'entraînement et de validation : avec l'objectif `multi:softprob` pour la classification multi-classe entre 3 suffixes ; la profondeur maximale de l'arbre est de 10 ; la métrique d'évaluation est AUC ; le taux d'apprentissage est de 0.1.

de données d'apprentissage, celui-ci est appliqué à l'ensemble de données de test pour estimer la précision des prédictions. Une forme spécifique de validation, offrant un équilibre optimal entre la qualité des résultats et les coûts de calcul, est la validation croisée à 10 passes (Gries, 2013, p.385-387).

Cette approche consiste à répartir aléatoirement les données en dix sous-ensembles, chacun représentant 10% de l'ensemble total. Le modèle est ensuite entraîné en utilisant neuf de ces sous-ensembles et testé sur le groupe restant. Ce processus est répété 10 fois en employant différentes combinaisons de sous-ensembles.

La validation croisée à 10 passes est une technique plus élaborée que la simple division des données en ensembles d'apprentissage et de test. Elle permet généralement de construire un modèle moins biaisé, en garantissant que toutes les observations de l'ensemble de données initial aient la même probabilité d'être présentes à la fois dans l'ensemble d'apprentissage et dans l'ensemble de test. Ainsi, chaque observation est utilisée pour l'évaluation du modèle, permettant d'obtenir une estimation plus précise de la performance du modèle. Nous procéderons à l'évaluation des modèles en utilisant cette technique.

7.1.2.2 Métriques

Pour évaluer la performance d'un modèle, il est courant d'utiliser une matrice de confusion. Cette matrice confronte les prédictions du classifieur (en colonnes) aux valeurs réelles observées (en lignes). Un exemple hypothétique de matrice de confusion est présenté dans le tableau 7.1.

		Prédictions		
		-n-	*-sk-*	*-Ov-*
Observations	*-n-*	146	3	9
	-sk-	4	284	3
	-Ov-	12	4	112

Tableau 7.1: Matrice de confusion fictive

Le tableau 7.1 expose les résultats d'une classification fictive pour les suffixes étudiés à travers une matrice de confusion. Cette dernière révèle que, parmi les 158 noms de base réellement associés au suffixe *-n-*, 146 ont été correctement classés, tandis que 3 ont été incorrectement classés avec le suffixe *-sk-* et 9 avec le suffixe *-Ov-*. La diagonale de la matrice de confusion représente les vrais positifs. Une meilleure performance du modèle se traduit par des valeurs élevées sur la diagonale et des valeurs proches de zéro dans les autres cases de la matrice.

Étant donné que nous utilisons une validation croisée à 10 passes, nous présenterons une matrice de confusion agrégée, combinant les matrices de confusion calculées pour chacun des 10 passes en une matrice de confusion globale reflétant les performances du modèle sur l'ensemble des données. Nous nous baserons sur la matrice de confusion

pour détecter notamment les cas d'erreurs.

La précision globale est une métrique permettant d'évaluer la performance d'un modèle en une seule valeur. Elle est calculée en divisant le nombre de prédictions correctes (la somme des valeurs sur la diagonale) par le nombre total d'éléments. Dans notre exemple, la précision est de $542/577 = 0.94$.

En outre, puisque nous avons recours à la validation croisée, les précisions rapportées correspondront aux valeurs moyennes obtenues à partir des résultats des 10 classifieurs lors de la validation croisée.

Pour une comparaison statistiquement significative, nous rapporterons un intervalle de confiance de 95% pour la valeur de précision. L'intervalle de confiance donne une plage de valeurs dans laquelle la précision est susceptible de se trouver, avec une probabilité de 95%[9]. Il dépend de la taille de l'échantillon et la distribution des données : une taille d'échantillon plus grande ou une variabilité plus faible entraînera un intervalle de confiance plus étroit, et vice versa. La précision accompagnée d'un intervalle de confiance nous permettra de détecter le modèle optimal.

7.1.2.3 Interprétation

Les modèles d'arbres boostés offrent des méthodes internes permettant d'évaluer l'importance relative des prédicteurs. Toutefois, interpréter cette évaluation de manière absolue demeure complexe. En effet, même si une variable est considérée comme relativement plus importante qu'une autre, cela ne fournit pas nécessairement des informations significatives quant à sa capacité prédictive absolue réelle[10].

Afin de résoudre le problème de relativité des prédicteurs, Fedden *et al.* (2021) se basent sur la précision des classifieurs, qui sont susceptibles de fournir un indice du pouvoir explicatif des variables indépendantes. La première méthode proposée par ces auteurs, méthode additive[11], consiste à introduire une variable dans le modèle. Si l'introduction de cette variable améliore la précision du modèle, cela signifie que cette variable possède un pouvoir explicatif pertinent pour la classification. Un exemple fictif de calcul du pouvoir explicatif selon la méthode additive (`PEA`) est présenté dans le tableau 7.2.

Modèles	Précision	Prédicteur	PEA
Tous les paramètres	0.9525	-	-
Tous sauf structure phonologique	0.9115	Structure phonologique	0.0410
Tous sauf classes flexionnelles	0.9347	Classes flexionnelles	0.0178
Tous sauf allomorphies vocaliques	0.9513	Allomorphies vocaliques	0.0012

Tableau 7.2: Exemple fictif du pouvoir explicatif ; méthode additive

[9] Nous calculerons un intervalle de confiance de 95% en utilisant la distribution *t*.

[10] Nous remercions Matías Guzmán Naranjo pour ses observations et ses commentaires à ce sujet.

[11] *Additive variable importance* (Fedden *et al.*, 2021)

Selon cette méthode, le pouvoir explicatif représente la différence entre la précision du modèle incluant tous les paramètres (n) et la précision du modèle excluant le paramètre à évaluer (n - 1). Par exemple, pour évaluer l'importance de la structure phonologique, on soustrait la précision du modèle sans cette variable (0.9115) de la précision du modèle incluant tous les paramètres (0.9525). Le pouvoir explicatif de la structure phonologique est donc de 0.0410, ce qui est également le plus élevé par rapport aux autres variables dans cet exemple fictif. Cependant, Fedden *et al.* (2021) soulignent que cette méthode ne prend pas en compte les interactions potentielles entre les prédicteurs. Par conséquent, il est possible que de nombreux noms, correctement traités par les modèles grâce à leur structure phonologique, pourraient également être bien classifiés grâce aux autres variables, comme les classes flexionnelles ou les allomorphies vocaliques, dans cet exemple fictif.

La deuxième méthode proposée par ces auteurs est une méthode soustractive[12]. Cette approche vise à éliminer toute redondance et à évaluer si un prédicteur apporte une valeur ajoutée au modèle, une valeur que les autres variables ne peuvent pas fournir. Après avoir isolé les noms qui sont mal catégorisés par les modèles avec un ensemble de prédicteurs sauf un (n - 1), on évalue le nombre de noms correctement classifiés parmi ces erreurs par chaque prédicteur restant. Les résultats de calcul fictif du pouvoir explicatif des variables selon la méthode soustractive (`PES`) sont présentés dans le tableau 7.3.

Modèle n - 1	Précision	Erreurs	Predicteur	Correct	PES
Classes flex	0.9347	397	Classes flex	266	0.6700
Structure phono	0.9115	416	Structure phono	212	0.5096
Allom voc	0.9513	374	Allom voc	172	0.4599

Tableau 7.3: Exemple fictif du pouvoir explicatif ; méthode soustractive

Pour estimer le pouvoir explicatif de la structure phonologique, nous commençons d'abord par évaluer les performances du modèle contenant toutes les variables à l'exception de ce paramètre. La précision du modèle sans structure syllabique est de 0.9115, générant 416 cas d'erreurs (les noms pour lesquels ce modèle prédit un suffixe erroné). Nous évaluons ensuite le modèle qui ne contient que la structure syllabique comme prédicteur sur ces cas erronés. Ce modèle classifie correctement 212 noms parmi les erreurs, le pouvoir explicatif de la structure syllabique correspond donc à 0.5096 (la précision de ce modèle). Par ailleurs, nous observons que le pouvoir explicatif des variables a changé. Si la méthode additive a fait ressortir la structure phonologique comme une propriété ayant le plus d'impact sur le choix du suffixe, la méthode soustractive a mis en valeur les classes sémantiques.

Dans la suite de notre étude, nous allons nous appuyer sur la méthodologie soustractive pour estimer le pouvoir explicatif des propriétés des noms de base dans l'analyse multivariée.

[12] *Subtractive variable importance* (Fedden *et al.*, 2021).

7.2 Modèles unifiés pour -*n*-, -*sk*- et -*ov*-

Dans cette section, nous aborderons la compétition entre les trois suffixes étudiés en examinant les données de haute et de basse fréquence.

La section 7.2.1 aura un double objectif. Nous commencerons par évaluer les performances des modèles intégrant toutes les propriétés des noms pour prédire l'un des trois suffixes. Ensuite, nous examinerons le pouvoir explicatif des propriétés des noms de base et analyserons l'influence de chaque variable sur la compétition entre les suffixes, en considérant simultanément les autres facteurs. Dans la section 7.2.2, nous chercherons à identifier les modèles les plus performants avec un nombre minimal de prédicteurs, capables de décrire les données de haute et de basse fréquence. Comme nous avons opté pour des arbres boostés, qui ne sont pas facilement interprétables, nous utiliserons des arbres de décision simples dans la section 7.2.3 pour mieux comprendre ces modèles optimaux. Enfin, la section 7.2.4 se focalisera sur l'évaluation des performances des modèles en examinant les matrices de confusion pour les modèles optimaux et en analysant les situations engendrant le plus d'erreurs.

7.2.1 Analyse multivariée

Comme nous l'avons constaté dans le chapitre 5, les variables indépendantes évaluées une par une ne sont par corrélées de la même manière avec les suffixes -*n*-, -*sk*- et -*Ov*-. Dans l'analyse multivariée, les variables prédictrices ont aussi rarement une pertinence équivalente pour la réponse. Très souvent, seules quelques-unes d'entre elles influencent substantiellement la variable dépendante. La grande majorité sont sans importance et peuvent ainsi être éliminées. Nous allons donc estimer l'importance de chaque variable lorsque tous les paramètres sont pris en compte dans une analyse multivariée.

7.2.1.1 Haute fréquence

La figure 7.4 illustre la précision de deux modèles. Le modèle `Tous les paramètres` qui combine toutes les propriétés des noms de base présente une précision de 0.8805. Les intervalles de confiance pour ce modèle sont très étroits. `Baseline`, c'est-à-dire le modèle de référence ou modèle naïf, correspond au modèle qui prédit systématiquement la classe la plus fréquente sans prendre en considération les caractéristiques des données. Dans le cas des données de haute fréquence, la baseline est équivalente à 0.4779, correspondant à la proportion des noms qui se combinent avec le suffixe -*n*- (les proportions pour -*sk*- et -*Ov*- étant respectivement de 0.3099 et 0.2122). Ainsi, le modèle combinant tous les paramètres est presque deux fois plus performant que le modèle naïf et affiche une très bonne précision.

Avant de procéder à l'analyse du pouvoir explicatif des propriétés des noms de base, il est à noter que certaines variables prédictrices sont redondantes. Ainsi, nous avons introduit deux variables supplémentaires non catégorielles pour la sémantique et l'étymologie (cf. chapitre 6). De plus, nous avons utilisé deux conventions pour

Modèles	Précision	Int. de confiance	
Tous les paramètres	0.8805	0.8784	0.8826
Baseline	0.4779	-	-

Tableau 7.4: Précision des modèles ; haute fréquence

encoder la position de l'accent (cf. section 5.2.2) et les classes flexionnelles (cf. section 5.2.3)[13]. La méthodologie soustractive que nous avons adoptée pour expliquer le pouvoir explicatif des variables requiert l'analyse des précisions des modèles du type `n - 1`. Cependant, l'application de cette méthodologie impliquerait, par exemple, l'évaluation de la sémantique catégorielle lorsque la sémantique non catégorielle est déjà prise en compte, et inversement, ce qui peut fausser les résultats. Pour résoudre ce problème, nous allons exclure deux variables au lieu d'une pour la sémantique, l'étymologie, la position de l'accent et les classes flexionnelles, et ensuite nous allons évaluer à quel point chaque variable parmi les deux restantes gère les cas d'erreurs.

Le tableau 7.5 présente les résultats triés par le pouvoir explicatif des propriétés des noms de base.

Modèle n - 1	Précision	Erreurs	Predicteur	Correct	PES
Sémantique	0.6551	860	Score_sem	730	0.8488
Sémantique	0.6551	860	ClSem	698	0.8116
SyllN	0.8332	416	SyllN	212	0.5096
Étymologie	0.8785	303	Score_etym	152	0.5017
DPhoR	0.8761	309	DPhoR	152	0.4919
Classes flexionnelles	0.8785	303	ClFlexZal	147	0.4851
Accent	0.8801	299	AccZal	134	0.4480
Classes flexionnelles	0.8785	303	ClFlex	132	0.4356
Genre	0.8765	308	Genre	134	0.4351
Accent	0.8801	299	AccSyllN	123	0.4114
Étymologie	0.8785	303	SourceN	122	0.4026
AllomC	0.8765	308	AllomC	122	0.3961
AllomV	0.8801	299	AllomV	115	0.3846

Tableau 7.5: Pouvoir explicatif des propriétés des noms de base ; haute fréquence

On remarque que la précision des modèles sans un paramètre est très élevée (plus de 0.8%), à l'exception du modèle qui exclut la sémantique (seulement 0.6551, avec

[13]Puisque les propriétés sémantiques sont les plus fortement corrélées au suffixe adjectival (cf. l'analyse univariée dans les sections 5.2.4 et 6.2.3), nous nous intéressons particulièrement à la comparaison du pouvoir explicatif entre ces deux variables. L'étymologie est modérément corrélée au choix du suffixe ; comme nous avons introduit une méthodologie distincte pour une annotation non catégorielle des propriétés étymologiques, nous nous intéressons aussi à évaluer l'apport que l'étymologie non catégorielle puisse avoir par rapport à l'étymologie catégorielle.

860 cas d'erreurs). Il n'est pas surprenant que les propriétés sémantiques aient le pouvoir explicatif le plus élevé dans l'analyse multivariée. Cependant, nous pouvons constater que c'est la sémantique non catégorielle qui est plus pertinente pour le choix des suffixes (730 prédictions correctes sur 860) par rapport à la sémantique catégorielle (698 noms correctement prédits sur 860), qui reste néanmoins très performante.

En ce qui concerne l'étymologie, on observe de nouveau un pouvoir explicatif plus élevé pour l'étymologie non catégorielle, comparée à l'étymologie catégorielle. L'écart est toutefois beaucoup plus marqué (le pouvoir explicatif est de 0.5017 contre 0.4026, respectivement). De plus, nous pouvons constater que la typologie de Zaliznjak est plus pertinente à la fois pour les classes flexionnelles et pour la position de l'accent (les mêmes tendances ont été également observées dans l'analyse univariée).

En général, les propriétés sémantiques, la structure syllabique, l'étymologie non catégorielle et les derniers phonèmes des radicaux sont les plus pertinents pour le choix du suffixe dans les données de haute fréquence. Ces résultats sont aussi globalement en accord avec l'analyse univariée effectuée dans la section 5.2[14] : nous retrouvons les mêmes propriétés, l'ordre changeant légèrement. Toutefois, l'étymologie fait exception. Dans l'analyse univariée, l'étymologie catégorielle était assez haut placée dans le classement des prédicteurs ; dans l'analyse multivariée, elle a cédé la place à l'étymologie non catégorielle.

7.2.1.2 Basse fréquence

Le tableau 7.6 représente les précisions des mêmes types de modèles dans les données de basse fréquence.

Modèles	Précision	Int. de confiance	
Tous les paramètres	0.8206	0.8146	0.8265
Baseline	0.4438	-	-

Tableau 7.6: Précision des modèles ; basse fréquence

D'abord, comparées aux données de haute fréquence, la précision du modèle intégrant tous les paramètres, soit 0.8206, est plus faible, ce qui pourrait être attribué à une plus grande variabilité des données. De plus, comme nous l'avons démontré dans le chapitre 5, les tailles d'effet et les valeurs des résidus sont également moindres pour les données de basse fréquence. De surcroît, l'intervalle de confiance est légèrement plus large ici, ce qui confirme l'idée que les différences entre les propriétés des noms de base et leurs préférences pour les suffixes sont moins marquées dans les données de basse fréquence. La baseline, qui s'établit à 0.4438, correspond à la proportion de données avec le suffixe -*sk*- (les proportions pour -*n*- et -*Ov*- étant respectivement de 0.3549 et 0.2013). Bien que les performances soient inférieures comparées à celles observées avec

[14] Cf. le tableau récapitulatif 5.17.

les données de haute fréquence, le modèle qui contient tous les paramètres demeure néanmoins près de deux fois plus performant que le modèle naïf.

Le tableau 7.7 répertorie les pouvoirs prédictifs des propriétés des noms de base dans les données de basse fréquence.

Modèle n - 1	Précision	Erreurs	Predicteur	Correct	PES
Sémantique	0.6467	437	Score_sem	354	0.8101
Sémantique	0.6467	437	ClSem	340	0.7780
DPhoR	0.8069	239	DPhoR	139	0.5815
AllomC	0.8230	219	AllomC	124	0.5662
Genre	0.8270	214	Genre	107	0.5000
SyllN	0.8028	244	SyllN	121	0.4965
AllomV	0.8270	214	AllomV	106	0.4942
Étymologie	0.8173	266	Score_etym	111	0.4905
Classes flexionnelles	0.8246	217	ClFlex	106	0.4885
Classes flexionnelles	0.8246	217	ClFlexZal	106	0.4885
Étymologie	0.8173	266	SourceN	116	0.4361
Accent	0.8165	277	AccSyllN	115	0.4152
Accent	0.8165	277	AccZal	115	0.4152

Tableau 7.7: Pouvoir explicatif des propriétés des noms de base ; basse fréquence

Les scores attribués au pouvoir explicatif des variables sont à nouveau plus faibles comparés à ceux obtenus avec les données de haute fréquence. Toutefois, le classement des variables est généralement en accord avec l'analyse univariée[15] : ainsi, nous retrouvons en haut de la liste les propriétés sémantiques et les derniers phonèmes des radicaux. Cependant, bien que l'étymologie ait occupé une place importante dans l'analyse univariée, il semble que son pouvoir explicatif diminue lorsqu'elle est en interaction avec d'autres paramètres. De plus, l'analyse multivariée a permis de faire ressortir l'importance des allomorphies consonantiques ainsi que des genres.

Comme dans les données de haute fréquence, la sémantique présente le pouvoir explicatif le plus élevé, et c'est encore la sémantique non catégorielle qui s'avère la plus pertinente (son pouvoir explicatif est de 0.8101, contre 0.7780 pour la sémantique catégorielle). De même, bien que l'étymologie soit moins importante dans les données de basse fréquence, les scores étymologiques sont plus pertinents pour le choix du suffixe que l'étymologie envisagée comme une propriété binaire. Contrairement aux données de haute fréquence, les deux conventions relatives aux classes flexionnelles et à la position de l'accent présentent le même pouvoir explicatif.

[15] Voir le tableau récapitulatif 5.17.

7.2.2 Modèles optimaux

Dans la section précédente, nous avons estimé le pouvoir explicatif des propriétés des noms de base pour le choix du suffixe dans l'analyse multivariée. Nous avons démontré que toutes les propriétés des noms de base n'ont pas le même impact sur le choix du suffixe. Nous avons aussi constaté que les modèles qui intègrent soit la totalité de prédicteurs (n), soit tous les prédicteurs sauf un (n - 1) sont très performants, mais leurs performances sont assez proches. Dans cette section notre objectif sera de retrouver les modèles qui contiennent le minimum de prédicteurs mais qui ne sont pas significativement moins performants que les modèles qui affichent de meilleurs scores.

Dans les analyses complexes, la sélection de modèles vise à déterminer les variables indépendantes appropriées et leurs interactions pour être incluses dans un modèle. Deux méthodes peuvent être utilisées à ce propos (Gries, 2013, p.366). La sélection progressive commence avec un modèle minimal, ajoutant et conservant les variables indépendantes uniquement si elles améliorent le modèle. À l'inverse, la sélection régressive par étapes commence avec un modèle maximal, incluant toutes les variables indépendantes potentiellement significatives, et supprime de manière itérative les prédicteurs les moins nécessaires jusqu'à ce qu'aucune amélioration supplémentaire ne soit obtenue. En fin de compte, des modèles plus simples sont adoptés dans la sélection régressive s'ils ne sont pas significativement moins performants que des alternatives plus complexes.

Nous allons utiliser une méthode de sélection régressive. Afin de trouver le modèle optimal, nous utiliserons la précision globale de chaque modèle et les intervalles de confiance. Un modèle peut être considéré comme significativement meilleur ou significativement moins bon que l'autre si leurs intervalles de confiance ne se chevauchent pas (Bonami et Pellegrini, 2022). Ainsi, le modèle optimal serait celui qui contient le minimum de prédicteurs et dont la précision se situe dans l'intervalle de confiance du modèle le plus performant.

7.2.2.1 Haute fréquence

Dans les données de haute fréquence analysées, le meilleur modèle est celui qui intègre l'ensemble des paramètres. Ce modèle présente une précision moyenne de 0.8805, avec un intervalle de confiance situé entre 0.8784 et 0.8826. Ainsi, la borne supérieure de l'intervalle de confiance du modèle optimal doit inclure 0.8784, qui correspond à la borne inférieure de l'intervalle de confiance du meilleur modèle.

Le modèle optimal pour les données de haute fréquence est celui qui combine les prédicteurs suivants : les scores sémantiques, la longueur des noms de base en syllabes et les derniers phonèmes des radicaux. Malgré le fait que l'étymologie non catégorielle présente un pouvoir explicatif relativement élevé (cf la section 7.2.1), ce paramètre n'est pas intégré au modèle optimal. Ce modèle affiche une précision moyenne de 0.8773, avec un intervalle de confiance compris entre 0.8758 et 0.8787. Bien que la précision de ce modèle soit inférieure à celle du meilleur modèle, les intervalles de

confiance se chevauchent, il n'est donc pas possible d'affirmer que sa performance est significativement plus faible.

7.2.2.2 Basse fréquence

Dans la continuité de l'analyse précédente, nous examinerons également les performances des modèles pour les données de basse fréquence. Une différence notable par rapport aux données de haute fréquence est que le modèle englobant l'ensemble des prédicteurs n'est pas le plus performant. Le modèle le plus performant est celui qui intègre l'ensemble des paramètres, à l'exception de la classe flexionnelle. Ce modèle présente une précision moyenne de 0.8279, avec un intervalle de confiance situé entre 0.8211 et 0.8345.

Contrairement aux données de haute fréquence, il faut avoir quatre prédicteurs pour montrer les même performances que le meilleur modèle : à côté des mêmes propriétés qu'on retrouve dans le modèle optimal pour les données de haute fréquence s'ajoute la présence des allomorphies consonantiques dans l'espace thématique des noms. Malgré le fait que les genres ont eu un score assez élevé du pouvoir explicatif (cf la section 7.2.1), cette variable ne figure pas dans le modèle optimal. Ce modèle affiche une précision moyenne de 0.8189, avec un intervalle de confiance compris entre 0.8147 et 0.8230.

Dans la suite de cette étude, nous procéderons à l'implémentation de modèles basés sur des arbres de décision simples, afin d'approfondir notre compréhension des interactions entre les propriétés des noms de base au sein de chacun des deux modèles optimaux identifiés.

7.2.3 Arbres de décision

7.2.3.1 Haute fréquence

Comme précisé dans la section 7.1.1, nous utilisons `rpart` (Therneau *et al.*, 2015) pour visualiser les arbres de décision simples et pour mieux comprendre l'intéraction des propriétés des noms de base dans le modèle optimal. Les graphiques de `rpart` sont légèrement différents que ceux utilisés en exemple (figure 7.2). La figure 7.5 présente un arbre de décision pour les données de haute fréquence. Chaque nœud de cet arbre représente une question, dont la réponse 'non' se trouve toujours sur la branche droite. Chaque feuille de l'arbre est caractérisée par trois informations : la décision associée à la feuille (l'un des trois suffixes), la probabilité prédite pour chaque suffixe et le pourcentage d'observations dans chaque feuille. Par exemple, la feuille la plus à droite permet de classer 25% de tous les noms avec le suffixe -*sk*- et ce suffixe concerne 92% de ces données, 5% et 3% de noms se combinent avec -*n*- et -*Ov*-, respectivement.

L'arbre de décision a été élaboré à partir d'un ensemble de données de 2 494 observations ; les variables clé de cet arbre de décision sont trois scores sémantiques sur quatre (le score pour les noms propres, pour les noms humains et pour les noms

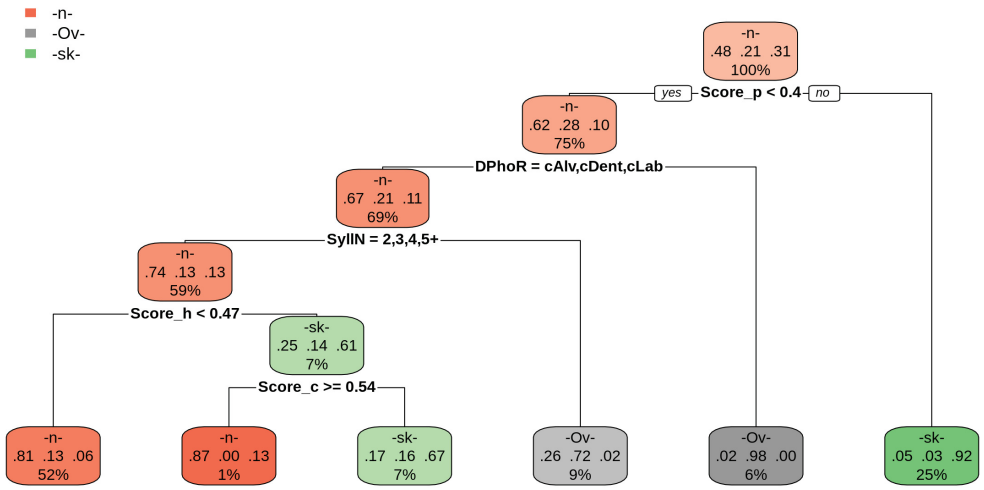

Figure 7.5: Arbre de décision ; haute fréquence

communs), ainsi que les derniers phonèmes des radicaux et la longueur des noms de base en syllabes.

L'arbre comporte plusieurs nœuds. Le nœud racine segmente les données selon le score sémantique des noms propres, envoyant 75% des observations à gauche et 25% à droite. La majorité des observations dans la partie droite appartient à la classe -*sk*-, dont la probabilité est de 0.9217, comme nous l'avons mentionné plus haut. Ainsi, cet arbre suggère que lorsque le score sémantique pour les noms propres est supérieur à 0.4, on peut prédire le suffixe -*sk*-. Ainsi, il n'est pas surprenant que 25% des données aient été classées dans la feuille la plus à droite de l'arbre. Cette feuille contient 626 noms, dont 469 ont été annotés manuellement comme appartenant à la classe des noms propres, et 577 se combinent effectivement avec le suffixe -*sk*- (92.17%).

Cette répartition des données concorde avec l'analyse univariée des résidus réalisée dans la section 5.2.4 : les noms propres et, dans une moindre mesure, les noms humains avaient tendance à privilégier le suffixe -*sk*-. Dans l'analyse multivariée, le poids des noms propres devient plus important, et l'isolation de la majorité des données avec le suffixe -*sk*- repose uniquement sur ce critère sémantique (nous reviendrons sur le poids des noms humains plus bas dans cette section).

Nous avons examiné le premier nœud permettant de différencier les suffixes, en particulier -*sk*-, sur la base du score sémantique des noms propres. Les deux nœuds suivants poursuivent la segmentation des données en fonction des derniers phonèmes des radicaux et de la longueur des noms de base, facilitant ainsi l'identification du parcours pour prédire le suffixe -*Ov*-. Dans le premier cas, le suffixe -*Ov*- est prédit lorsque le score sémantique des noms propres est inférieur à 0.4 et que le dernier

phonème des radicaux est une consonne vélaire ou une voyelle. Cela concerne 151 noms de base, concrets et abstraits, dont 148 (98.01%) se combinent effectivement avec -*Ov*-. L'analyse univariée a révélé que ce sont principalement les consonnes vélaires qui présentent une forte préférence pour -*Ov*- ; l'analyse multivariée met en lumière une interaction entre les consonnes vélaires et le score sémantique.

Dans le deuxième cas, le suffixe -*Ov*- est prédit lorsque le score sémantique des noms propres est toujours inférieur à 0.4, mais si le radical se termine par une consonne alvéolaire, dentale ou labiale, alors le nom ne doit pas comporter plus de 2 syllabes. 9% des données sont concernées, parmi lesquelles figurent 233 noms de base, répartis de manière uniforme entre les noms concrets et abstraits, dont 168 se combinent réellement avec -*Ov*- (72.10%). L'analyse univariée a également mis en évidence la préférence des noms monosyllabiques pour -*Ov*-. L'arbre de décision permet de révéler d'autres propriétés majoritairement présentes chez ces noms, notamment en termes de dernier phonème des radicaux et de score sémantique.

La segmentation dans la partie gauche de l'arbre révèle que le score sémantique élevé des noms propres n'est pas le seul facteur influençant la prédiction du suffixe -*sk*-. Pour prédire ce suffixe, le score des noms propres peut également être inférieur à 0.4, à condition que les critères suivants soient remplis : le dernier phonème des radicaux doit être une consonne alvéolaire, dentale ou labiale ; le nom doit comporter au moins deux syllabes ; le score des noms humains doit être supérieur à 0.47 et le score des noms concrets doit être inférieur à 0.54. L'idée qu'il existe des noms avec le score humain élevé et le score concret bas est peu intuitive ; une analyse plus approfondie des données est donc nécessaire. Dans ce cas, il s'agit des noms désignant des êtres humains, lesquels font référence majoritairement à des professions (ДИЗАЙНЕР (DIZAJNER) 'designer', МЕНЕДЖЕР (MENEDŽER) 'manager', РЕКТОР (REKTOR) 'recteur'), mais aussi à des ethniques (БАШКИР (BAŠKIR) 'Bachkir', ЦЫГАН (CYGAN) 'Tzigane') ou à des caractéristiques individuelles des personnes (ДИЛЕТАНТ (DILETANT) 'dilettante', ЩЁГОЛЬ (ŠČËGOL') 'dandy'). Il est important de souligner que le `Score_h` est systématiquement supérieur au `Score_c`, ce qui suggère que les noms en question peuvent être considérés comme majoritairement humains.

Ce parcours, assez complexe, concerne 7% des données (164 noms), dont 110 se combinent réellement avec le suffixe -*sk*- (67.07%). Cependant, comme mentionné précédemment, le score sémantique des noms humains (ou la classe sémantique des noms humains) ne suffisent en soi pour prédire le suffixe -*sk*-, contrairement aux noms propres. L'interaction avec d'autres critères, notamment phonologiques, est nécessaire pour y parvenir.

Dans tous les autres cas, le suffixe -*n*- est prédit. En résumé, le fait qu'un nom soit un nom propre est suffisant en soi pour isoler une grande partie de données avec le suffixe -*sk*-. La discrimination entre -*n*- et -*Ov*- est moins évidente, la distinction suivante entre -*n*- et -*sk*- est aussi complexe. Contrairement à -*Ov*- les noms plutôt communs doivent être polysyllabiques et avoir une consonne alvéolaire, dentale ou labiale finale pour se combiner avec -*n*- ou -*sk*-. De plus, entre ces noms, les noms avec

le score humain bas priviligient -*n*-. S'il s'agit des noms avec le score humain élevé, il faut alors qu'ils soient dotés d'un score concret élevé pour pouvoir se combiner avec -*n*-. De manière similaire au cas du suffixe -*sk*- mentionné précédemment, une meilleure compréhension du concept des noms à la fois humains et concrets est nécessaire. L'analyse des données révèle qu'il s'agit principalement de noms collectifs (АРМИЯ (ARMIJA) 'armée', ПЕХОТА (PEXOTA) 'infanterie', СТАДО (STADO) 'troupeau'), dont le `Score_c` est effectivement supérieur à celui des noms humains qui désignent des personnes et des animaux individuellement (qui se combinent avec -*sk*-). Il est également important de noter que, dans le cas des noms qui préfèrent -*n*-, le `Score_c` est plus élevé que le `Score_h`, ce qui indique qu'il s'agit davantage de noms concrets que de noms humains.

7.2.3.2 Basse fréquence

Étant donné que le modèle optimal pour les données de basse fréquence contient quatre prédicteurs, l'arbre de décision (figure 7.6) entraîné présente plus de complexité que celui qui décrit les intéractions dans les données de haute fréquence.

Tout d'abord, un score élevé pour les noms propres (>0.4) n'est plus suffisant pour isoler les noms qui se combinent avec le suffixe -*sk*-. L'évaluation du score sémantique pour les noms concrets est nécessaire : si ce score est bas, le suffixe -*sk*- est prédit, sinon, le modèle prédit le suffixe -*n*-. Malgré la complexité accrue, la prédiction du suffixe -*sk*- couvre davantage de cas par rapport aux données de haute fréquence (36% contre 25%) et présente une précision supérieure (96% contre 92%). Les 1% de données classées avec le suffixe -*n*-, marquées par des scores élevés (`Score_p` et `Score_c`), concernent quelques noms géographiques (ГОЛЬФСТРИМ (GOL'FSTRIM) 'Gulf Stream', ТРУБЕЖ (TRUBEŽ) 'Trubež'), mais représentent majoritairement des noms communs dont la distribution syntaxique est similaire à celle des toponymes (АВЕНЮ (AVENJU) 'avenue', ОКЕАН (OKEAN) 'océan', КОСМОС (KOSMOS) 'cosmos', КОНТИНЕНТ (KONTINENT) 'continent'), ce qui pourrait expliquer le `Score_p` relativement élevé.

Comme pour les données de haute fréquence, les derniers phonèmes des radicaux qui correspondent à une consonne vélaire ou une voyelle permettent de prédire le suffixe -*Ov*- dans la grande majorité des cas. Toutefois, si le score sémantique des noms propres est élevé (entre 0.36 et 0.40), c'est le suffixe -*sk*- qui est prédit. Contrairement aux toponymes classés avec le suffixe -*sk*- mentionnés précédemment, les 1% de données combinées avec le suffixe -*sk*- ici sont les noms propres des personnes (ВЕКСЕЛЬБЕРГ (VEKSEL'BERG) 'Vekselberg', МАЗДАК (MAZDAK) 'Mazdak'). De plus, c'est le seul nœud dans les arbres de haute et de basse fréquence permettant une distinction directe entre les suffixes -*sk*- et -*Ov*-.

La présence d'allomorphies consonantiques, un facteur supplémentaire par rapport aux données de haute fréquence, ne sert qu'à discriminer le suffixe -*n*- : 10% des données sont concernées (la feuille la plus à gauche). Lorsque les allomorphies ne sont pas présentes, la classification ultérieure se fait toujours entre les trois suffixes concurrents.

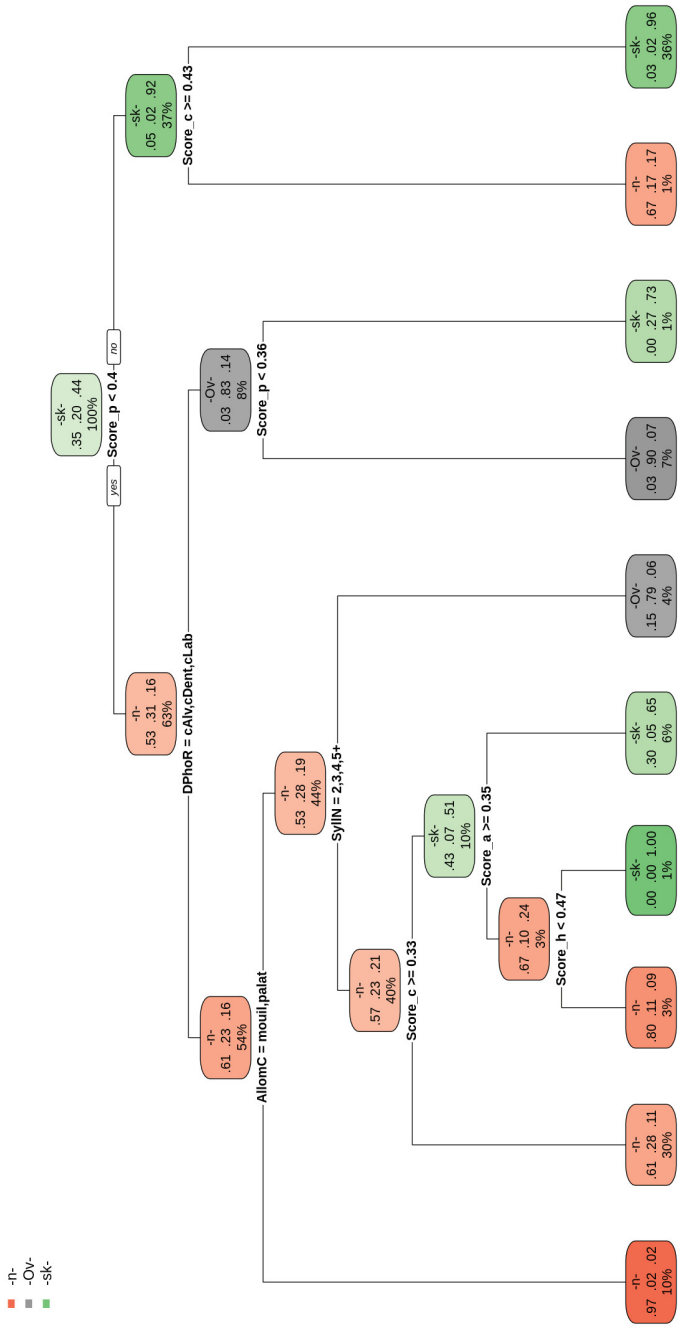

Figure 7.6: Arbre de décision ; basse fréquence

De manière similaire aux données de haute fréquence, le suffixe -*Ov*- est également prédit lorsque le score des noms propres est bas (<0.4), le dernier phonème des radicaux est une dentale, alvéolaire ou labiale, et lorsque le nom de base est monosyllabique. 4% des données sont concernées. Pour résumer, le suffixe -*Ov*- peut apparaître avec n'importe quel type de phonème final. S'il s'agit de vélaires, ce sont principalement des noms propres de personnes, indépendamment de leur structure syllabique ; s'il s'agit d'alvéolaires, dentales ou labiales, ce sont principalement des noms monosyllabiques, sans indication particulière sur leur classe sémantique.

La division à des niveaux plus profonds (pour les noms polysyllabiques non propres dont le radical se termine par une alvéolaire, dentale ou labiale) se fait principalement entre les suffixes -*n*- et -*sk*- sur la base de trois scores sémantiques. Un facteur supplémentaire, le `Score_a`, est ajouté, contrairement aux données de haute fréquence. Le suffixe -*sk*- est principalement prédit lorsque les scores `Score_c` et `Score_a` sont bas (<0.33 et <0.35 respectivement), ce qui concerne 6% des données. Si le `Score_a` est élevé, alors le `Score_h` doit également être élevé (>0.47) pour pouvoir prédire le suffixe -*sk*-. Les 1% de données concernées désignent effectivement des humains (ЧУДАК (ČUDAK) 'original$_N$', БЮРОКРАТ (BJUROKRAT) 'bureaucrate') ou des métiers (АРБИТР (ARBITR) 'arbitre', БОДИБИЛДЕР (BODIBILDER) 'culturiste'). Dans les autres cas, lorsque les scores des noms concrets et abstraits sont élevés (>0.33 et >0.35 respectivement) et lorsque le score des noms humains est bas (<0.47), c'est le suffixe -*n*- qui est prédit. Toutefois, la majorité des prédictions pour le suffixe -*n*- concerne uniquement les noms concrets (30% des données), mais la précision est relativement basse (61%).

7.2.4 Analyse des erreurs

Nous avons passé en revue les propriétés des noms de base qui favorisent l'un des trois suffixes adjectivaux. Cependant, les modèles visant à discriminer ces suffixes dans les données de haute et basse fréquence ne sont pas parfaits, pour rappel, leurs précisions sont de 0.8773 et de 0.8189, respectivement. Dans cette section, nous nous concentrerons sur l'analyse des erreurs, et notamment sur les propriétés des noms qui compliquent la tâche des classifieurs.

7.2.4.1 Haute fréquence

Dans la matrice de confusion 7.8, on observe la répartition des prédictions et des observations pour les suffixes -*n*-, -*sk*- et -*Ov*- dans le contexte des données de haute fréquence.

Le modèle optimal semble mieux classer les suffixes -*n*- et -*sk*- par rapport au suffixe -*Ov*-, comme le montre le nombre plus élevé d'observations correctement classées pour ces deux catégories. Les erreurs de classification les plus fréquentes concernent les cas où le suffixe -*Ov*- est incorrectement classé comme -*n*- (180 cas).

		Prédictions		
		-n-	*-sk-*	*-Ov-*
Observations	*-n-*	1 082	53	57
	-sk-	74	692	7
	-Ov-	180	30	319

Tableau 7.8: Matrice de confusion ; haute fréquence

Concernant les erreurs de prédiction du suffixe *-Ov-*, celles-ci concernent principalement les noms monosyllabiques.

Sur les 7 noms qui se combinent réellement avec *-sk-*, 5 sont concernés (182).

(182) Лувр ‘Louvre’
LUVR
плоть ‘chair’
PLOT’
скит ‘ermitage’
SKIT
спас ‘sauveur’
SPAS
флот ‘flotte’
FLOT

Lorsque le suffixe réellement attesté est *-n-*, 52 noms monosyllabiques sur 57 erreurs ont été mal classifiés avec *-Ov-* (183), pour en citer quelques-uns.

(183) бар ‘bar’
BAR
взвод ‘peloton’
VZVOD
глаз ‘œil’
GLAZ
гнев ‘colère’
GNEV
знать ‘noblesse’
ZNAT’
кровь ‘sang’
KROV’
лес ‘forêt’
LEC
медь ‘cuivre’
MED’
сон ‘sommeil’

SON
СЫР ‘fromage’
SYR
ТРОН ‘trône’
TRON

Il est à noter que les consonnes vélaires qui favorisent le suffixe -*Ov*- sont très peu représentées (184). Cela suggère que c'est principalement la structure phonologique inhabituelle des noms pour la distribution des suffixes -*n*- et, dans une moindre mesure, du suffixe -*sk*- qui influence le classifieur en faveur de -*Ov*-.

(184) ВЫПУСК ‘émission’
VYPUSK
РОЗЫСК ‘recherche’
ROZYSK

En ce qui concerne les erreurs de prédiction du suffixe -*sk*-, la variation est plus importante.

Dans le cas du suffixe -*n*- observé, le suffixe -*sk*- est incorrectement prédit lorsque le nom de base est polysyllabique et plutôt humain ou propre, dans 34 cas sur 53 (185).

(185) АНОНИМ ‘anonyme’
ANONIM
ГЛАВА ‘chef’
GLAVA
КАНАРЕЙКА ‘canari’
KANAREJKA
Луна ‘Lune’
Luna
НАРОД ‘peuple’
NAROD
СУМАСБРОД ‘excentrique’
SUMASBROD

Dans le cas du suffixe réel -*Ov*-, les erreurs de prédiction de -*sk*- concernent les noms bi- et trisyllabiques qui désignent principalement des animaux ou sont des noms propres, dans 17 cas sur 30 (186).

(186) ДЕЛЬФИН ‘dauphin’
DEL'FIN
ЖУРАВЛЬ ‘grue’
ŽURAVL'
ЛОСОСЬ ‘saumon’
LOSOS'

мамонт ‘mammouth’
mamont
Марс ‘Mars’
Mars
осётр ‘esturgeon’
osëtr
Рыжик ‘Ryžik’
Ryžik
страус ‘autruche’
straus

Dans ces deux situations, les scores sémantiques élevés pour les noms propres et les noms humains sont fortement associés à -*sk*-, il n’est donc pas surprenant que le modèle commette des erreurs.

Pour les prédictions erronées avec -*n*-, la variation est encore plus grande.

Lorsque le suffixe réellement attesté est -*Ov*-, le suffixe -*n*- est mal attribué dans 169 cas sur 180 lorsque le nom de base est bi- ou trisyllabique et que le dernier phonème des radicaux correspond à une consonne dentale, alvéolaire ou labiale. Ces noms ont également un score `Score_c` élevé et désignent principalement des plantes (187a), des tissus (187b), des matières (187c), des éléments et des substances chimiques (187d), des minéraux (187e).

(187) a. вишня ‘cerise’
višnja
липа ‘tilleul’
lipa
осина ‘tremble’
osina
перец ‘poivre’
perec
сандал ‘santal’
sandal
черешня ‘griotte’
čerešnja
эвкалипт ‘eucalyptus’
èvkalipt

b. вельвет ‘velours’
vel’vet
дерматин ‘simili cuir’
dermantin
капрон ‘nylon’
kapron
марля ‘gaze’

MARLJA
ПАРЧА ‘brocart’
PARČA
САТИН ‘satin’
SATIN

c. ПЕНОПЛАСТ ‘mousse de polystyrène’
PENOPLAST
ТЕФЛОН ‘téflon’
TEFLON
ФАЯНС ‘faïence’
FAJANS
ЭМАЛЬ ‘émail’
ÈMAL’

d. ГЕРОИН ‘héroïne’
GEROIN
ИНСУЛИН ‘insuline’
INSULIN
КАРБОН ‘carbone’
KARBON
МЕТАН ‘méthane’
METAN
ТРОТИЛЛ ‘trinitrotoluène’
TROTILL

e. АСБЕСТ ‘amiante’
ASBEST
БИРЮЗА ‘turquoise’
BIRJUZA
КРЕМЕНЬ ‘silex’
KREMEN’
ОПАЛ ‘opale’
OPAL
ЯШМА ‘jaspe’
JAŠMA

En ce qui concerne le suffixe réel *-sk-*, *-n-* est prédit dans 54 cas sur 74 erreurs lorsqu'il s'agit de noms dont le score abstrait est élevé (188a). Pour huit mois référencés dans les données de haute fréquence, ceux-ci sont également prédits avec le suffixe *-n-* au lieu de *-sk-* (188b).

(188) a. ЗЕМСТВО ‘zemstvo’
ZEMSTVO
МЕДИЦИНА ‘médecine’
MEDICINA

мезозой ‘mésozoïque’
MEZOZOJ
шариат ‘charia’
ŠARIAT

b. декабрь ‘décembre’
DEKABR’
июль ‘juillet’
IJUL’
май ‘mai’
MAJ
ноябрь ‘novembre’
NOJABR’
октябрь ‘octobre’
OKTJABR’
сентябрь ‘septembre’
SENTJABR’
февраль ‘février’
FEVRAL’
январь ‘janvier’
JANVAR’

7.2.4.2 Basse fréquence

La matrice de confusion pour les données de basse fréquence (figure 7.9) est similaire à celle des données de haute fréquence : la majorité des erreurs se concentre sur le suffixe *-Ov-* et le classement des noms concernés avec le suffixe *-n-* (104 cas).

		Prédictions		
		-n-	*-sk-*	*-Ov-*
Observations	*-n-*	378	40	21
	-sk-	46	496	7
	-Ov-	104	25	120

Tableau 7.9: Matrice de confusion ; haute fréquence

Le suffixe *-sk-* est prédit lorsque les noms qui se combinent réellement avec *-n-* ont majoritairement un score abstrait assez bas (23 noms concernés sur 40) : il s’agit des noms propres désignant des organisations (189a), des personnages (189b) ou des toponymes (189c) et des noms humains désignant les métiers (189d).

(189) a. Антанта ‘Entente’
ANTANTA
Любэ ‘Ljubè (groupe musical)’
LJUBÈ

Рено ‘Renault’
Reno

b. Валькирья ‘Valkyrie’
Val’kir’ja
Кастанеда ‘Castaneda’
Kastaneda
Чебурашка ‘Tchebourachka’
Čeburaška

c. Монблан ‘Mont-Blanc’
Monblan

d. коневод ‘éleveur de chevaux’
konevod
следопыт ‘éclaireur’
sledopyt
судмедэксперт ‘médecin légiste’
sudmedèkspert

Les noms de ce type correspondent en effet à ceux qui se combinent généralement avec *-sk-*, d’autant plus que антантский (antantskij), монбланский (monblanskij), коневодский (konevodskij), валькирьский (val’kir’skij), следопытский (sledopytskij) et судмедэкспертский (sudmedèkspertskij) ne violent pas les règles phonologiques et peuvent être considérés comme des mots potentiels non attestés dans le corpus (contrairement à ?любэский (ljubèskij), ?реноский (renoskij) et ?чебурашкский (čeburašskij) qui sont difficilement acceptables). Toutefois, dans RusCorpora, ce sont les adjectifs en *-n-* qui sont répertoriés (190)[16]. On remarque que ces formations sont propres aux styles poétique et littéraire du début du XXe siècle.

(190) a. *Но, поднатужась до пота, Всё же, к двадцатым годам, Даже* ***антантным*** *дредноутам Кепка дала по шеям!* [Н. Я. Агнивцев. Кепка (1923-1926)]
No, podnatužas’ do pota, Vsë že, k dvadcatym godam, Daže ***antantnym*** *drednoutam Kepka dala po šejam!*
‘Mais, en faisant un effort jusqu’à transpirer, dans les années ’20, la Casquette a donné du fil à retordre même aux cuirassés **de l’Entente** !’ [N. Ja. Agnivcev. La Casquette (1923-1926)]

b. *Вы пили вино мятежно, Вы брали* ***монбланную*** *ноту!* [И. Северянин. Сонаты в шторм (1911)]
Vy pili vino mjatežno, Vy brali ***monblannuju*** *notu!*
‘Vous buviez du vin avec rébellion, vous preniez la note **de Mont-Blanc** !’ [I. Severjanin. Sonates dans la tempête (1911)]

[16]Dans ce chapitre et le chapitre suivant, les références bibliographiques pour les exemples issus de RusCorpora sont présentées sans modification ni homogénéisation.

c. *... следовало-бы им объединиться в кооперативные ячейки (машинные, семенные,* ***коневодные*** *и т. д.* [Нужна плановость // «Нижегородский кооператор», 1928]
... sledovalo-by im ob"edinit'sja v kooperativnye jačejki (mašinnye, semennye, ***konevodnye*** *i t. d.*
'... [elles] devraient se regrouper en cellules coopératives (s'occupant des machines, semences, élevage **de chevaux**, etc.).' [Il faut de la planification // «Le coopérateur de Nijni-Novgorod», 1928]

d. *Рука — вырвалась из моих рук,* ***валькирийный****, гневно-крылатый шлем — где-то далеко впереди.* [Е. И. Замятин. Мы (1920)]
Ruka — vyrvalas' iz moix ruk, ***val'kirijnyj****, gnevno-krylatyj šlem — gde-to daleko vperedi.*
'La main – s'est échappée de mes mains, le casque **de Walkyrie**, furieux et ailé – quelque part loin devant.' [E. I. Zamjatin. Nous (1920)]

e. *Недосказ – стихотворное коварство, Чутьё* ***следопытное*** *народное.* [Н. А. Клюев. «Суровое булыжное государство...» (1921)]
Nedoskaz – stixotvornoe kovarstvo, Čut'ë ***sledopytnoe*** *narodnoe.*
'Non-dit est une ruse poétique, flair **de piste** populaire.' [N. A. Klujev. «L'État rude et pavé...» (1921)]

f. *Когда скелетированные останки доставили в краевое* ***судмедэкспертное*** *бюро, то их исследование показало, что он, возможно, принадлежит Анжеле Бурдаковой.* [Корзенников Сергей соб. корр. 'Труда'. ПРЕСТУПЛЕНИЕ БЕЗ СВИДЕТЕЛЕЙ // Труд-7, 2000.10.19]
Kogda skeletirovannye ostanki dostavili v kraevoe ***sudmedèkspertnoe*** *bjuro, to ix issledovanie pokazalo, čto on, vozmožno, prinadležit Anžele Burdakovoj.*
'Lorsque les restes squelettiques ont été livrés au bureau régional d'expertise **médico-légale**, leur examen a montré qu'il pourrait appartenir à Angela Burdakova.' [Sergej Korzenikov journaliste de 'Trud'. CRIME SANS TÉMOINS // Trud-7, 2000.10.19]

Lorsque le suffixe réel est *-Ov-*, mais le suffixe prédit est *-sk-*, il s'agit majoritairement des noms polysyllabiques avec les scores propre et humain assez élevés (13 cas sur 25). Il s'agit des noms d'animaux (191a) et d'organisations (191b).

(191) a. ОРАНГУТАН 'orang-outan'
ORANGUTAN
САПСАН 'faucon pèlerin'
SAPSAN

b. Норникель 'Nornickel'
Nornikel'

Tout comme dans des exemples précédents, les formations avec *-sk-* sont valables : ОРАНГУТАНСКИЙ (ORANGUTANSKIJ), САПСАНСКИЙ (SAPSANSKIJ), НОРНИКЕЛЬСКИЙ (NORNIKEL'SKIJ).

(192) a. *... и всё клала свою дрожащую ручку на мою* ***орангутановую*** *лапу.* [В. В. Набоков. Лолита (1967)]
... i vsë klala svoju drožaščuju ručku na moju ***orangutanovuju*** *lapu.*
'et [elle] posait sa main tremblante sur ma patte **d'orang-outan**.' [V. V. Nabokov. Lolita (1967)]

b. *Он знает в Подмосковье два* ***сапсановых*** *гнезда и летом заберёт птенцов, вставших на крыло.* [Михаил Бутов. Свобода // «Новый Мир», 1999]
On znaet v Podmoskov'e dva ***sapsanovyx*** *gnezda i letom zaberët ptencov, vstavšix na krylo.*
'Il connaît deux nids **de faucons** dans la région de Moscou et, en été, il récupérera les oisillons qui ont pris leur envol.' [Mixail Butov. Liberté // «Novyj Mir», 1999]

c. *... а «****норникелевый****» выдвиженец станет в этом деле помехой.* [Павел Вощанов. Проект «враги народа» (2003) // «Новая газета», 2003.01.02]
*... a «****nornikelevyj****» vydviženec stanet v ètom dele pomexoj.*
'... et le candidat **de "Nornikel"** deviendra un obstacle dans cette affaire.' [Pavel Voščanov. Projet «les ennemis du peuple» (2003) // «Novaja Gazeta», 2003.01.02]

Lorsque le suffixe attesté est *-sk-*, les noms sont mal classés avec *-Ov-* s'ils sont polysyllabiques ; le dernier phonème du thème n'est pas une dentale (4 cas sur 7). Ce sont des noms humains (193a), abstraits (193b) ou concrets (193c).

(193) a. МОНАРХ 'monarque'
MONARX
ШЛЮХА 'prostituée'
ŠLJUXA

b. КАРАОКЕ 'karaoké'
KARAOKE

c. СУСЕК 'mangeoire'
SUSEK

Cette distribution est plutôt atypique pour les noms qui se combinent avec *-Ov-*, puisque, comme nous l'avons vu dans la section 7.2.3, ce suffixe a des préférences pour différentes combinaisons des propriétés sémantiques et phonologiques, mais pas spécialement pour les noms polysyllabiques avec un radical se terminant par une consonne autre que dentale.

(194) a. *... схватились в яростной, беспощадной* ***караокской*** *схватке конкурсанты из Якутска, Краснодара, ...* [Александр МЕШКОВ. Даже овцы и сороки - все поют под караоке! // Комсомольская правда, 2013.08.19]
... sxvatilis' v jarostnoj, bespoščadnoj ***karaokskoj*** *sxvatke konkursanty iz Jakutska, Krasnodara, ...*
'... les concurrents de Iakoutsk, Krasnodar, ... se sont affrontés dans une bataille **de karaoké** féroce et impitoyable.' [Alexandre MEŠKOV. Même les moutons et les pies tous chantent au karaoké! // Komsomolskaja Pravda, 2013.08.19]

b. *Но сам напиток чересчур сладкий Только как прохладительный напиток «****Монархский****» Производителям — хвала за честность.* [Оксана КАСАТКИНА. Выбираем квас // Комсомольская правда, 2005.06.30]
*No sam napitok čeresčur sladkij Tol'ko kak proxladitel'nyj napitok «****Monarxskij****» Proizvoditeljam — xvala za čestnost'.*
'Mais la boisson elle-même est trop sucrée, seulement comme une boisson rafraîchissante **"Monarch"** Les fabricants sont loués pour leur honnêteté.' [Oksana KASATKINA. Choisir son kvas // Komsomol'skaja Pravda, 2005.06.30]

c. *В ресторане со скромным названием «****Сусекские*** *вампиры» подавали сациви под аккомпанемент песен из репертуара Сосо Павлиашвили.* [Владимир Абрамов. Игры патриотов. Спортивный агент Владимир Абрамов написал рассказ о Лиге СНГ // Советский спорт, 2013.02.09]
*V restorane so skromnym nazvaniem «****Susekskie*** *vampiry» podavali sacivi pod akkompanement pesen iz repertuara Soso Pavliašvili.*
'Au restaurant au nom modeste "Les vampires **du Susek**", on servait du satsivi accompagné de chansons du répertoire de Soso Pavliašvili.' [Vladimir Abramov. Les jeux des patriotes. L'agent sportif Vladimir Abramov a écrit un récit sur la Ligue de la CEI // Sovetskij Sport, 2013.02.09]

d. *Как это он меня назвал недавно —* ***шлюхская*** *блядь?..* [Михаил Гиголашвили. Экобаба и дикарь (1998-2007) // «Зарубежные записки», 2009]
Kak èto on menja nazval nedavno — ***šluxskaja*** *bljad'?..*
'Comment m'a-t-il appelée récemment – une putain **de salope** ?..' [Mixail Gigolašvili. L'écofemme et le sauvage (1998-2007) // «Notes étrangères», 2009]

Les noms se combinant avec *-n-* qui ont été mal classés avec *-Ov-* sont majoritairement monosyllabiques et abstraits (12 sur 21), exemplifiés en (195) et contextualisés en (196).

(195) БЛОНД ‘blond$_N$’
BLOND
ВЗОР ‘regard’
VZOR
ХРУСТ ‘craquement’
XRUST

(196) a. *В незабудковом вуальном платье, С белорозой в* ***блондных*** *волосах, Навещаешь ты в седьмой палате Юношу, побитого в горах...* [И. Северянин. В госпитале (1911)]
V nezabudkovom vual'nom plat'e, S belorozoj v ***blondnyx*** *volosax, Naveščaeš' ty v sed'moj palate Junošu, pobitogo v gorax...*
‘Dans une robe voilée de myosotis, avec une rose blanche dans tes cheveux **blonds**, tu rends visite dans la septième chambre au jeune homme battu dans les montagnes...’ [I. Severjanin. À l'hôpital (1911)]

b. *Ну, тогда напоследок погордимся ещё царём Василием Ивановичем Шуйским, которого самозванец при всём честном народе выпорол плетьми на* ***взорном*** *месте.* [Анатолий Мариенгоф. Циники (1928)]
Nu, togda naposledok pogordimsja eščë carëm Vasiliem Ivanovičem Šujskim, kotorogo samozvanec pri vsëm čestnom narode vyporol plet'mi na ***vzornom*** *meste.*
‘Eh bien, enfin, soyons fiers du roi Vassilij Ivanovič Šujskij, que l'imposteur a fouetté avec des lanières devant tout le peuple honorable en un **échafaud**.’ [Anatoli Mariengof. Les cyniques (1928)]

c. *... К Вам, проспекты, где дома, как баки, Где в* ***хрустном*** *лае трамвайной собаки Сумрак щупает у алкоголиков пульсы.* [В. Г. Шершеневич. «К Вам несу мое сердце в оберточной бумаге...» (1913-1915)]
... K Vam, prospekty, gde doma, kak baki, Gde v ***xrustnom*** *lae tramvajnoj sobaki Sumrak ščupaet u alkogolikov pul'sy.*
‘... à vous, avenues où les maisons sont comme des réservoirs, où dans le **crépitement** du tramway, le crépuscule tâte le pouls des alcooliques.’ [V.G. Šeršenevič. «Je vous apporte mon cœur enveloppé dans du papier d'emballage...» (1913-1915)]

Finalement, les suffixe *-n-* est prédit majoritairement pour les noms abstraits (38 sur 46) qui se combinent réellement avec *-sk-* (197, 198).

(197) АСПИРАНТУРА ‘études doctorales’
ASPIRANTURA
БАКАЛАВРИАТ ‘licence’
BAKALAVRIAT
КЛОУНАДА ‘bouffonnerie’

KLOUNADA
ЦИГУН ‘qigong’
CIGUN

(198) a. *г) Как Нэткины [Анны Гитерман]* ***аспирантурские*** *дела?* [Юлий Даниэль. Письма из заключения (1966-1970)]
g) Kak Nètkiny [Anny Giterman] ***aspiranturskie*** *dela?*
‘g) Comment la formation **doctorale** de Netka [Anna Giterman] se passe-t-elle ?’ [Julij Daniel’. Lettres de prison (1966-1970)]

b. *Неэффективные инженерные вузы понизятся в звании и перейдут на четырёхлетнюю* ***бакалаврiатскую*** *систему.* [Павел Панов. В России предлагают оставить лишь несколько элитных технических вузов // Известия, 2014.05.30]
Neèffektivnye inženernye vuzy ponizjatsja v zvanii i perejdut na četyrëxletnjuju ***bakalavriatskuju*** *sistemu.*
‘Les écoles d’ingénieurs inefficaces seront rétrogradées et passeront à un système **de licence** de quatre ans.’ [Pavel Panov. En Russie, il est proposé de ne conserver que quelques établissements techniques d’élite // Izvestia, 2014.05.30]

c. *— «Неужели крайние правые только и способны, чтобы ежедневно устраивать* ***клоунадские*** *буффонады?!»* [неизвестный. Государственная Дума. Аграрный понедельник «Национализация капиталов» (1907.04.09) // «Петербургская газета», 1907]
— «Neuželi krajnie pravye tol’ko i sposobny, čtoby ežednevno ustraivat’ ***klounadskie*** *buffonady?!»*
‘– "Les extrémistes de droite ne sont-ils vraiment capables que d’organiser quotidiennement des bouffonneries **de clowns** ?!"’ [inconnu. Douma d’État. Lundi agraire "Nationalisation des capitaux" (1907.04.09) // "Petersburg Gazette", 1907]

d. *Могу порекомендовать, например, книги Мантэка Чиа о* ***цигунских*** *практиках «внутренняя улыбка» и «целительные звуки».* [Илья Шабшин. В чем тайна «Секрета» // «Психология на каждый день», 2010]
Mogu porekomendovat’, naprimer, knigi Mantèka Čia o ***cigunskix*** *praktikax «vnutrennjaja ulybka» i «celitel’nye zvuki».*
‘Je peux recommander, par exemple, les livres de Mantak Čia sur les pratiques **du Qigong** "sourire intérieur" et "sons guérisseurs".’ [Ilya Šabšin. Quel est le mystère du "Secret" // "Psychologie pour tous les jours", 2010]

De plus, *-n-* est prédit pour les noms se combinant réellement avec *-Ov-* si ces noms sont polysyllabiques et leur radical se termine par une consonne autre que vélaire (96 sur 104). Il s’agit de noms qui sont soit concrets (199a), soit abstraits (199b).

(199) a. ГЛИНТВЕЙН ‘vin chaud’
GLINTVEJN
ПЛАЗМА ‘plasma’
PLAZMA
ФУКСИЯ ‘fuchsia’
FUKSIJA

b. БРИТЬЁ ‘rasage’
BRIT’Ë
РАКУРС ‘angle de vue’
RAKURS
ЦЕНТНЕР ‘quintal’
CENTNER

(200) a. *… я думаю, что я могу и не дожить до этого* ***глинтвейнового*** *рая.* [Александр МЕШКОВ. Как «пенсионер» Мешков работу искал // Комсомольская правда, 2011.08.04]
… ja dumaju, čto ja mogu i ne dožit’ do ètogo ***glintvejnovogo*** *raja.*
‘… je pense que je pourrais mourir avant ce paradis **de vin chaud**.’ [Alexandre MEŠKOV. Comment "retraité" Meshkov cherchait du travail // Komsomolskaïa Pravda, 2011.08.04]

b. *Здесь мощные, совершенно необъяснимые всплески магнитного поля сбивали приборы и расщепляли* ***плазмовый*** *шнур в реакторе фотонных ракет.* [Аркадий Стругацкий, Борис Стругацкий. Полдень. XXII век (1961-1967)]
Zdes’ moščnye, soveršenno neob"jasnimye vspleski magnitnogo polja sbivali pribory i rasščepljali ***plazmovyj*** *šnur v reaktore fotonnyx raket.*
‘Ici, des pics puissants et totalement inexplicables du champ magnétique perturbaient les instruments et sectionnaient le flux **de plasma** dans le réacteur des fusées à photons.’ [Arkadij Strugackij, Boris Strugackij. Midi. XXIIe siècle (1961-1967)]

c. *Пёстрая карамелька Если, шагая из весны в лето, вам неохота стягивать полюбившийся немаркий свитерок цвета пьяной вишни или любимые* ***фуксиевые*** *брючки, не особенно переживайте.* [Елена ЛЕВИНА. Платье в клеточку, чулки в полосочку // Комсомольская правда, 2001.05.11]
Pëstraja karamel’ka Esli, šagaja iz vesny v leto, vam neoxota stjagivat’ poljubivšijsja nemarkij sviterok cveta p’janoj višni ili ljubimye ***fuksievye*** *brjučki, ne osobenno pereživajte.*
‘Bonbon bariolé Si, en passant du printemps à l’été, vous n’avez pas envie de retirer votre pull préféré et discret de couleur cerise ivre ou vos pantalons **fuchsia** préférés, ne vous inquiétez pas trop.’ [Elena LEVINA. Robe à carreaux, bas rayés // Komsomolskaïa Pravda, 2001.05.11]

d. *... а мои руки* ***бритьевого*** *акта виднелись сквозь начавший выцветать фон, и четверорукое чудо возникало на жести вывески.* [К. С. Петров-Водкин. Моя повесть. Часть 2. Пространство Эвклида (1932)]
... a moi ruki ***brit'evogo*** *akta vidnelis' skvoz' načavšij vycvetat' fon, i četverorukoe čudo voznikalo na žesti vyveski.*
'... et mes mains de l'acte **de rasage** étaient visibles à travers le fond qui commençait à se faner, et la merveille à quatre mains apparaissait sur le panneau en tôle.' [K. S. Petrov-Vodkin. Mon histoire. Partie 2. L'espace d'Euclide (1932)]

e. *А во-вторых, совершенно очевидно, что Кайботт изображал не изнурённых рабочих, а натурщиков, которых — ради достижения эффекта выпуклых мышц — он и поставил в такие* ***ракурсовые****, трудовые позы.* [Николай Молок. Французский поцелуй // Известия, 2006.04.24]
A vo-vtoryx, soveršenno očevidno, čto Kajbott izobražal ne iznurënnyx rabočix, a naturščikov, kotoryx — radi dostiženija èffekta vypuklyx myšc — on i postavil v takie ***rakursovye****, trudovye pozy.*
'Et deuxièmement, il est tout à fait évident que Caillebotte représentait non pas des travailleurs épuisés, mais des modèles que – pour obtenir l'effet de muscles saillants – il avait placés dans des positions de travail **avec un certain angle de vue**.' [Nikolaj Molok. Baiser français // Izvestia, 2006.04.24]

f. *Подхожу, открываю дверь, а там сидит* ***центнеровая*** *девица в красных штанах и совершенно по-свински ржёт: "Тебе поди ехать надо?!"* [Орлов Павел. ХОД ФРАНЦУЗСКОЙ СВИНЬЕЙ // Труд-7, 2006.12.26]
Podxožu, otkryvaju dver', a tam sidit ***centnerovaja*** *devica v krasnyx štanax i soveršenno po-svinski ržët: "Tebe podi exat' nado?!"*
'Je m'approche, j'ouvre la porte et je vois une fille **d'une centaine de kilos** en pantalon rouge qui glousse comme un cochon : "Tu dois y aller, hein ?!' [Pavel Orlov. LE PAS DU COCHON FRANÇAIS // Trud-7, 2006.12.26]

Conclusion

Au cours de ce chapitre, nous avons employé deux modèles afin d'évaluer la concurrence suffixale dans les données de haute et basse fréquence : les arbres boostés et les arbres de décision simples.

Les arbres de décision boostés ont permis d'appréhender la capacité à prédire le suffixe dans les deux groupes de données, d'identifier les propriétés des noms de base les plus influentes dans la classification et de déterminer les modèles optimaux.

Globalement, les modèles affichent une excellente précision. Néanmoins, cette précision est supérieure pour les données de haute fréquence (0.8773, intervalle de confiance [0.8758, 0.8787]) par rapport aux données de basse fréquence (0.8189, intervalle de confiance [0.8147, 0.8230]). Une précision accrue accompagnée d'un intervalle de confiance plus réduit pour les données de haute fréquence suggère que celles-ci présentent des caractéristiques plus marquées et des schémas plus nets, facilitant ainsi l'identification des relations entre les variables par le modèle d'apprentissage automatique et permettant une meilleure généralisation à partir des données d'entraînement. Autrement dit, les données de haute fréquence révèlent davantage de régularité dans la distribution des propriétés des noms de base. En revanche, les données de basse fréquence peuvent comporter des caractéristiques moins prononcées, des schémas moins définis ou des distributions plus complexes. Par conséquent, les propriétés des noms de base imposent moins de contraintes sur le choix du suffixe et engendrent une plus grande variation.

L'analyse multivariée a confirmé l'importance des variables non catégorielles pour la sémantique et l'étymologie en ce qui concerne le choix des suffixes, en comparaison avec les propriétés catégorielles. Cependant, l'étymologie ne semble pas avoir un rôle significatif pour la classification. Seules trois propriétés suffisent pour prédire correctement le suffixe dans les données de haute fréquence : les scores sémantiques, la structure syllabique et les derniers phonèmes des radicaux. Dans les données de basse fréquence, un quatrième prédicteur est nécessaire : la présence d'allomorphies consonantiques dans l'espace thématique des noms. Ainsi, nous avons identifié des modèles optimaux pour décrire les données de haute et de basse fréquence.

Les arbres de décision simples ont permis d'analyser les interactions entre les propriétés dans chacun des deux modèles optimaux. De manière générale, le suffixe -*sk*- se distingue des deux autres suffixes grâce à des propriétés sémantiques (le score des noms propres et, dans une moindre mesure, le score des noms communs). La différenciation du suffixe -*Ov*- repose sur le niveau phonologique, notamment le dernier phonème du radical et la structure syllabique : -*Ov*- présente une prédilection pour les consonnes vélaires et les noms monosyllabiques. En ce qui concerne -*n*-, il peut être prédit comme suffixe par défaut dans tous les autres cas : les noms communs, avec le dernier phonème alvéolaire, dental ou labial, polysyllabiques, avec un score bas pour les noms humains et un score élevé pour les noms concrets. Les allomorphies consonantiques biaisent les prédictions également vers le suffixe -*n*-.

Comme nous l'avons observé, les données de basse fréquence manifestent une variation plus importante, d'où un arbre de décision plus élaboré. Néanmoins, cet arbre présente les mêmes schémas que celui décrivant les données de haute fréquence, avec des distinctions additionnelles visant à saisir la variation supplémentaire. Concernant les allomorphies consonantiques, elles ne permettent que de différencier les noms s'associant au suffixe -*n*-.

La dérivation des adjectifs dénominaux suit effectivement des schémas génériques permettant de prédire le suffixe avec une bonne précision, comme l'illustrent les modèles

d'arbres boostés. Toutefois, la présence d'erreurs suggère également que ces schémas ne sont pas absolus. Une variation ne se conformant pas nécessairement aux règles générales est observée, tant dans les données de haute fréquence que de manière plus marquée dans les données de basse fréquence. Cela révèle que la dérivation des adjectifs demeure flexible et peut s'adapter à des situations spécifiques ou à des cas particuliers.

Les erreurs dans la prédiction des suffixes adjectivaux mettent également en lumière la complexité inhérente à la langue russe. Ainsi, la distribution des propriétés d'un nom de base peut être typique pour un suffixe donné, alors que le nom en question s'associe en réalité à un autre suffixe. Ces cas représentent les défis les plus importants pour la modélisation, les modèles d'apprentissage automatique étant limités dans leur capacité à représenter cette complexité.

Chapitre 8

Productivité des suffixes

Sommaire

Introduction

Dans le chapitre précédent, nous avons axé notre étude de la concurrence sur les données de haute et de basse fréquence afin d'examiner comment les différentes propriétés des noms de base influencent le choix d'un suffixe adjectival. Dans ce chapitre, nous utiliserons l'intégralité du jeu de données RuDénom, qui comprend des adjectifs et leurs noms de base correspondants, ainsi que leurs fréquences respectives. Nous aborderons la concurrence du point de vue de la productivité des suffixes. Dans la section 3.1.2, nous avons montré que la productivité peut être considérée comme une propriété scalaire, ce qui ouvre la voie à des études quantitatives et statistiques de la productivité des différents affixes. Les suffixes *-n-*, *-sk-* et *-Ov-* étant tous les trois très productifs en synchronie, l'objectif consistera à quantifier leur productivité.

La productivité d'un affixe pourrait être mesurée selon différents critères : elle peut correspondre aux intuitions linguistiques et refléter la facilité avec laquelle l'affixe

se combine avec de nouveaux mots (Aronoff, 1976 ; Baayen, 1992). La productivité peut également être associée à la capacité des affixes à former de nouveaux mots, en particulier des hapax, nouveaux lexèmes supposément jamais produits auparavant (Plag, 2006). En outre, selon Hay et Baayen (2002), la productivité est liée au nombre de lexèmes construits décomposables, c'est-à-dire des lexèmes sémantiquement transparents, qui peuvent facilement être divisés en radical et en affixe. De plus, les mesures de productivité ne se limitent pas simplement au décompte des types, tokens ou hapax (Baayen, 1993 ; Plag, 2006), mais prennent également en compte leurs ratios différents (Baayen, 1993 ; Gaeta et Ricca, 2006 ; Baayen, 2009)[1].

Dans ce chapitre, notre attention sera focalisée sur deux indices de productivité. Tout d'abord, dans la section 8.1, nous examinerons les fréquences des adjectifs formés avec les trois suffixes en question, en prenant en compte les fréquences absolues (types et tokens) ainsi que les fréquences relatives. Nous discuterons également de la perception des adjectifs en tant qu'unités indécomposables ou d'unités composées d'un radical et d'un suffixe, en évoquant la question de la décomposabilité des adjectifs. Ensuite, dans la section 8.2, nous aborderons la productivité basée sur les hapax en présentant différentes méthodes d'évaluation que ces derniers offrent. Nous montrerons que la productivité des suffixes étudiés n'est pas homogène et varie en fonction des particularités linguistiques des noms de base.

8.1 Fréquences

8.1.1 Types et tokens

L'extraction des fréquences des lexèmes est une méthode couramment utilisée en lexicographie. Dans la mesure où la taille des corpus en ligne augmente, ils deviennent de plus en plus représentatifs de la langue étudiée. Cette augmentation de la représentativité des corpus contribue ainsi à améliorer la précision des méthodes d'analyse linguistique basées sur le calcul des fréquences.

L'une des mesures de la productivité des affixes consiste à comptabiliser le nombre de types formés à l'aide de chaque suffixe. Cette mesure est appelée le degré d'emploi (Baayen, 1993 ; Plag, 2006) ou la productivité réalisée (Baayen, 2009)[2]. Le tableau 8.1 présente la distribution des types (`V`) et des tokens (`N`) selon les suffixes -*n*-, -*sk*- et -*Ov*-.

La première observation que l'on peut faire est que les types d'adjectifs formés avec le suffixe -*n*- sont les plus nombreux, tandis que ceux en -*Ov*- sont les moins représentés.

[1] D'autres méthodes sont aussi possibles. Ainsi, Arutjunova (1961, pp.54-64) propose de comparer le nombre de mots pouvant servir de base à la dérivation (noms) avec le nombre de dérivés existants (adjectifs). Cette méthode est pertinente lorsqu'elle est appliquée à des bases clairement définies et facilement identifiables. Cependant, notre base de données RuDenom n'inclut pas tous les noms existants dans RusCorpora pour les comparer aux noms qui servent réellement de base pour les adjectifs dérivés. La méthode de Arutjunova (1961) n'est donc pas applicable à notre recherche.

[2] *Extent of use* (Baayen, 1993 ; Plag, 2006) et *realized productivity* (Baayen, 2009).

	-n-	*-sk-*	*-Ov-*
V	**5 039**	4 643	2 910
N	**8 920 404**	3 811 021	1 752 216

Tableau 8.1: Distribution des types et des tokens ; RuDénom

Dans son étude des adjectifs russes, Alekseeva (2011, p.90) constate des tendances similaires pour les types adjectivaux. Les mêmes distributions sont également observées pour les tokens[3].

Dans son étude des verbes russes construits avec les préverbes[4] *po-* et *voz-*, Antic (2012) avance l'idée que, de manière intuitive, le préverbe *po-* est bien plus productif que *voz-*, appuyant ce point de vue par le fait qu'il existe bien plus de verbes avec *po-* qu'avec *voz-*[5]. Lorsqu'on prend en compte les fréquences brutes des suffixes *-n-*, *-sk-* et *-Ov-*, c'est le suffixe *-n-* qui apparaît comme le plus productif. Toutefois, cela contredit les résultats des études précédentes sur la productivité des suffixes : comme mentionné dans la section 3.2.1, ce sont plutôt *-sk-* (Nemčenko, 1976) ou *-Ov-* (Alekseeva, 2011) qui sont considérés comme les plus productifs. De plus, comme nous l'avons présenté dans les sections 2.3 et 3.2.3, l'intuition linguistique suggère que c'est le suffixe *-Ov-* qui est le plus productif. Cette hypothèse est soutenue par sa capacité à se combiner avec un inventaire de bases de plus en plus diversifié, ainsi que par sa disponibilité à former des doublets pour des adjectifs déjà existants avec le suffixe *-n-*, afin d'exprimer un sens strictement relationnel. En outre, les fréquences obtenues par un simple décompte des types et des tokens reflètent plutôt la productivité passée des affixes (Plag, 2006 ; Baayen, 2009)[6] : si un affixe a été largement utilisé dans le passé, ce qui a conduit à la présence d'un grand nombre de types le contenant, cela ne permet pas de conclure que cet affixe est toujours productif dans la langue contemporaine (Varvara, 2017). Ces observations doivent être complétées par une analyse des fréquences plus détaillée.

La distribution des fréquences des unités lexicales dans un corpus linguistique est généralement hétérogène. En effet, on observe un petit nombre de mots à très haute fréquence et une majorité de mots à fréquence relativement faible. Cette distribution s'inscrit dans ce que Chitashvili et Baayen (1993) appellent une 'grande quantité d'événements rares'[7], avec les hapax représentant environ la moitié du corpus. La loi de Zipf (Zipf, 1935 ; Arapov *et al.*, 1975) établit une relation inversement proportionnelle entre le rang d'un mot dans la liste des fréquences et sa fréquence d'apparition. Un

[3]Cependant, Baayen et del Prado Martín (2005) montrent que la fréquence des tokens ne fournit pas d'informations pertinentes pour la productivité des affixes, et que la fréquence des types offre une meilleure approximation.

[4]Le terme *préverbe* est réservé par Guiraud-Weber (2004, pp.35-36) pour les préfixes verbaux qui perfectivent un verbe imperfectif simple, tout en changeant son sens lexicale (cf. петь (pet') 'chanter' / запеть (zapet') 'se mettre à chanter').

[5]Antic (2012) fournit les fréquences suivantes : 4 278 types pour *po-* et 1 236 pour *voz-*.

[6]*Past productivity* (Plag, 2006 ; Baayen, 2009).

[7]*Large Number of Rare Events distribution* (Chitashvili et Baayen, 1993, p.57).

mot de rang 1 est le plus fréquent dans le corpus, un mot de rang 2 est le deuxième plus fréquent, et ainsi de suite. Dans le cas de mots ayant une fréquence identique, leur classement est effectué de manière aléatoire.

La distribution zipfienne peut s'appliquer non seulement à un corpus de textes, où les mots les plus fréquents sont souvent des articles, des prépositions ou des conjonctions, mais aussi à des corpus de lexèmes tels que RuDénom, qui répertorie tous les adjectifs extraits de RusCorpora. Dans le tableau 8.2, nous pouvons observer les distributions des fréquences des suffixes par quartiles. Les résultats montrent que la moitié des données ont des fréquences inférieures à 15 (pour le suffixe -*n*-), voire même à 10 (pour les suffixes -*Ov*- et -*sk*-). En revanche, un peu plus de 25% des adjectifs ont des fréquences supérieures à 100 (pour -*n*- et -*Ov*-), seuil que nous avons défini pour délimiter notre sous-ensemble de lexèmes de haute fréquence.

Stats	-*Ov*-	-*n*-	-*sk*-
total	2 910	5 039	4 643
min	1	1	1
25%	2	2	1
50%	9	14	6
75%	112	195	47
max	121 739	385 082	431 267

Tableau 8.2: Fréquences absolues ; RuDénom

La figure 8.1 présente le spectre de fréquence (Baayen, 2001, p.8) pour les trois suffixes -*sk*-, -*n*-, et -*Ov*-, en fonction des classes de fréquences sur une échelle logarithmique (base 10)[8]. Les hapax, 10^0, sont plus fréquents pour le suffixe -*sk*- que pour le suffixe -*n*- et -*Ov*-. En général, les adjectifs formés avec le suffixe -*Ov*-, ayant une fréquence faible (jusqu'à 10 occurrences, 10^1), sont moins représentés que ceux formés avec les suffixes -*sk*- ou -*n*-.

Le calcul du nombre d'hapax pour chaque suffixe dans le corpus fournit une mesure alternative de la productivité, appelée productivité en expansion (Baayen, 2009)[9] : plus le nombre d'hapax est élevé, plus la règle morphologique correspondante est utilisée, et plus le suffixe contribue à l'augmentation du vocabulaire dans le corpus.

Il est à noter qu'un hapax n'est pas nécessairement un néologisme, il peut également correspondre à un mot rare mais ancien. De plus, certains auteurs utilisent souvent des mots spécifiques qui leur sont propres, ce qu'on nomme des hapax d'auteur

[8]La raison principale de l'utilisation de l'échelle logarithmique dans les figures de ce chapitre est d'améliorer leur lisibilité. Cependant, d'après Hay et Baayen (2002), il existe des indices du fait que les humains traitent la fréquence de manière logarithmique. En effet, les différences entre les fréquences plus basses sont perçues de manière plus marquée que les différences équivalentes entre les fréquences plus élevées. Par exemple, la différence entre 10 et 20 est perçue comme plus importante que la différence entre 1 010 et 1 020.

[9]*Expanding productivity* (Baayen, 2009).

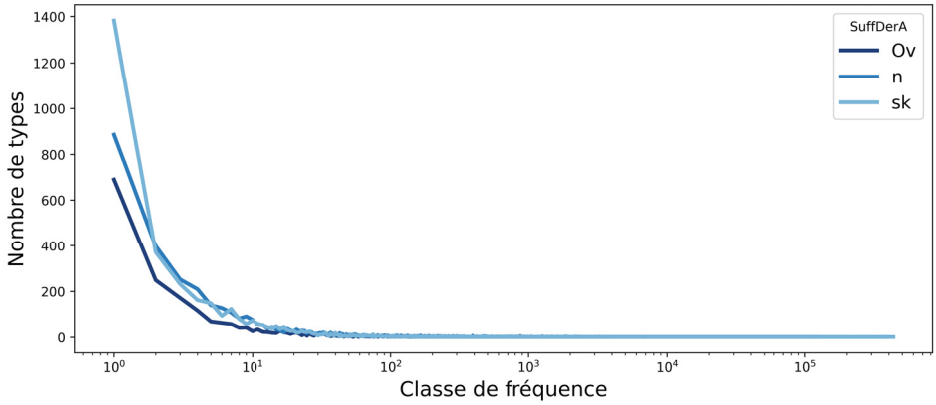

Figure 8.1: Fréquences absolues ; RuDénom

(Dal *et al.*, 2008)[10]. Cependant, les études montrent que plus la taille du corpus est grande, plus la proportion de néologismes parmi les hapax augmente (Plag, 2006). En outre, le nombre de néologismes le plus élevé est précisément retrouvé parmi les hapax (Baayen et Renouf, 1996). Dans cette optique, les hapax sont pertinents pour l'étude de la productivité.

Selon cette approche, c'est le suffixe -*sk*- qui est considéré comme étant le plus productif, car il offre davantage de possibilités de former de nouveaux mots (fréquence 10^0 sur la figure 8.1). Le tableau 8.3 présente le nombre total d'hapax (V_1) pour chaque suffixe[11].

	-*n*-	-*sk*-	-*Ov*-
V_1	886	**1 382**	689

Tableau 8.3: Distribution des hapax ; RuDénom

Il est à noter que 73% des adjectifs formés avec le suffixe -*sk*- sont construits à partir des toponymes. Cette particularité explique en grande partie le nombre significatif de hapax associés à ce suffixe.

Afin d'adapter ces chiffres qui présentent les fréquences absolues des hapax, Dal *et al.* (2008) proposent de mesurer le ratio du nombre d'hapax et le nombre total d'hapax dans le corpus. Cette approche permet de ramener les fréquences à une échelle de 0 à 1 et facilite son interprétation. Les ratios sont ainsi les suivants : -*n*- : 0.30, -*sk*- : 0.47, -*Ov*- : 0.23.

[10] Cf. la section 4.2.1 pour la discussion sur les occasionnalismes et les néologismes.

[11] Le total des hapax diffère de celui présenté dans le tableau 4.6. En effet, ce dernier référence les données sélectionnées pour l'étude des adjectifs de basse fréquence, sans considérer des doublets, qui sont répertoriés dans un sous-corpus distinct. Le tableau 8.3 fournit les détails sur les hapax dans l'ensemble du corpus RuDénom, y compris les doublets.

Une autre mesure de productivité consiste à évaluer la facilité des suffixes *-n-*, *-sk-* et *-Ov-* à se combiner avec de nouveaux mots, en comptant les hapax nominaux. Le nombre de base nominales hapax pour chaque suffixe est présenté dans le tableau 8.4.

	-n-	*-sk-*	*-Ov-*
Hapax nominaux	50	**155**	50

Tableau 8.4: Distribution des hapax nominaux ; RuDénom

Cette métrique fait de nouveau ressortir le suffixe *-sk-* comme le plus productif : il se combine avec 155 noms dont la fréquence est 1 dans RusCorpora, tandis que les suffixes *-n-* et *-Ov-* comptent chacun 50 hapax nominaux parmi leurs noms de base.

Nous avons présenté différentes mesures de productivité basées sur les fréquences absolues, incluant le nombre de types, de tokens, ainsi que d'hapax adjectivaux et nominaux. Cependant, ces mesures ne fournissent pas de conclusions uniformes concernant la productivité des différents suffixes.

Les fréquences absolues des dérivés ont traditionnellement été reliées à la transparence de ces derniers : plus la fréquence est élevée, plus la transparence sémantique est faible et moins le suffixe est productif (Baayen, 1993 ; Bybee, 1995). Toutefois, selon Hay (2001), la fréquence relative plutôt que la fréquence absolue est corrélée à la transparence et à la décomposabilité morphologique, qui ont impact direct sur la productivité des suffixes. Dans ce qui suit, nous allons donc combiner les fréquences des noms et les fréquences des adjectifs en une seule mesure. Avant de procéder à ces calculs, il convient d'introduire le concept de décomposabilité des lexèmes construits.

8.1.2 Décomposabilité

Le débat scientifique sur la manière dont nous accédons aux mots construits oppose deux solutions : certains chercheurs défendent l'idée que les mots construits sont perçus en tant qu'unités indécomposables, comme une seule unité lexicale (Butterworth, 1983), tandis que d'autres soutiennent que les mots sont perçus comme composés d'un radical et d'un affixe (Taft, 1985). Des modèles plus récents indiquent que ces deux approches sont possibles et qu'elles sont complémentaires (Hay, 2001 ; Hay et Baayen, 2002) : certains mots avec le même affixe peuvent être traités comme une seule unité lexicale, tandis que d'autres sont décomposables en leurs constituants.

Un des facteurs qui impacte le choix entre deux voies est la fréquence des lexèmes dérivés et de leurs lexèmes de base (Hay, 2004). Des expériences récentes en psycholinguistique montrent que la fréquence relative de la forme dérivée par rapport à la fréquence de sa base est un indicateur fiable pour déterminer si le mot sera reconnu dans son ensemble ou via ses composants : les dérivés plus fréquents que leurs bases sont plus susceptibles d'être reconnus comme des mots entiers, tandis que les dérivés moins fréquents sont plutôt traités par décomposition en radical et affixe (Hay, 2001 ;

Plag et Baayen, 2009).

La fréquence relative est ainsi le rapport entre la fréquence du mot dérivé et celle de sa base. Par exemple, dans la base de données RuDénom, l'adjectif криминальный (kriminal'nyj) (11 054) est beaucoup plus fréquent que le nom криминал (kriminal) 'crime' (989). En conséquence, cet adjectif serait traité comme une unité entière plutôt que décomposable en radical et suffixe. En revanche, l'adjectif абажурный (abažurnyj) (13) est beaucoup moins fréquent que sa base абажур (abažur) 'abat-jour' (1 691). Dans ce cas, cet adjectif serait plutôt accédé par voie de décomposition : via le radical <абажур>- et le suffixe -<н>-.

Il est à noter que, selon cette approche, l'accès à un mot via ses composants ou en tant qu'unité entière se fait exclusivement en fonction des fréquences relatives, indépendamment de la fréquence absolue du dérivé. Par conséquent, les formes peu fréquentes peuvent être reconnues comme un mot entier si leur base est encore moins fréquente. De même, les formes de haute fréquence peuvent être analysées comme décomposables si leurs bases sont encore plus fréquentes (Hay et Baayen, 2002).

Plag et Baayen (2009) mettent également en évidence les implications de cette perspective sur les mots construits pour la notion de complexité morphologique. Tout d'abord, le rôle du même suffixe dans l'accès lexical varierait en fonction des fréquences relatives de la base et du dérivé (comme nous l'avons montré précédemment avec les exemples de абажурный (abažurnyj) et криминальный (kriminal'nyj), tous les deux construits avec le suffixe -*n*-). De plus, les dérivés qui sont plus fréquents que leurs bases et qui sont caractérisés par un accès direct seraient moins transparents sur le plan sémantique que les mots qui sont moins fréquents que leurs bases et qui sont accédés via leurs constituants (Hay et Baayen, 2002). Enfin, plus le nombre de formes décomposables est élevé, plus un affixe est susceptible d'être productif. À l'inverse, un affixe représenté par de nombreux mots caractérisés par un accès direct est peu susceptible d'être productif, comme le soulignent Hay et Baayen (2002).

8.1.3 Fréquences relatives

Afin d'évaluer les fréquences relatives des adjectifs et de leurs noms de base, nous utiliserons plusieurs mesures. Les diagrammes de dispersion nous permettront de visualiser les fréquences des adjectifs et des noms. Le coefficient de corrélation de Pearson sera utilisé pour estimer si la corrélation entre les fréquences est significative. La droite de régression présentera également des mesures telles que la pente et l'intercept. Selon Hay et Baayen (2002), une corrélation positive significative, une pente forte et une valeur élevée de l'intercept sont caractéristiques d'un affixe plus productif. De plus, nous évaluerons les proportions d'adjectifs qui sont plus fréquents que leurs noms de base (non décomposables) et ceux qui sont moins fréquents (décomposables) pour chaque suffixe. Finalement, nous traiterons la transparence des adjectifs, qui correspond à l'inclusion du nom de base dans leur définition dans des dictionnaires.

La figure 8.2 illustre trois diagrammes de dispersion des fréquences relatives des

adjectifs et de leurs noms de base sur une échelle logarithmique[12], en fonction de trois suffixes différents.

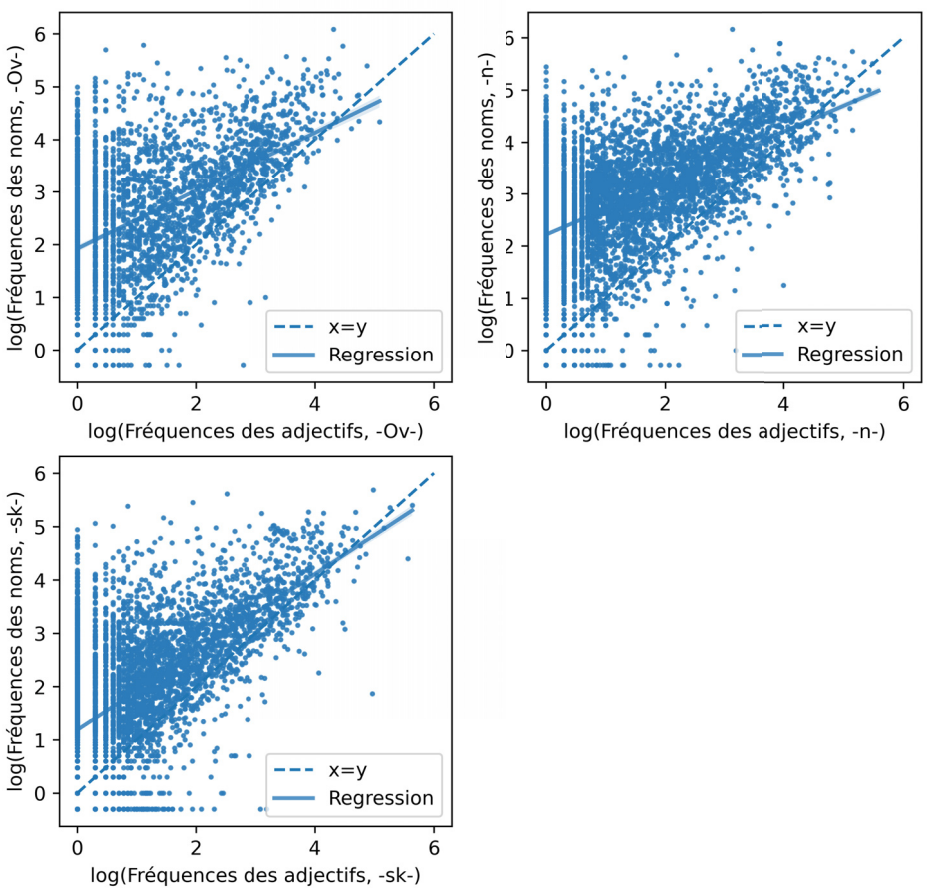

Figure 8.2: Fréquences relatives ; RuDénom

La ligne en pointillé `x=y` indique les cas où la fréquence des noms de base est équivalente à celle des adjectifs dérivés. Les points situés au-dessus de la ligne `x=y` représentent les données où la fréquence des adjectifs est inférieure à celle de leurs noms de base, tandis que les points situés en-dessous de cette ligne correspondent au cas où la fréquence des adjectifs est supérieure à celle de leurs noms de base.

La ligne continue représente la droite de régression : une représentation graphique de l'estimation de la dépendance linéaire théorique entre les fréquences adjectivales et les fréquences nominales. La formule de la droite de régression est la suivante :

[12]Certains noms de base ne sont pas répertoriés dans RusCorpora, ce qui entraîne des fréquences de 0. Nous les traitons donc comme ayant une fréquence de 0.5 afin de pouvoir appliquer une échelle logarithmique.

$$y = \beta x + \alpha$$

où :

β = la pente, l'inclinaison de la droite,
α = l'intercept, la valeur de y quand $x = 0$.

Comme il s'agit de variables continues, nous utiliserons le coefficient r de Pearson pour mesurer la force de la corrélation linéaire.

$$r = \frac{\sum_{i=1}^{n}(n_i - \bar{n})(a_i - \bar{a})}{\sqrt{\sum_{i=1}^{n}(n_i - \bar{n})^2 \sum_{i=1}^{n}(a_i - \bar{a})^2}}$$

où :

n = le nombre d'observations,
n_i, a_i = les valeurs des fréquences des noms et des adjectifs pour l'observation i
$\bar{n}, \bar{a}$ = les moyennes des fréquences des noms et des adjectifs.

Le coefficient de corrélation de Pearson peut prendre des valeurs dans l'intervalle [-1, 1]. Si $r = 0$, cela signifie que les deux variables sont indépendantes. Plus la valeur du coefficient se rapproche de 1 ou de -1, plus la relation de corrélation linéaire entre les deux variables est forte. Le signe du coefficient indique le sens de la corrélation (corrélation positive ou négative).

Les résultats indiquent une corrélation positive et significative pour les trois suffixes étudiés (tableau 8.5). Ces résultats confirment l'idée que *-n-*, *-sk-* et *-Ov-* sont tous productifs en synchronie et qu'ils sont réguliers sur le plan sémantique. Le coefficient r plus élevé pour *-sk-* suggère que ce suffixe est plus régulier que les deux autres, et donc aussi plus productif.

	-n-	*-sk-*	*-Ov-*
Coefficient r	0.55	**0.61**	0.50
Valeur p	<0.001	<0.001	<0.001

Tableau 8.5: Corrélation entre les fréquences des adjectifs et les fréquences de leurs noms de base ; RuDénom

Le coefficient de corrélation entre les fréquences des mots de base et celles de leurs dérivés est un aspect important à considérer du point de vue de la production. En termes de perception, ce sont l'intercept et la pente de la droite de régression qui présentent des mesures plus intéressantes, selon Hay et Baayen (2002). Un intercept plus élevé et une pente plus raide sont susceptibles de nous renseigner sur les taux de décomposabilité des adjectifs dérivés. Ces coefficients sont présentés dans le tableau 8.6.

	-n-	*-sk-*	*-Ov-*
Pente β	0.49	**0.73**	0.55
Intercept α	**2.21**	1.19	1.92

Tableau 8.6: Coefficients de régression pour les fréquences des adjectifs et les fréquences de leurs noms de base ; RuDénom

La pente la plus raide est observée pour le suffixe *-sk-* ; toutefois, la valeur d'intercept la plus élevée est observée pour *-n-*. Ces deux mesures ne sont pas concordantes, il est donc difficile d'estimer à partir de ces valeurs seuls quels adjectifs seraient plus susceptibles à être décomposables : ceux formés avec *-n-* ou ceux formés avec *-sk-*.

Comme nous l'avons mentionné plus haut, une corrélation significative entre la fréquence de la base et la fréquence du dérivé indique une productivité élevée et une régularité sémantique. Si un affixe est productif et régulier, alors la fréquence d'utilisation d'une forme dérivée peut être directement prédite à partir de la fréquence du mot de base. À l'inverse, une relation moins régulière et moins prévisible entre les bases et les dérivés entraîne une prédiction moins fiable de la fréquence des dérivés à partir de la fréquence des bases. Nous avons observé une corrélation forte et significative entre les fréquences des adjectifs avec les trois suffixes en question et leurs noms de base. Il est ainsi possible de prédire la fréquence des adjectifs dérivés, en renseignant les valeurs de la pente et de l'intercept dans la formule de régression.

Des études antérieures ont mis en évidence que les formes dérivées ont tendance à être moins fréquentes que leurs bases (Harwood et Wright, 1956 ; Hay, 2001 ; Sims et Parker, 2015). Les mêmes tendances ont été observées en russe pour les données préverbées (Antic, 2012). Les tendances observées sur les données des adjectifs dénominaux sont encore plus saillantes. Le tableau 8.7 présente le taux d'adjectifs pour chaque suffixe qui se trouvent au-dessus de la ligne `x=y` (qui sont moins fréquents que leurs bases), et au-dessous de cette ligne (plus fréquents que leurs bases).

	-n-	*-sk-*	*-Ov-*
$F_N > F_A$	4 566 (**0.91%**)	3 672 (0.79%)	2 571 (0.88%)
$F_N \leq F_A$	473 (0.09%)	971 (0.21%)	339 (0.12%)

Tableau 8.7: Décomposabilité des adjectifs ; RuDénom

De manière générale, la grande majorité des données sont au-dessus de cette ligne (entre 79 et 91%), ce qui démontre que la fréquence des adjectifs est souvent inférieure à celle de leur nom de base. La plupart des adjectifs qui se trouvent au-dessus de la ligne `x=y` sont donc décomposables et sont sémantiquement plus transparents. Les adjectifs qui se trouvent, à leur tour, en dessous de la ligne `x=y` sont donc susceptibles d'être lexicalisés.

On constate que 21% des adjectifs suffixés en *-sk-* se trouvent en dessous de la ligne x=y (971 types), tandis que les pourcentages sont plus faibles pour *-Ov-* (12%,

339 adjectifs) et encore plus faibles pour -*n*- (9%, 473 adjectifs). En d'autres termes, les adjectifs suffixés en -*n*- présentent la plus forte probabilité d'être accédés via leurs constituants et pas comme des entités uniques.

Cette conclusion peut paraître contre-intuitive. En effet, on pourrait s'attendre à ce que les adjectifs avec le suffixe -*n*- se caractérisent par un taux de décomposabilité plus faible et, par conséquent, par une plus forte perception en tant qu'unités entières. Ceci est dû au fait que ce sont ces adjectifs en -*n*- qui développent le sens qualificatif et deviennent ainsi moins transparents. Cependant, le fait que ces adjectifs puissent être considérés comme hautement décomposables confirme l'hypothèse selon laquelle, malgré son potentiel à construire des adjectifs qui peuvent être qualificatifs, le suffixe -*n*- reste disponible pour la formation d'adjectifs relationnels et conserve donc une grande productivité. Comme le montre le tableau 8.7, la productivité des affixes associée au taux d'adjectifs décomposables stipule alors que c'est précisément le suffixe -*n*- qui est le plus productif.

Le fait que les adjectifs soient plus fréquents que leurs bases implique non seulement qu'ils ont tendance à être perçus comme des unités non décomposables. Selon Hay et Baayen (2002), ces dérivés se sont détachés des propriétés de leurs bases ; ils sont donc moins transparents sémantiquement et peuvent subir un glissement sémantique. Selon Hay (2001), un adjectif est transparent si sa base est présente dans sa définition. Le tableau 8.8 présente la distribution des adjectifs transparents (le nom de base est présent dans la définition de l'adjectif dans un dictionnaire de référence[13] ; `B⊆Déf`) et non transparents (`B⊈Déf`) selon leurs fréquences relatives aux noms de base.

	-n-		*-sk-*		*-Ov-*	
	B⊈Déf	B⊆Déf	B⊈Déf	B⊆Déf	B⊈Déf	B⊆Déf
$F_N \leq F_A$	90	242	1	260	2	193
	27.11%	72.89%	0.38%	99.62%	1.03%	98.97%
$F_N > F_A$	138	2 915	9	1 352	36	1 432
	4.52%	95.48%	0.66%	99.34%	2.45%	97.54%
Total	228	3 157	10	1 612	38	1 625
	7%	93%	1%	**99%**	2%	98%

Tableau 8.8: Transparence des adjectifs ; RuDénom

La première observation révèle que les adjectifs sont généralement très transparents (ligne `Total`), avec un taux de plus de 90%. Cependant, il est à noter que les adjectifs en -*n*- présentent un taux de transparence légèrement inférieur (93%, comparé à 98% et 99%). Comme cette analyse est basée sur les définitions, elle est en accord avec le fait que les adjectifs en -*n*- sont plus enclins à la lexicalisation et à l'évolution vers un sens qualificatif. Le suffixe -*sk*- construit alors des adjectifs les plus transparents

[13]Pour estimer la transparence des adjectifs, nous avons extrait leurs définitions de Wiktionaire. Comme tous les adjectifs n'y sont pas référencés, le volume d'échantillon analysé est inférieur au volume de la base Rudénom.

(99%).

Cependant, les tendances dans les distributions de la transparence des adjectifs en fonction de la fréquence relative sont difficilement discernables (lignes $F_N \leq F_A$ et $F_N > F_A$).

Nous allons examiner les adjectifs de très haute fréquence ($\log(A)>5$) en termes de transparence. Ces adjectifs présentent la même ambiguïté constatée dans l'ensemble de données. En effet, il n'est pas systématique que les adjectifs plus fréquents que leurs bases soient lexicalisés. De même, les adjectifs moins fréquents que leurs bases ne sont pas nécessairement transparents. Cette observation est particulièrement marquée dans le cas des adjectifs en *-n-*, qui sont présentés dans le tableau 8.9.

	Base	F_N	Adjectif	F_A	B⊆Déf
$F_N <$	*глава* 'chef'	219 971	*главный* 'principale'	385 082	–
	glava		*glavnyj*		
F_A	*известие* 'nouvelle'	98 769	*известный* 'connu'	263 153	–
	izvestie		*izvestnyj*		
	лето 'été'	89 670	*летний*	114 589	+
	leto		*letnij*		
	обычай 'coutume'	16 679	*обычный* 'habituel'	141 843	–
	obyčaj		*obyčnyj*		
	род 'genre'	96 302	*родной* 'natal'	198 410	–
	rod		*rodnoj*		
	сбор 'collecte'	43 663	*сборный*	127 837	+
	sbor		*sbornyj*		
	страх 'peur'	61 244	*страшный* 'effrayant'	155 748	–
	strax		*strašnyj*		
$F_N >$	*вид* 'vue'	284 308	*видный* 'visible'	104 632	–
	vid		*vidnyj*		
F_A	*война* 'guerre'	157 045	*военный*	135 460	+
	vojna		*voennyj*		
	конец 'fin'	318 183	*конечный*	304 080	+
	konec		*konečnyj*		
	место 'lieu'	539 484	*местный*	140 534	–
	mesto		*mestnyj*		
	сила 'force'	339 048	*сильный* 'fort'	217 842	+
	sila		*sil'nyj*		
	труд 'travail'	168 493	*трудный* 'difficile'	114 185	–
	trud		*trudnyj*		

Tableau 8.9: Fréquences relatives et absolues, $\log(A)>5$; *-n-*

Pour les adjectifs plus fréquents que leurs bases ($F_N < F_A$), la présence de la base dans la définition de l'adjectif (+ dans la colonne B⊆Déf) est faible (2 cas sur 7). Ceci semble appuyer l'idée que ces adjectifs sont moins transparents. Par contre, pour les adjectifs

moins fréquents que leurs bases ($F_N > F_A$), le nombre de cas de présence et d'absence de la base dans la définition de l'adjectif sont équivalents. Ainsi, les adjectifs transparents et non transparents peuvent être plus ou moins fréquents que leurs bases. Par exemple, l'adjectif ЛЕТНИЙ (LETNIJ) est plus fréquent que sa base mais est également considéré comme transparent, tandis que l'adjectif ВИДНЫЙ (VIDNYJ) est moins fréquent que sa base mais n'est pas considéré comme transparent. De plus, la distinction entre adjectifs qualificatifs (accompagnés d'une traduction) et adjectifs de relation (sans traduction) ne semble pas dépendre des fréquences relatives : les deux types d'adjectifs peuvent se retrouver tant au-dessus qu'au-dessous de la ligne x=y.

En outre, il est à noter que la non transparence est typique des adjectifs qualificatifs, mais ne se limite pas à ceux-ci. Par exemple, l'adjectif de relation МЕСТНЫЙ (MESTNYJ) ne contient pas sa base МЕСТО (MESTO) dans sa définition (*постоянно пребывающий или встречающийся в данном регионе* (*postojanno prebyvajuščij ili vstrečajuščijsja v dannom regione*) 'résidant ou se trouvant constamment dans une région donnée'). À l'inverse, l'adjectif qualificatif СИЛЬНЫЙ (SIL'NYJ) intègre sa base СИЛА (SILA) dans sa définition (*обладающий силой* (*obladajuščij siloj*) 'possédant de la force').

Des exemples avec le suffixe *-sk-* sont présentés dans le tableau 8.10.

	Base	F_N	Adjectif	F_A	B⊆Déf
$F_N < F_A$	*Россия* 'Russie' *Rossija*	250 468	*российский* *rossijskij*	431 267	+
	Русь 'Rus' *Rus'*	25 149	*русский* *russkij*	364 436	+
	совет 'soviet' *sovet*	94 546	*советский* *sovetskij*	144 464	+
$F_N > F_A$	*Москва* 'Moscou' *Moskva*	227 173	*московский* *moskovskij*	182 116	+

Tableau 8.10: Fréquences relatives et absolues, log(A)>5 ; *-sk-*

Dans le cas des adjectifs en *-sk-*, ils sont tous transparents, indépendamment des fréquences relatives. Cela semble aller à l'encontre de l'hypothèse générale qui postule que les adjectifs plus fréquents que leurs bases sont généralement moins transparents. Un seul exemple avec le suffixe *-Ov-* (tableau 8.11) infirme aussi cette hypothèse.

	Base	F_N	Adjectif	F_A	B⊆Déf
$F_N < F_A$	*финансы* 'finances' *finansy*	21 752	*финансовый* *finansovyj*	121 739	+

Tableau 8.11: Fréquences relatives et absolues, log(A)>5 ; *-Ov-*

En guise de conclusion préliminaire, nous pouvons constater que les résultats issus des métriques basées sur les fréquences absolues et les fréquences relatives sont contradictoires. Ces métriques mettent tour à tour en évidence la productivité la plus

élevée du suffixe -*n*- ou du suffixe -*sk*-. Nous allons maintenant faire un point sur des mesures nécessitant une analyse plus approfondie des hapax.

8.2 Hapax

Dans cette section, nous allons évaluer la productivité des suffixes en tant que capacité à former de nouveaux mots. Pour ce faire, nous allons utiliser plusieurs mesures qui prennent en compte les hapax. Comme discuté dans la section précédente (8.1), les hapax ne correspondent pas toujours à des néologismes, mais ils peuvent être considérés comme une estimation fiable du statut néologique d'un mot.

Comme nous l'avons montré dans la section 8.1.1, un simple décompte des hapax indique que le suffixe -*sk*- est le plus productif. Toutefois, ces résultats ne tiennent pas compte du nombre d'hapax par rapport à toutes les occurrences d'un suffixe donné (tokens), ni de la taille globale du corpus. Dans cette section, nous allons prendre en compte ces deux facteurs.

8.2.1 Hapax et tokens

La première métrique, proposée par Baayen (1993, 2001) pour estimer la productivité, emploie le nombre d'hapax, mais sous forme de ratio par rapport au nombre total de tokens contenant l'affixe concerné. Cette mesure représente la productivité au sens restreint (Plag, 2006)[14] ou encore l'indice de productivité (Dal *et al.*, 2008).

La formule pour calculer la productivité est ainsi la suivante :

$$P = \frac{V_1}{N}$$

où :

V_1 = le nombre de hapax pour un affixe donné,
N = le nombre total de tokens contenant ce même affixe.

Cette métrique évalue la probabilité de rencontrer un nouveau type non attesté auparavant, c'est-à-dire un hapax, après avoir échantillonné N tokens contenant affixe donné (Baayen et Lieber, 1991). Le tableau 8.12 expose la productivité des suffixes -*n*-, -*sk*- et -*Ov*- en fonction du nombre de hapax et de tokens. Contrairement aux métriques employées dans la section précédente selon lesquelles les suffixes -*sk*- ou -*n*- ressortaient comme les plus productifs, le calcul basé sur le ratio des hapax au nombre total de tokens révèle que c'est le suffixe -*Ov*- qui est le plus productif des trois.

La productivité mesurée de cette manière est considérée par Hay et Baayen (2002) comme représentative de la décomposabilité morphologique abordée dans la section

[14] *Productivity in the narrow sense* (Plag, 2006) ; *potential productivity* (Baayen, 2009) ; ou encore *category-conditionned productivity* (Baayen, 1993).

	-*n*-	-*sk*-	-*Ov*-
V_1	886	1 382	689
N	8 920 404	3 811 021	1 752 216
P($\times 10^3$)	0.0993	0.3627	**0.3932**

Tableau 8.12: Productivité (ratio des hapax) ; RuDénom

8.1. Ces auteurs démontrent que les valeurs de productivité des affixes sont fortement corrélées au taux de décomposabilité : plus les mots sont accédés via leurs constituants, plus l'affixe est disponible pour la création de nouveaux mots. Toutefois, dans la section 8.1.2 nous avons constaté que le taux de décomposabilité des adjectifs avec le suffixe -*Ov*- est inférieur au taux de décomposabilité des adjectifs avec le suffixe -*n*- : ces derniers présentent le taux de décomposabilité le plus élevé. Ainsi, la corrélation observée par Hay et Baayen (2002) sur les données en anglais n'est pas constatée pour les adjectifs dénominaux russes[15].

Cette mesure de productivité présente d'autres désavantages. Tout d'abord, elle peut parfois afficher des valeurs contre-intuitives (Varvara, 2019). De plus, comme le soulignent Gaeta et Ricca (2006), la taille du corpus doit être suffisamment importante pour que les hapax soient majoritairement composés de néologismes. Par ailleurs, le nombre de tokens est également à prendre en compte : cette mesure accroîtrait la productivité pour les suffixes ayant un faible nombre de tokens présents dans le dénominateur de l'équation (pour des valeurs plus grandes de N, P diminuera, et vice versa.). Une critique fréquemment mentionnée est que cette mesure repose sur une comparaison entre les affixes ayant des fréquences de tokens très différentes, ce qui a des répercussions significatives sur la productivité. Afin de résoudre ce problème, Gaeta et Ricca (2006) ajustent la mesure de productivité proposée par Baayen en la calculant pour un nombre de tokens équivalent pour différents affixes.

8.2.2 Productivité potentielle

Comme mentionné précédemment, les principales objections concernant la mesure de productivité proposée par Baayen mettent en avant le fait que les résultats ne sont pas significatifs si, dans un corpus donné, on compare des affixes ayant des fréquences de tokens très différentes. Dans le cas du corpus RuDénom, le nombre total de tokens pour chaque suffixe dépasse le million d'occurrences. Cependant, le nombre de tokens pour le suffixe -*n*- est plus de cinq fois supérieur à celui du suffixe -*Ov*-, ce qui représente une différence non négligeable.

La solution intuitive pour calculer la productivité des suffixes pour un nombre équivalent de tokens consiste à prendre le nombre de tokens le plus faible (dans le cas

[15]Il est à noter que nous fondons notre analyse de décomposabilité uniquement en fonction de la position d'un adjectif par rapport à la ligne `x=y` abordée dans la section 8.1.3. Hay et Baayen (2002) affinent cette mesure et développent un seuil de fréquence des bases et des dérivés en fonction duquel le dérivé sera perçu comme une unité entière ou composé d'un radical et d'un affixe.

de RuDénom, les tokens des adjectifs en -*Ov*-). Toutefois, le rapport initial entre les hapax et le nombre de tokens pour les suffixes les plus fréquents serait perdu. Gaeta et Ricca (2006) proposent deux méthodes pour contourner ce problème : travailler avec un corpus 'variable' et estimer la productivité par interpolation et extrapolation.

La première solution repose sur l'utilisation d'un corpus général pouvant être facilement divisé en sous-corpus. À cet effet, les auteurs utilisent le journal *La Stampa*, de 1996 à 1998, qui permet alors une division en 36 sous-corpus chacun correspondant à un mois. Pour chaque sous-corpus, une liste complète des formes fléchies a été créée de manière indépendante avec leur fréquence de tokens respective. Ensuite, les formes pour les affixes étudiés ont été lemmatisées et un tableau final a été constitué pour calculer la productivité de chaque affixe avec le nombre de tokens, de types et d'hapax.

Un des inconvénients de cette méthode est qu'elle nécessite d'une vérification manuelle approfondie afin d'éliminer le bruit. De plus, elle n'est pas applicable à notre recherche, étant donné que nous n'avons pas accès à l'intégralité du RusCorpora. En outre, nous avons extrait les données de cinq corpus : général, journaux, multimédia, poétique et oral ; même si l'accès à la totalité des textes de ces corpus avait été possible, la division de ce corpus hypothétique en sous-corpus homogènes aurait été difficile. Comme le soulignent Gaeta et Ricca (2006), la méthode du corpus variable ne présenterait des résultats linguistiquement significatifs que si les sous-corpus sont uniformes en termes de typologie textuelle.

Toutefois, la méthode du corpus variable a montré que la fréquence de tokens reste globalement stable lorsque le nombre de sous-corpus augmente. En d'autres termes, pour chaque affixe donné, le nombre de tokens peut être évalué avec précision comme étant directement proportionnel au nombre total de tokens dans le (sous-)corpus. Cela implique que la productivité de chaque suffixe peut être estimée pour n'importe quelle valeur de N.

Cette conclusion a permis de développer la deuxième méthode, le calcul de productivité par interpolation et extrapolation (Baayen, 2001, pp.63-76) : en conditionnant le spectre de fréquence à une taille d'échantillon donnée N_0, et en travaillant avec d'autres tailles d'échantillon N plus petites ou plus grandes que N_0. Le calcul par interpolation et extrapolation émet ainsi une hypothèse sur la productivité d'un affixe pour un certain nombre de tokens, en observant les proportions effectivement attestées dans le corpus.

Gaeta et Ricca (2006) démontrent que les deux métriques, productivité calculée selon l'approche du corpus variable et productivité potentielle, s'alignent de manière significative. La productivité potentielle permet alors de s'affranchir de la conception d'un corpus variable et peut être appliquée à des données qui ne contiennent que les informations sur les fréquences, comme la base RuDénom.

Pour calculer la productivité potentielle des suffixes -*n*-, -*sk*- et -*Ov*- par interpolation et extrapolation, nous utiliserons la famille des modèles `LNRE` (Large-Number-of-Rare-Events) (Baayen, 2001). Le modèle qui nous intéresse particulièrement est le modèle Zipf-Mandelbrot fini (`fZM`) (Evert, 2004), implémenté dans la bibliothèque

`zipfR` (Evert et Baroni, 2007) en R.

Le tableau 8.13 présente les résultats du calcul, où la productivité potentielle des suffixes est calculée pour trois valeurs de N correspondant au nombre de tokens attestés pour chaque suffixe : P(N$_{\text{-}n\text{-}}$), P(N$_{\text{-}sk\text{-}}$) et P(N$_{\text{-}Ov\text{-}}$).

	-n-	*-sk-*	*-Ov-*
P(N$_{\text{-}Ov\text{-}}$)	0.2091	**0.2750**	0.2221
P(N$_{\text{-}sk\text{-}}$)	0.1945	**0.2717**	0.2175
P(N$_{\text{-}n\text{-}}$)	0.1709	**0.2688**	0.2133

Tableau 8.13: Productivité (potentielle) ; RuDénom

La productivité potentielle ne coïncide pas avec la productivité estimée avec le ratio des hapax et des tokens : alors que cette dernière considère *-Ov-* comme étant le plus productif, la productivité potentielle met en avant *-sk-* pour les trois tailles de corpus, P(N$_{\text{-}n\text{-}}$), P(N$_{\text{-}sk\text{-}}$) et P(N$_{\text{-}Ov\text{-}}$). Les deux mesures sont toutefois en accord concernant le suffixe *-n-* : celui-ci est le moins productif des trois.

En évaluant la productivité potentielle, nous pouvons établir un classement de la productivité des suffixes de la manière suivante : *-sk-* surpassant *-Ov-*, qui lui-même surpasse *-n-*.

8.2.3 Productivité conditionnée

Nous avons présenté les différentes mesures de productivité pour les trois suffixes étudiés en nous basant sur l'ensemble de données extraites de RusCorpora. Cependant, comme le souligne Plag (2006), il est généralement admis que certains types de suffixes dérivationnels sont plus pertinents dans certains types de textes que dans d'autres, mais les affirmations concernant la productivité d'un suffixe donné sont généralement faites sans prendre en compte cette différence de pertinence. L'objectif de la présente recherche n'est pas d'étudier la distribution des suffixes en fonction des registres (notre base de données n'a pas été conçue spécifiquement pour répondre à cette problématique), cependant, dans cette section, nous ferons le point sur la productivité des suffixes adjectivaux en fonction de leurs préférences pour les noms de base.

Dans les chapitres 5, 6 et 7, nous avons démontré que la sémantique des noms de base est la propriété la plus étroitement liée au choix entre *-n-*, *-sk-* et *-Ov-*. Notamment, les noms propres et les noms humains ont une forte préférence pour *-sk-*. Dans la section précédente, 8.2.2, la mesure la plus robuste de la productivité, la productivité potentielle, estime que c'est précisément le suffixe *-sk-* qui est le plus productif. Plag (2006) dans son étude de la productivité des suffixes dans différents registres estime que, au sein d'un seul registre, différents suffixes peuvent varier énormément dans leur productivité ; de même, un suffixe donné peut afficher des différences considérables de productivité entre les différents registres. En appliquant cette logique à notre étude, nous pouvons alors nous demander si la productivité du suffixe *-sk-* n'est pas

conditionnée par la présence d'un grand nombre d'adjectifs dérivés notamment à partir de toponymes.

Le sous-ensemble de lexèmes auquel une règle morphologique peut être appliquée est généralement défini comme son domaine de productivité, ou contrainte sur la productivité ; cela correspond aux caractéristiques qu'un lexème doit posséder pour être la base d'une règle morphologique considérée (Varvara, 2017).

Comme nous l'avons vu dans la section 5.2.4, les noms propres représentent une catégorie à part entière. Certaines études en linguistique de corpus les excluent même de l'analyse, par exemple, Nemčenko (1976) ou Alekseeva (2011), d'autres les intègrent (Chovanová, 2011 ; Strnadová, 2014). Zemskaja (2011, pp.235-236) affirme que le décompte exhaustif des adjectifs dérivés de toponymes est impossible. Leur quantité est inconnue puisque les toponymes sont aussi très nombreux. Ainsi, il est impossible de savoir si toutes les combinaisons ont été réalisées ou s'il reste des mots à créer. Comme la représentativité est l'une des caractéristiques principales de RusCorpora, nous ne pouvons pas nous attendre à y retrouver tous les adjectifs possibles dérivés de toponymes. De plus, comme l'objectif de RuDénom a été de référencer le maximum des adjectifs présents dans RusCorpora, tous les adjectifs ont une fréquence d'au moins 1. Cependant, les noms de base correspondants ne sont pas forcément référencés dans RusCorpora, leurs fréquences peuvent être égales à zéro. Au total, il y a 553 noms dans RuDenom avec une fréquence de 0. Plus de deux tiers de ces noms (63.11%) sont des toponymes qui dérivent des adjectifs avec le suffixe -*sk*-.

Dans cette section, nous allons exclure les adjectifs qui sont formés à partir de noms propres et calculer la productivité des suffixes lorsqu'ils construisent des adjectifs à partir des noms communs uniquement.

Le tableau 8.14 présente la distribution des types et des tokens pour les adjectifs dérivés avec les trois suffixes en question uniquement à partir des noms communs. Comparé aux distributions génériques (tableau 8.1), les trois suffixes ont moins de types et de tokens. Cependant, si les suffixes -*n*- et -*Ov*- ont perdu moins de 100 types chacun, le suffixe -*sk*- est représenté par à peine la moitié des types initiaux.

	-*n*-	-*sk*-	-*Ov*-
Types (V)	4 985	1 944	2 873
Tokens (N)	8 855 823	1 666 957	1 748 938

Tableau 8.14: Distribution des types et des tokens (noms communs) ; RuDénom

Le tableau 8.15 présente non seulement le nombre d'hapax pour chaque suffixe, mais également la mesure de productivité traditionnelle et les mesures de la productivité potentielle. Une fois les adjectifs hapaxiques formés à partir des toponymes éliminés, le suffixe -*sk*- ne conserve qu'un quart des effectifs.

La mesure de productivité traditionnelle fournit des résultats similaires à ceux observés pour l'ensemble des adjectifs : -*Ov*- est plus productif que -*sk*-, qui est lui-même plus productif que -*n*-. Toutefois, la productivité potentielle diffère : contrairement à

	-n-	*-sk-*	*-Ov-*
V_1	852	374	669
N	8 855 823	1 666 957	1 748 938
$P(\times 10^3)$	0.0962	0.2244	**0.3825**
$P(N_{\text{-}Ov\text{-}})$	0.2071	0.1717	**0.2208**
$P(N_{\text{-}sk\text{-}})$	0.2079	0.1751	**0.2211**
$P(N_{\text{-}n\text{-}})$	0.1658	0.0346	**0.2120**

Tableau 8.15: Productivité (noms communs) ; RuDénom

l'ensemble des adjectifs où *-sk-* était le plus productif, lorsqu'il s'agit de la dérivation des adjectifs à partir des noms communs, c'est le suffixe *-Ov-* qui est le plus productif des trois. Quant au suffixe *-sk-*, non seulement il perd en productivité, mais il occupe également la troisième place, se situant derrière *-n-*.

Ainsi, la productivité des suffixes adjectivaux est variable : si un adjectif est formé à partir d'un nom commun, *-Ov-* et, dans une moindre mesure, *-n-* affichent un degré de productivité élevé. En revanche, si l'étude inclut un grand nombre de toponymes, le suffixe *-sk-* se distingue comme étant le plus productif pour ce type de noms.

Conclusion

Dans ce chapitre, nous avons entrepris d'évaluer la productivité des suffixes *-n-*, *-sk-* et *-Ov-*. Bien qu'il soit généralement accepté que ces trois suffixes sont très productifs en synchronie, notre objectif a été de proposer un classement de leur productivité. À cette fin, nous avons présenté diverses mesures.

L'analyse de la productivité basée sur les fréquences absolues offre des résultats contradictoires, mettant en évidence soit le suffixe *-n-*, soit *-sk-*. De surcroît, les fréquences relatives ne semblent pas spécialement adaptées à notre recherche portant sur des suffixes très productifs : les résultats sont également variables, privilégiant soit *-n-*, soit *-sk-*. L'analyse de décomposition et de transparence ne nous a pas non plus permis d'aboutir à des conclusions homogènes : la première a mis en valeur le suffixe *-n-* qui construiraient les adjectifs hautement décomposables, tandis que la deuxième a fait ressortir *-sk-*.

Les mesures les plus robustes de la productivité prennent en compte le nombre d'hapax et le nombre total de tokens pour chaque suffixe spécifique. Une approche le plus fiable consiste à estimer la productivité de chaque suffixe pour un nombre donné de tokens. Ainsi, il est possible de conditionner le calcul de productivité pour chaque suffixe à un nombre équivalent de tokens. Appliquée à l'ensemble du corpus, cette mesure a conduit au classement suivant en termes de productivité : *-sk-* > *-Ov-* > *-n-*. Cependant, l'ensemble du corpus contient à la fois les noms propres et les noms communs. Comme nous l'avons observé dans les chapitres précédents, le suffixe *-sk-* est le plus enclin à se combiner avec les noms propres, notamment des toponymes.

Le calcul de la productivité potentielle pour chaque suffixe basé uniquement sur les noms communs a révélé que la productivité du suffixe -*sk*- est effectivement restreinte au domaine des noms propres. En ce qui concerne les noms communs, il est le moins productif des trois. Le classement de la productivité des suffixes pour les noms communs se présente donc comme suit : -*Ov*- > -*n*- > -*sk*-.

Chapitre 9

Exploration des doublets

Sommaire

Introduction

Dans le cadre de notre recherche, nous avons examiné la concurrence des adjectifs dénominaux en fonction des propriétés des noms de base et de la productivité des suffixes. Il nous reste à présent à aborder la question des doublets, c'est-à-dire des cas attestés dans RusCorpora où deux adjectifs, formés à partir de la même base, coexistent[1].

Le présent chapitre marque une déviation par rapport aux deux chapitres précédents sur la concurrence (7) et la productivité (8) des suffixes en ce qu'il adopte une approche

[1]Dans la section 4.2.2, nous avons brièvement évoqué la question des triplets en présentant quelques exemples. Étant donné que notre étude se concentre sur les doublets et que les triplets sont relativement rares, nous allons les répartir entre les trois groupes de doublets (*-n-/-Ov-*, *-n-/-sk*, et *-sk-/-Ov-*) pour les analyses présentées dans ce chapitre.

plus exploratoire. Il vise à esquisser un cadre initial d'investigation des doublets à partir d'analyses quantitatives, nous permettant d'extraire des conclusions préliminaires.

Dans la section 9.1, nous évaluerons ces doublets en nous appuyant sur les mêmes métriques que celles que nous avons utilisées pour examiner les données de haute et de basse fréquence, à savoir la distribution des propriétés des noms de base, et les données génériques de RuDénom, à savoir le degré de décomposition. De plus, nous évaluerons les fréquences relatives des adjectifs au sein de chaque paire de doublets.

La section 9.2 fera le point sur la nature des doublets. Nous analyserons en particulier trois situations : les faux doublets (doublets impliquant des adjectifs qualificatifs et relationnels, ou bien une homonymie ou une polysémie des noms de base), les doublets hapax et les doublets occasionnels (où un adjectif est très fréquent et un autre correspond à un hapax).

Enfin, dans la section 9.3, nous explorerons dans quelle mesure les doublets sont sémantiquement et fonctionnellement équivalents, en nous appuyant sur leur similarité contextuelle et propriétés distributionnelles.

9.1 Distributions

9.1.1 Propriétés des noms de base

Dans le chapitre 7, nous avons établi que trois propriétés – la structure syllabique, le dernier phonème des radicaux et les scores sémantiques – sont suffisantes pour prédire le suffixe adjectival dans les données de haute fréquence. De plus, il s'est avéré que l'analyse des allomorphies thématiques était également nécessaire pour les données de basse fréquence. Dans le chapitre 5, nous avons en outre conclu que ces quatre propriétés se caractérisent par des distributions spécifiques dans les noms qui se combinent avec différents suffixes. Dans cette section, nous analyserons la distribution de ces quatre propriétés dans les noms de base formant des doublets.

9.1.1.1 Structure syllabique

Le tableau 9.1 présente la distribution des noms formant des doublets selon la structure syllabique. Les doublets avec les suffixes *-n-*/*-Ov-* sont rarement formés avec des noms de base ayant quatre syllabes ou plus.

SyllN	*-n-*/*-Ov-*	*-n-*/*-sk-*	*-sk-*/*-Ov-*
1	106	15	20
2	297	133	29
3	180	152	23
4	29	41	3
5+	2	25	0

Tableau 9.1: Distribution du nombre de syllabes ; doublets

Comme dans le cas des données de haute et de basse fréquence (cf. tableau 5.5), les bases de deux et trois syllabes sont majoritaires à former les adjectifs doublets.

Suivant la même méthodologie que nous avons mise en place dans le chapitre 5, nous allons également analyser les résidus pour la meilleure compréhension des tendances. La figure 9.1 présente les résidus pour le nombre de syllabes dans les données de doublets.

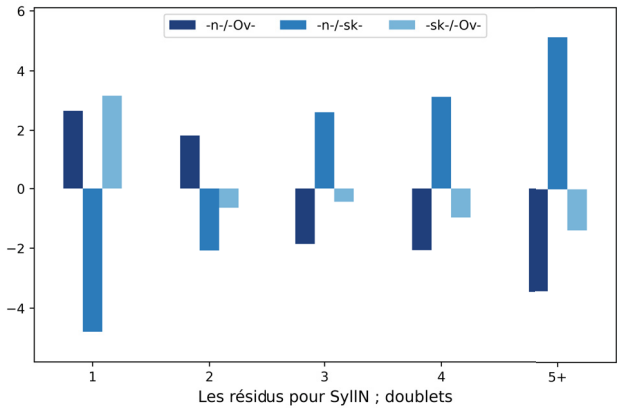

Figure 9.1: Résidus pour le nombre de syllabes ; doublets

L'analyse des résidus révèle que plus le nom de base est long, plus il a tendance à formes des adjectifs avec les suffixes *-n-/-sk*. Les doublets avec *-n-/-sk-* sont, au contraire, défavorisés pour les noms monosyllabiques.

9.1.1.2 Derniers phonèmes des radicaux

En ce qui concerne les derniers phonèmes (tableau 9.2), les radicaux se terminant par une consonne dentale (`cDent`) sont les plus courantes, dans les trois cas les doublets sont les plus fréquemment construits sur ce type de bases. Encore une fois, les radicaux avec une finale dentale sont les plus nombreux parmi les adjectifs doublets, comme c'est cas où il n'y a qu'un seul adjectif construit attesté (cf. tableau 5.8).

DPhoR	*-n-/-Ov-*	*-n-/-sk-*	*-sk-/-Ov-*
cAlv	151	43	9
cDent	400	287	50
cLab	44	34	8
cVel	19	2	8

Tableau 9.2: Distribution du dernier phonème des radicaux ; doublets

La figure 9.2 présente la distribution des résidus pour le dernier phonème des radicaux.

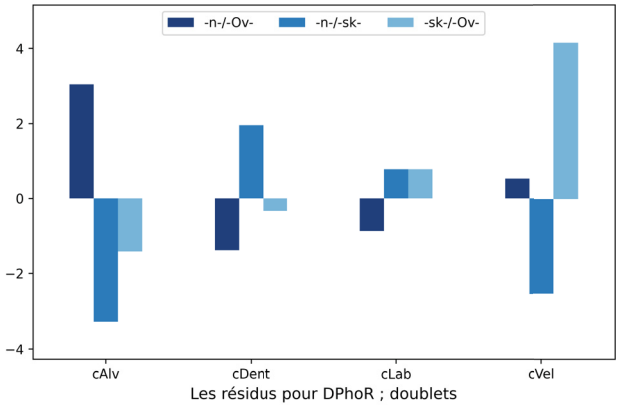

Figure 9.2: Résidus pour le dernier phonème des radicaux ; doublets

Les résidus montrent une préférence pour les finales alvéolaires dans les doublets *-n-/-Ov-* et pour les finales vélaires dans les doublets *-sk-/-Ov-*. Les finales dentales favorisent plutôt *-n-/-sk-*.

9.1.1.3 Allomorphies consonantiques

Généralement des doublets ne présentent pas d'allomorphies consonantiques, dans une moindre mesure ils présentent une palatalisation, la mouillure est minoritaire (tableau 9.3).

AllomC	*-n-/-Ov-*	*-n-/-sk-*	*-sk-/-Ov-*
0	468	321	67
mouil	41	35	6
palat	105	10	2

Tableau 9.3: Distribution des allomorphies consonantiques ; doublets

La figure 9.3 présente les résidus pour les allomorphies consonantiques. Les résidus montrent que la palatalisation favorise *-n-/-Ov-* et défavorise à la fois *-n-/-sk-* et *-sk-/-Ov-*.

9.1.1.4 Sémantique

Finalement, la figure 9.4 présente la distribution des scores sémantiques pour les doublets. Nous pouvons constater que les moyennes des scores pour les noms propres sont identiques pour les doublets *-n-/-sk-* et *-sk-/-Ov-*, elles sont supérieures à la moyenne du même score pour les doublets *-n-/-Ov-*. De même, les scores pour les noms humains sont plus élevés pour *-n-/-sk-* et *-sk-/-Ov-*, comparé aux doublets *-n-/-Ov-*. Par contre, on peut constater que le score des noms communs est beaucoup plus élevé

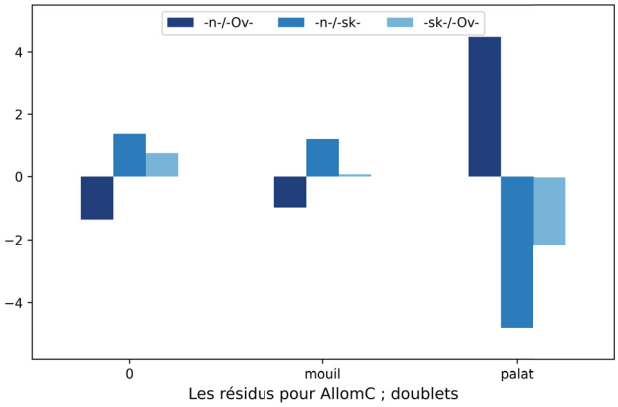

Figure 9.3: Résidus pour les allomorphies consonantiques ; doublets

pour les doublets *-n-/-Ov-*. En ce qui concerne les scores des noms abstraits, ils sont assez similaires pour les trois sous-ensembles de données de doublets.

Une analyse plus détaillée de la structure sémantique des noms qui constituent la base des doublets révèle que, parmi les doublets *-n-/-Ov-*, les noms de base correspondent principalement à des entités concrètes : des outils (201a), des plantes (201b), des substances et des éléments chimiques (201c), des matériaux et des matières (201d). Les noms abstraits sont aussi largement présents (201e).

(201) a. ИГЛА ‘aiguille’ → ИГЛОВОЙ / ИГОЛЬНЫЙ
IGLA IGLOVOJ IGOL’NYJ
СТАНОК ‘machine’ → СТАНКОВЫЙ / СТАНОЧНЫЙ
STANOK STANKOVYJ STANOČNYJ

b. МАЛИНА ‘framboise’ → МАЛИНОВЫЙ / МАЛИННЫЙ
MALINA MALINOVYJ MALINNYJ

c. ФЕНОЛ ‘phénol’ → ФЕНОЛОВЫЙ / ФЕНОЛЬНЫЙ
FENOL FENOLOVYJ FENOL’NYJ

d. ЦЕМЕНТ ‘ciment’ → ЦЕМЕНТОВЫЙ / ЦЕМЕНТНЫЙ
CEMENT CEMENTOVYJ CEMENTNYJ

e. ЛИСТОПАД ‘chute des feuilles’ → ЛИСТОПАДОВЫЙ / ЛИСТОПАДНЫЙ
LISTOPAD LISTOPADOVYJ LISTOPADNYJ

Les bases des doublets en *-n-/-sk-* sont principalement représentés par des noms caractérisant les individus sur le plan physique et intellectuel (202a), ainsi que par des noms désignant des professions (202b). Les noms faisant référence à des lieux (202c) et à des artefacts (202d) sont moins fréquents ; les noms abstraits sont également présents (202e).

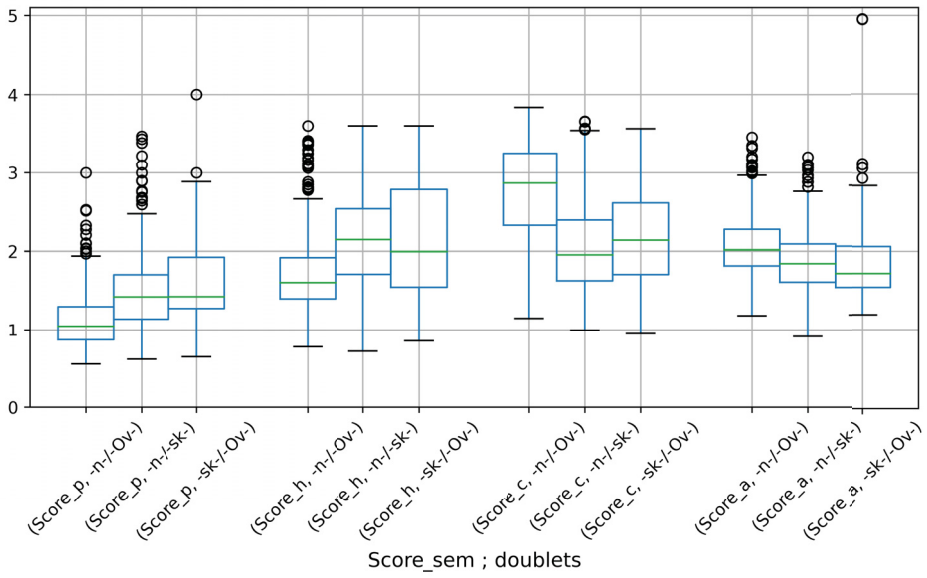

Figure 9.4: Scores sémantiques ; doublets

(202) a. МУЛАТ ‘mulâtre’ → МУЛАТНЫЙ / МУЛАТСКИЙ
MULAT MULATNYJ MULATSKIJ
ИДИОТ ‘idiot’ → ИДИОТНЫЙ / ИДИОТСКИЙ
IDIOT IDIOTNYJ IDIOTSKIJ

b. ПЕКАРЬ ‘boulanger’ → ПЕКАРНЫЙ / ПЕКАРСКИЙ
PEKAR’ PEKARNYJ PEKARSKIJ

c. ЛАБОРАТОРИЯ ‘laboratoire’ → ЛАБОРАТОРНЫЙ / ЛАБОРАТОРСКИЙ
LABORATORIJA LABORATORNYJ LABORATORSKIJ

d. ТРАМВАЙ ‘tramway’ → ТРАМВАЙНЫЙ / ТРАМВАЙСКИЙ
TRAMVAJ TRAMVAJNYJ TRAMVAJSKIJ

e. ШИРПОТРЕБ ‘biens de consommation courante’ → ШИРПОТРЕБНЫЙ / ШИРПОТРЕБСКИЙ
ŠIRPOTREB ŠIRPOTREBNYJ ŠIRPOTREBSKIJ

Les noms qui forment des doublets avec *-sk-*/*-Ov-* désignent des animaux (203a) et des humains (203b) ; des entités abstraites (203c), les noms désignant des lieux (203d) sont moins fréquents.

(203) a. СКОРПИОН ‘scorpion’ → СКОРПИОНСКИЙ / СКОРПИОНОВЫЙ
SKORPION SKORPIONSKIJ SKORPIONOVYJ

b. ХАМ ‘goujat, rustre’ → ХАМСКИЙ / ХАМОВЫЙ
XAM XAMSKIJ XAMOVYJ

c. ЮАНЬ ‘yuan (monnaie)’ → ЮАНЬСКИЙ / ЮАНЕВЫЙ
JUAN’ JUAN’SKIJ JUANEVYJ

d. АД ‘enfer’ → АДСКИЙ / АДОВЫЙ
AD ADSKIJ ADOVYJ

On observe une présence de noms abstraits dans les trois distributions, ce qui renforce l’idée que cette propriété n’est pas discriminante dans le cas des doublets.

Il est à noter que les doublets ne partagent ni la même signification, ni la même combinabilité. Ainsi, l’usage de *станковый пулемёт* (*stankovyj pulemët*) ‘mitrailleuse montée’ est approprié, tandis que l’expression ?*станочный пулемёт* (*stanočnyj pulemët*) l’est moins. De la même manière, *малиновый пиджак* (*malinovyj pidžak*) ‘veste framboise’ est couramment utilisé, tandis que l’expression ?*малинный пиджак* (*malinnyj pidžak*) est moins commune ou incorrecte. Nous allons revenir sur la combinabilité des doublets dans la section 9.3.

Pour résumer, la distribution des propriétés clé des noms de base dans les doublets résulte de la combinaison des distributions pour les suffixes individuels. Dans le cas des doublets *-n-*/*-Ov-*, la présence d’une consonne alvéolaire en fin de radical et la nature concrète des noms sont spécifiques au suffixe *-n-*, tandis que la prédominance de noms monosyllabiques est caractéristique du suffixe *-Ov-*.

De même, pour les doublets *-sk-*/*-Ov-*, les noms monosyllabiques et la finale vélaire des radicaux sont attribuables à *-Ov-*, tandis que la présence d’une consonne dentale en fin de radical et la prévalence de noms propres ou humains sont typiques du suffixe *-sk-*.

Enfin, dans le cas des doublets *-n-*/*-sk-*, les caractéristiques mentionnées pour *-sk-* demeurent identiques, tandis que les particularités associées à *-n-* comprennent les noms polysyllabiques, la terminaison des radicaux par une consonne dentale et la présence d’allomorphies consonantiques.

9.1.2 Fréquences des doublets

Dans le chapitre 8, nous avons évalué les fréquences absolues et les fréquences relatives des adjectifs. Dans cette section, nous allons également analyser les fréquences absolues pour chaque couple de doublets, mais nous limiterons la discussion des fréquences relatives entre les noms et les adjectifs à la décomposabilité des adjectifs, les coefficients de corrélation et de régression étant utilisés exclusivement pour mesurer la productivité. De plus, étant donné que nous traitons systématiquement de deux adjectifs doublets, nous pourrons mettre en place une analyse des fréquences relatives entre ces adjectifs.

9.1.2.1 Fréquences absolues des adjectifs

La figure 9.5 présente les spectres de fréquences pour les adjectifs doublets.

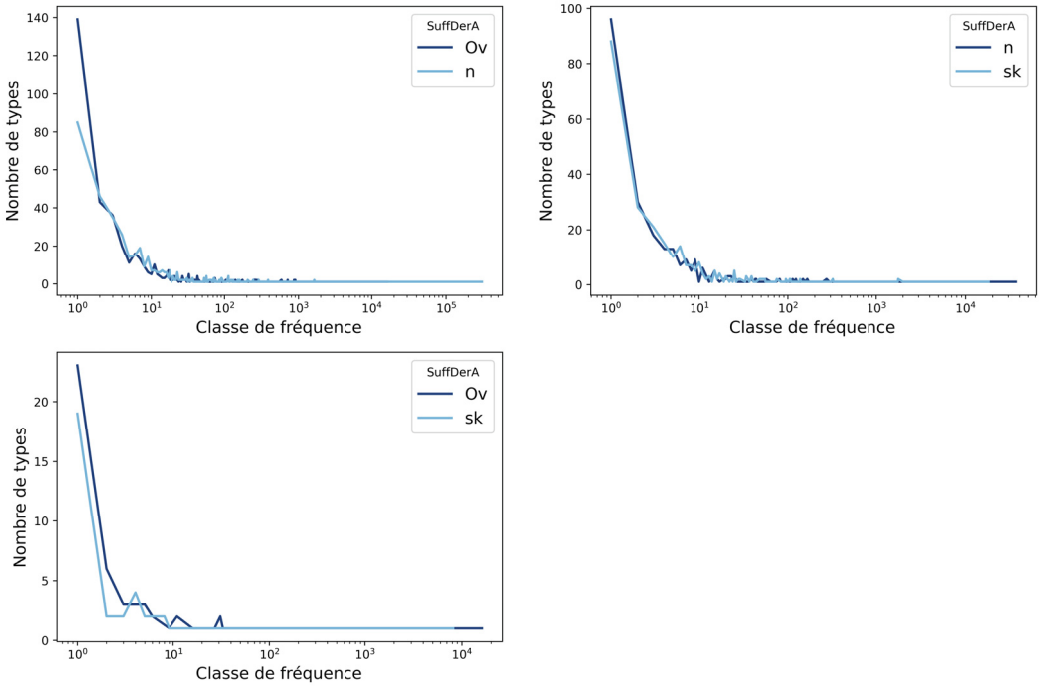

Figure 9.5: Spectres de fréquence ; doublets

Lorsque l'on compare les distributions pour les doublets -*n*-/-*Ov*-, on constate que les adjectifs formés avec le suffixe -*n*- sont plus fréquents que ceux formés avec -*Ov*- (la fréquence maximale pour -*n*- est de 38 421, contre 18 105 pour -*Ov*-). Cependant, le suffixe -*Ov*- a produit davantage d'hapax (143) que le suffixe -*n*- (85).

Pour ce qui est des doublets -*n*-/-*sk*-, leurs spectres de fréquences sont presque identiques. Les adjectifs formés avec le suffixe -*n*- présentent à la fois un nombre plus élevé d'hapax (la différence est moins marquée que pour -*n*-/-*Ov*-, avec 96 hapax pour -*n*- contre 89 pour -*sk*-) et des fréquences plus élevées (la fréquence maximale est de 38 421 pour -*n*-, contre 18 105 pour -*sk*-).

Enfin, dans le cas des doublets -*sk*-/-*Ov*-, ce sont les adjectifs formés avec le suffixe -*Ov*- qui ont à la fois le plus grand nombre d'hapax (23 contre 19) et les fréquences les plus élevées (75 067 contre 14 012).

9.1.2.2 Fréquences relatives entre adjectifs et noms

Le tableau 9.4 résume les proportions d'adjectifs qui sont plus fréquents ou moins fréquents que leurs noms de base.

		-*n*-/-*Ov*-		-*n*-/-*sk*-		-*sk*-/-*Ov*-	
		-*n*-	-*Ov*-	-*n*-	-*sk*-	-*sk*-	-*Ov*-
x=y	$F_N > F_A$	0.96%	0.97%	0.95%	0.99%	0.97%	0.96%
	$F_N < F_A$	0.04%	0.03%	0.05%	0.01%	0.03%	0.04%

Tableau 9.4: Métriques pour les fréquences relatives ; doublets

Concernant la décomposition des adjectifs, les données reflètent les tendances déjà observées : la majorité des adjectifs se situent au-dessus de la ligne `x=y`, c'est-à-dire qu'ils sont moins fréquents que leurs noms de base et sont donc caractérisés par un accès en décomposition plutôt que par un accès direct. Un nombre restreint d'adjectifs se trouve cependant systématiquement en dessous de la ligne `x=y`, ces adjectifs sont plus fréquents que leurs bases et sont donc perçus en tant qu'unité indivisible.

Alors que dans le corpus RuDénom, tous les suffixes pouvaient être ordonnés selon le taux de décomposition des adjectifs construits de la manière suivante : -*n*- > -*Ov*- > -*sk*-, les données des doublets ne suivent pas les mêmes tendances. Pour les doublets -*n*-/-*Ov*-, le taux de décomposition est plus élevé pour -*Ov*- (97%) que pour -*n*- (96%). Concernant les doublets -*n*-/-*sk*-, l'écart est plus prononcé en faveur de -*sk*- (99% contre 95% pour -*n*-). Si on compare les doublets -*sk*-/-*Ov*-, c'est le suffixe -*sk*- qui construit le plus d'adjectifs décomposables (97%) par rapport à -*Ov*- (96%).

Finalement, par rapport à la base de données RuDénom dont nous avons analysé les fréquences relatives dans la section 8.1.3, les adjectifs doublets se distinguent par un taux de décomposition plus élevé que les adjectifs n'ayant pas de doublon. Pour rappel, ces derniers affichaient un taux de décomposition variant entre 79% (pour le suffixe -*sk*-) et 91% (pour le suffixe -*Ov*-). En revanche, le taux de décomposition des doublets dépasse systématiquement les 95%. Il apparaît donc que les doublets sont moins fréquents que leurs noms de base.

Nous allons maintenant examiner les adjectifs qui sont plus fréquents que leurs bases ; qui se situent donc au-dessus de la ligne `x=y` et se caractérisent par une complexité sémantique élevée. Il est important de souligner qu'aucun nom ne produit deux adjectifs plus fréquents que la base : soit un seul adjectif du couple de doublets est non décomposable, soit les deux sont décomposables[2].

Concernant les doublets -*n*-/-*Ov*-, on compte 26 adjectifs (4%) avec le suffixe -*n*- et 16 adjectifs (3%) avec le suffixe -*Ov*- qui sont plus fréquents que leurs bases. Pour

[2]La seule exception concerne la dérivation de людный (ljudnyj) (1720) et людской (ljudskoj) (5415) à partir de люди (ljudi) 'gens' (181). Il est à noter que le mot люди (ljudi) représente le pluriel supplétif du nom человек (čelovek) 'homme, personne'. Cependant, le nom de base люди (ljudi) a été conservé dans les deux cas, car dans les dictionnaires tels que Wiktionaire, les entrées pour человек (čelovek) et люди (ljudi) sont séparées ; de plus, la recherche par lemmes sur RusCorpora présente les résultats à la fois pour человек (čelovek) et pour люди (ljudi). Si l'on considère que le nom de base dans les deux cas est человек (čelovek) (1 440 939), alors les deux adjectifs en question seraient moins fréquents que ce nom de base et seraient ainsi considérés comme décomposables.

-*n*-, nous retrouvons à la fois les adjectifs qualificatifs (204a) et relationnels (204b). En général, ces adjectifs et leurs noms de base présentent des fréquences très élevées (pour les adjectifs qualificatifs) et élevées (pour les adjectifs relationnels) ; néanmoins, il y a quelques cas d'adjectifs de relation peu fréquents, dont les noms de base sont encore moins fréquents (204c).

(204) a. БОЛЬ 'douleur' (37 359) → БОЛЬНОЙ 'malade' (41 674)
BOL' BOL'NOJ
ГРОЗА 'orage' (10 558) → ГРОЗНЫЙ 'menaçant' (25 512)
GROZA GROZNYJ
ГРЯЗЬ 'saleté' (23 516) → ГРЯЗНЫЙ 'sale' (27 095)
GRJAZ' GRJAZNYJ
ДУРЬ 'sottise' (1 846) → ДУРНОЙ 'stupide' (18 633)
DUR' DURNOJ
ЖИР 'graisse' (6 948) → ЖИРНЫЙ 'gras' (8 279)
ŽIR ŽIRNYJ
СТРАХ 'peur' (61 244) → СТРАШНЫЙ 'effrayant' (155 748)
STRAX STRAŠNYJ

b. ВИСКОЗА 'viscose' (146) → ВИСКОЗНЫЙ (158)
VISKOZA VISKOZNYJ
ЖЕЛЧЬ 'bile' (1 748) → ЖЕЛЧНЫЙ (1 805)
ŽELČ' ŽELČNYJ
ПУРПУР 'pourpre$_N$' (945) → ПУРПУРНЫЙ (1 370)
PURPUR PURPURNYJ
ХРУСТАЛЬ 'cristal' (1 845) → ХРУСТАЛЬНЫЙ (3 661)
XRUSTAL' XRUSTAL'NYJ

c. АУСТЕНИТ 'austénite' (8) → АУСТЕНИТНЫЙ (18)
AUSTENIT AUSTENITNYJ
ОЛДСКУЛ 'old school' (2) → ОЛДСКУЛЬНЫЙ (22)
OLDSKUL OLDSKUL'NYJ
ТИАЗОЛ 'thiazole' (1) → ТИАЗОЛЬНЫЙ (3)
TIAZOL TIAZOL'NYJ

Parmi les adjectifs en -*Ov*- dans les doublets -*n*-/-*Ov*-, les 16 adjectifs plus fréquents que leurs bases sont uniquement des adjectifs relationnels. Leurs fréquences sont élevées (205a), mais on retrouve également des fréquences plus faibles (205b).

(205) a. ДЕБЕТ 'débit' (277) → ДЕБЕТОВЫЙ (284)
DEBET DEBETOVYJ
КОРАЛЛ 'corail' (1 179) → КОРАЛЛОВЫЙ (1 230)
KORALL KORALLOVYJ
МАЛИНА 'framboise' (3 326) → МАЛИНОВЫЙ (3 878)
MALINA MALINOVYJ

пластик ‘plastique’ (4 533) → пластиковый (7 464)
PLASTIK PLASTIKOVYJ
резина ‘caoutchouc’ (3 826) → резиновый (7 319)
REZINA REZINOVYJ
эбонит ‘ébonite’ (160) → эбонитовый (235)
ÈBONIT ÈBONITOVYJ

b. канаус ‘canaus (tissue) ’ (20) → канаусовый (30)
KANAUS KANAUSOVYJ

Concernant les doublets *-n-*/*-sk-*, pour *-n-*, on constate 17 adjectifs non décomposables, avec une présence de quelques adjectifs qualificatifs (206a), mais les adjectifs relationnels sont cette fois-ci plus nombreux (206b).

(206) a. интеллектуал ‘intellectuel$_N$’ (1 870) →
INTELLEKTUAL
интеллектуальный ‘intellectuel$_A$’ (11 475)
INTELLEKTUAL’NYJ
интеллигент ‘personne cultivée’ (6 578) →
INTELLIGENT
интеллигентный ‘cultivé’ (8 077)
INTELLIGENTNYJ
люди ‘gens’ (181) → людный ‘fréquenté’ (1 720)
LJUDI LJUDNYJ
мода ‘mode’ (13 337) → модный ‘à la mode’ (13 993)
MODA MODNYJ

b. кулинар ‘cuisinier’ (265) → кулинарный (2 544)
KULINAR KULINARNYJ
санитар ‘sanitaire’ (2 956) → санитарный (10 970)
SANITAR SANITARNYJ
чужеземец ‘étranger$_N$’ (768) → чужеземный (855)
ČUŽEZEMEC ČUŽEZEMNYJ
ювелир ‘bijoutier’ (2 009) → ювелирный (4 917)
JUVELIR JUVELIRNYJ
янтарь ‘ambre’ (800) → янтарный (2 137)
JANTAR’ JANTARNYJ

De plus, tous ces adjectifs affichent des fréquences élevées, à l’exception d’un seul : абстинент (ABSTINENT) ‘abstinent$_N$’ (15) → абстинентный (ABSTINENTNYJ) (53).

Seuls 4 adjectifs formés avec le suffixe *-sk-* dans les couples de doublets *-n-*/*-sk-* sont non décomposables (207).

(207) АЗИАТ ‘Asiatique$_N$’ (1 216) → АЗИАТСКИЙ (9 121)
AZIAT AZIATSKIJ
БУТАФОРИЯ ‘trompe-l’œil’ (425) → БУТАФОРСКИЙ (818)
BUTAFORIJA BUTAFORSKIJ
КОЛЛЕЖ ‘collège’ (71) → КОЛЛЕЖСКИЙ (1 300)
KOLLEŽ KOLLEŽSKIJ
ЛЮДИ ‘gens’ (463) → ЛЮДСКОЙ (5 415)
LJUDI LJUDSKOJ

Enfin, parmi les doublets *-sk-*/*-Ov-*, seulement 3 adjectifs ne seraient pas considérés comme décomposables, formés avec les suffixes *-sk-* (208a) et *-Ov-* (208b), respectivement.

(208) a. АНГОРА ‘angora’ (58) → АНГОРСКИЙ (118)
ANGORA ANGORSKIJ
b. КАШЕМИР ‘cachemire’ (284) → КАШЕМИРОВЫЙ (295)
KAŠEMIR KAŠEMIROVYJ
ЦИГЕЙКА ‘manteau’ (48) → ЦИГЕЙКОВЫЙ (67)
CIGEJKA CIGEJKOVYJ

9.1.2.3 Fréquences relatives entre les adjectifs

Pour calculer les fréquences relatives des adjectifs au sein d’un couple de doublets, nous allons simplement diviser la fréquence de chaque adjectif par la somme des fréquences des adjectifs dans le couple de doublets. Par exemple, dans le dernier cas en (208b), la fréquence de l’adjectif ЦИГЕЙКОВЫЙ (CIGEJKOVYJ) est de 67 ; son doublet ЦИГЕЙСКИЙ (CIGEJSKIJ) est un hapax avec une fréquence de 1. Ainsi, la fréquence relative de ЦИГЕЙКОВЫЙ (CIGEJKOVYJ) est de 0.985, tandis que la fréquence relative de ЦИГЕЙСКИЙ (CIGEJSKIJ) est de 0.015.

Il est à noter que les fréquences relatives calculées de cette manière reflètent les proportions des fréquences des adjectifs au sein de chaque couple de doublets, mais que les fréquences absolues peuvent varier considérablement pour les mêmes valeurs des fréquences relatives. Pour illustrer ce point, le tableau 9.5 présente un échantillon de doublets *-n-*/*-Ov-* avec des fréquences relatives de 0.7/0.3, ainsi que les fréquences absolues correspondantes.

La figure 9.6 illustre les fréquences relatives des adjectifs doublets, arrondies au dixième près.

La majorité des types se situent aux extrémités de la figure, ce qui indique qu’il existe généralement une forte préférence pour un adjectif par rapport à son doublet. Ces doublets peuvent être considérés comme occasionnels et seront analysés plus en détail dans la section 9.2.3. En ce qui concerne la partie centrale de chaque figure (entre 0.4 et 0.6), les différences entre le nombre de types sont moins prononcées. Ces doublets sont généralement représentés par un nombre de types assez faible, à

Nom	A1	F1	R1	A2	F2	R2
пурпур *purpur* 'pourpre$_N$'	*пурпурный* *purpurnyj*	1 370	0.7	*пурпуровый* *purpurovyj*	512	0.3
верба *verba* 'saule'	*вербный* *verbnyj*	690	0.7	*вербовый* *verbovyj*	247	0.3
аммоний *ammonij* 'ammonium'	*аммонийный* *ammonijnyj*	53	0.7	*аммониевый* *ammonievyj*	23	0.3
окурок *okurok* 'mégot'	*окурочный* *okuročnyj*	6	0.7	*окурковый* *okurkovyj*	3	0.3

Tableau 9.5: Exemples des fréquences relatives ; *-n-*/*-Ov-*

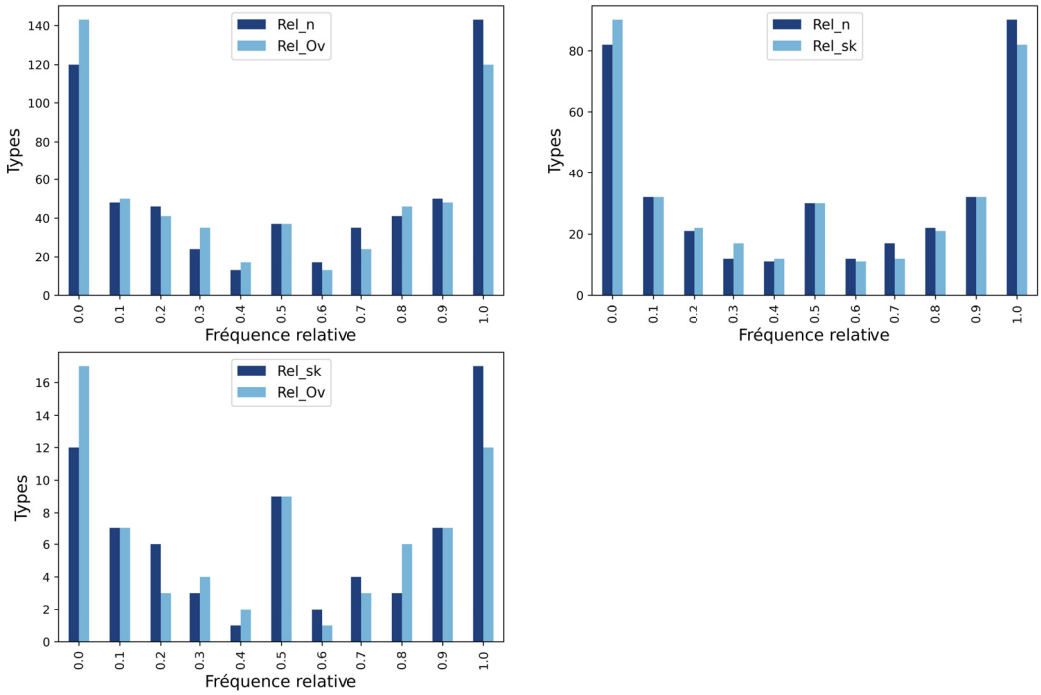

Figure 9.6: Fréquences relatives des adjectifs ; doublets

l'exception des cas où les fréquences relatives sont équivalentes (0.5/0.5). Strnadová (2014, pp.98-107), dans son étude de la concurrence entre les adjectifs dénominaux en français et les constructions prépositionnelles, considère ces cas comme des exemples de variation libre. Il n'y pas de préférence nette pour une forme, il s'agit donc de la

concurrence plus marquée. Nous examinerons ces cas plus en détail dans la section 9.2.2[3].

Avant de passer à une analyse plus approfondie, il est important de préciser la nature des adjectifs présents dans la base de données de doublets. Comme nous l'avons constaté plus tôt dans cette section, certains 'doublets' formels concernent les adjectifs qualitatifs et les adjectifs relationnels. Cependant, dans ces cas, on ne peut pas considérer qu'il s'agit des doublets à proprement parler.

9.2 Nature des doublets

Dans cette section, notre première tâche consistera à distinguer les doublets véritables des faux. Les faux doublets seront examinés plus en détail dans la section 9.2.1. Une fois ces cas mis de côté, nous nous concentrerons sur certaines caractéristiques spécifiques des doublets véritables : en particulier, nous nous intéresserons aux doublets concurrents, dont les fréquences relatives se situent entre 0.4 et 0.6. Ces derniers feront l'objet d'une analyse détaillée dans la section 9.2.2, avec une attention particulière portée aux doublets hapax, qui présentent une fréquence relative de 0.5. Dans la section 9.2.3, notre attention se portera sur les cas où les fréquences relatives sont extrêmes, c'est-à-dire entre 0.9 et 1. Un intérêt particulier sera accordé aux doublets occasionnels, où la fréquence d'un adjectif dépasse 100, tandis que le second adjectif du couple est un hapax.

9.2.1 Faux doublets

Dans la section précédente, nous avons noté que certains couples d'adjectifs doublets comprennent un adjectif qualificatif et un adjectif relationnel. Toutefois, ces cas ne correspondent pas à la définition de doublets, dans la mesure où ils ne représentent pas des adjectifs en concurrence directe exprimant le même sens lexical référentiel. De plus, la notion de doublets suppose que ces derniers sont formés à partir non seulement de la même base morphologique, mais également de la même base sémantique. Ainsi, dans cette section, nous allons aussi identifier les cas où les noms de base sont homonymes. Dans de telles situations, les adjectifs dérivés ne peuvent pas non plus être considérés comme des doublets.

Le tableau 9.6 présente le nombre de faux doublets dans le corpus.

Type	*-n-/-Ov-*	*-n-/-sk-*	*-sk-/-Ov-*
Qualificatif/Relationnel	36	2	0
Homonymie/Polysémie	4	2	3

Tableau 9.6: Faux doublets

[3]Dans cette même section, nous reviendrons également sur les pics visibles sur chaque figure au niveau des fréquences relatives 0.5/0.5.

Comme nous l'avons souligné à plusieurs reprises, les adjectifs formés avec le suffixe *-n-* sont plus susceptibles de développer un sens qualificatif. Cette tendance est particulièrement prononcée dans le cas des doublets en *-n-*/*-Ov-*, où nous observons le plus grand nombre de ces adjectifs. Dans tous ces cas, les adjectifs correspondants en *-Ov-* sont des adjectifs relationnels. Des exemples d'adjectifs qualificatifs formés avec *-n-* ont été présentés en (204a) dans la section 9.1.2.

Parmi les doublets *-n-*/*-sk-*, seuls deux adjectifs avec le suffixe *-n-* (à savoir МИРНЫЙ (MIRNYJ) 'pacifique' et МОДНЫЙ (MODNYJ) 'à la mode') ont été identifiés comme ayant un sens qualificatif. En revanche, tous les adjectifs parmi les doublets *-sk-*/*-Ov-* sont relationnels.

Il est à noter que les limites entre les classes d'adjectifs ne sont pas rigides (voir la section 2.1.2). Ainsi, certains adjectifs relationnels peuvent développer un sens qualificatif. Comme le font remarquer Hénault et Sakhno (2015), la sémantique des adjectifs formés avec le suffixe *-n-* : les adjectifs en (209a) contient une quantification de l'attribut ('beaucoup de neige', 'beaucoup de son'), ce qui les rapproche des adjectifs qualificatifs (209b). Cependant, ils peuvent également fonctionner comme des adjectifs de relation (209c)[4].

(209) a. СНЕЖНЫЙ 'enneigé'
SNEŽNYJ
ЗВУЧНЫЙ 'sonore'
ZVUČNYJ

b. *Такой в этом году февраль снежный!* 'Quel février enneigé cette année !'
Takoj v ètom godu fevral' snežnyj!

c. *снежное пространство* 'un espace enneigé'
snežnoe prostranstvo

Les adjectifs avec les suffixes *-sk-* et *-Ov-* peuvent également développer un sens qualificatif. L'adjectif АДСКИЙ (ADSKIJ) 'infernal' fonctionne comme un adjectif relationnel lorsqu'il signifie 'comme en enfer' (210a). En tant qu'adjectif qualificatif, ce mot acquiert un sens figuré : 'extrême, exceptionnel' (210b). De la même manière, АДОВЫЙ (ADOVYJ) peut également avoir un sens qualificatif 'dur, lourd' (210c).

(210) a. *адские пытки* 'les tortures infernales'
adskie pytki

b. *адское терпение* 'patience exceptionnelle'
adskoe terpenie

c. *адовые муки* 'tourments de l'enfer'
adovye muki

[4]Exemples tirés de Hénault et Sakhno (2015).

Les exemples en (211), (212) et (213) listent les cas de faux doublets construits à partir de différents sens de noms homonymiques ou polysémiques, avec des exemples de contexte.

(211) a. ДИАГОНАЛЬ ‘diagonale (tissu)’ → ДИАГОНАЛЕВЫЙ
DIAGONAL’ DIAGONALEVYJ
диагоналевая ткань ‘tissu en diagonale’
diagonalevaja tkan’
ДИАГОНАЛЬ ‘diagonale’ → ДИАГОНАЛЬНЫЙ
DIAGONAL’ DIAGONAL’NYJ
диагональное положение ‘position diagonale’
diagonal’noe položenie

b. ДРОБЬ ‘fraction’ → ДРОБОВОЙ
DROB’ DROBOVOJ
дробовой заряд ‘charge de grenaille’
drobovoj zarjad
ДРОБЬ ‘grenaille’ → ДРОБНЫЙ
DROB’ DROBNYJ
дробное число ‘nombre fractionnaire’
drobnoe číslo

c. СВОД ‘récapitulatif’ → СВОДНЫЙ
SVOD SVODNYJ
сводные данные ‘données récapitulatives’
svodnye dannye
СВОД ‘voûte’ → СВОДОВЫЙ
SVOD SVODOVYJ
сводовые структуры ‘structures voûtées’
svodovye struktury

d. СТОЛ ‘capitale’ → СТОЛЬНЫЙ
STOL STOL’NYJ
стольный град ‘ville capitale’
stol’nyj grad
СТОЛ ‘table’ → СТОЛОВЫЙ
STOL STOLOVYJ
столовые приборы ‘couverts de table’
stolovye pribory

(212) a. АТЛАС ‘satin’ → АТЛАСНЫЙ
ATLAS ATLASNYJ
атласное платье ‘robe en satin’
atlasnoe plat’e
АТЛАС ‘Atlas’ → АТЛАССКИЙ
ATLAS ATLASSKIJ

атласские горы ‘Montagnes de l’Atlas’
atlasskie gory

b. КВАРТАЛ ‘quartier’ → КВАРТАЛЬСКИЙ
KVARTAL KVARTAL’SKIJ
квартальские гангстеры ‘gangsters du quartier’
kvartal’skie gankstery
КВАРТАЛ ‘trimestre’ → КВАРТАЛЬНЫЙ
KVARTAL KVARTAL’NYJ
квартальный отчёт ‘rapport trimestriel’
kvartal’nyj otčët

(213) a. ГРАФ ‘graph’ → ГРАФОВЫЙ
GRAF GRAFOVYJ
графовые структуры ‘structures de graphes’
grafovye struktury
ГРАФ ‘comte’ → ГРАФСКИЙ
GRAF GRAFSKIJ
графская конюшня ‘écurie du comte’
grafskaja konjušnja

b. СВЕТ ‘monde’ → СВЕТСКИЙ
SVET SVETSKIJ
светские разговоры ‘conversations mondaines’
svetskie razgovory
СВЕТ ‘lumière’ → СВЕТОВОЙ
SVET SVETOVOJ
световой день ‘jour lumineux’
svetovoj den’

c. ФРАНК ‘franc’ → ФРАНКОВЫЙ
FRANK FRANKOVYJ
франковый гонорар ‘honoraires en francs’
frankovyj gonorar
ФРАНК ‘Franc (peuple)’ → ФРАНКСКИЙ
FRANK FRANKSKIJ
франкский король ‘roi des Francs’
frankskij korol’

Nous avons identifié en (211-213) les cas les plus saillants d’homonymie ou de divergence sémantique. Cependant, la tâche de distinguer les homonymes et de les écarter reste complexe. Par exemple, ДУМА (DUMA) peut signifier à la fois ‘pensée’ et ‘assemblée, conseil’. C’est généralement le second sens qui prédomine dans les adjectifs doublets. Ainsi, ДУМСКИЙ (DUMSKIJ) fait référence à la Douma d’État, la chambre basse du parlement russe actuel (214a), tandis que ДУМНЫЙ (DUMNYJ) est un terme historique se référant à la Douma des boyards – le conseil suprême auprès du chef

de l'État (prince, grand prince, tsar), un organe consultatif et législatif qui existait du XIVe au XVIIIe siècle (214b)[5]. Comme ces deux adjectifs font référence au même concept d'organe législatif, mais à différentes périodes, ils ont été conservés dans la base de données.

(214) a. *думские фракции* 'les factions de la Douma'
dumskie frakcii
b. *думный дворянин* 'noble de la Douma'
dumnyj dvorjanin

Ainsi, ДУМА ne réalise pas le sens 'pensée' dans les adjectifs doublets, il est à noter que tous les homonymes ou mots polysémiques ne construisent pas non plus des adjectifs doublets à partir de tous leurs sens. Par exemple, КУКОЛКА (KUKOLKA) prend le sens 'nymphe' lors de la formation de doublets, et non le sens 'marionnette' (215).

(215) *куколочные шкурки* 'peaux de nymphes'
kukoločnye škurki
куколочные яйца 'œufs de nymphes'
kukoločnye jajca
куколковое состояние 'état de nymphe'
kukolkovoe sostojanie

De plus, certains cas de doublets potentiels ont été vérifiés lors de la constitution et du nettoyage de la base de données RuDenom (cf. le chapitre 4). Ainsi, КАРАКУЛЕВЫЙ (KARAKULEVYJ) et КАРАКУЛЬСКИЙ (KARAKUL'SKIJ) se réfèrent tous les deux à КАРАКУЛЬ (KARAKUL') 'astrakan' (216), et non à КАРАКУЛЯ (KARAKULJA) 'gribouillage'.

(216) *каракулевая шуба*, 'manteau en astrakan'
karakulevaja šuba
каракульское овцеводство 'élevage de moutons astrakan'
karakul'skoe ovcevodstvo

9.2.2 Doublets concurrents

Le tableau 9.7 montre le nombre de doublets dont les fréquences relatives se situent entre 0.4 et 0.6 inclus, ainsi que le décompte des hapax parmi eux.

La présence de doublets hapax révèle une certaine instabilité dans la formation des adjectifs dénominaux. Alekseeva (2007, pp.90-92) associe ce phénomène à l'intégration de noms d'origine étrangère dans la langue russe. Nos données confirment cette hypothèse, avec une majorité de noms de base ayant un score étymologique négatif (9

[5]Cf. Sakhno (2011) pour une étude plus approfondie en synchronie et diachronie du mot ДУМА (DUMA).

Type	*-n-/-Ov-*	*-n-/-sk-*	*-sk-/-Ov-*
$0.4 \leq R \leq 0.6$	67	53	12
Hapax	14	22	8

Tableau 9.7: Doublets concurrents

sur 14 pour *-n-/-Ov-*, 18 sur 22 pour *-n-/-sk-*, 6 sur 8 pour *-sk-/-Ov-*). Les adjectifs dérivés semblent d'abord émerger dans le langage parlé, souvent informel, avant de se propager dans le discours public et les textes imprimés. À ce stade, leur forme peut encore être non finalisée, ce qui peut générer des doublets.

En ce qui concerne le style poétique, le choix entre les doublets adjectivaux *-n-/-Ov-* et *-sk-/-Ov-* peut être influencé par la structure syllabique des adjectifs : en effet, les doublets en *-Ov-* comprennent une syllabe supplémentaire. Ainsi, en fonction du rythme du vers, l'un des adjectifs peut être favorisé (217a-217b). Dans d'autres situations, c'est la contrainte de la rime qui détermine le choix de la forme adjectivale (217c-217d).

(217) a. АВРОРА 'aurore' → АВРОРОВЫЙ
AVRORA AVROROVYJ
авроровый цвет 'couleur d'aurore'
avrorovyj cvet
АВРОРА 'aurore' → АВРОРНЫЙ
AVRORA AVRORNYJ
научи Узреть нетленными очами, Как отрок в огненной печи Цветёт ***авроpными*** *лучами.* [М. А. Кузмин. Пещной отрок [Сны, 3] (1921)]
nauči Uzret' netlennymi očami, Kak otrok v ognennoj peči Cvetët ***avrornymi*** *lučami*
'Apprends à voir avec des yeux immortels, comment un enfant dans une fournaise ardente fleurit de rayons **d'aurore**.' [M. A. Kuzmin. Le garçon de la grotte [Rêves, 3] (1921)

b. ФАЙДЕШИН 'faïdachine' → ФАЙДЕШИНОВЫЙ
FAJDEŠIN FAJDEŠINOVYJ
файдешиновый халат 'robe de faïdachine'
fajdešinovyj xalat
ФАЙДЕШИН 'faïdachine' → ФАЙДЕШИННЫЙ
FAJDEŠIN FAJDEŠINNYJ
Мы сейчас / увяжем вас в ***файдешинный*** */ самовяз!* [С. И. Кирсанов. «Щиплет, щиплет / ноги снег...» [Моя именинная, 7] (1927)]
My sejčas / uvjažem vas v ***fajdešinnyj*** */ samovjaz*
'Nous allons maintenant / vous attacher avec un nœud coulant de **ficelle à rôtir** !' [S. I. Kirsanov. «Il pince, il pince / les pieds dans la neige...» [Mon anniversaire, 7] (1927)]

c. ГИПОТЕНУЗА ‘hypoténuse’ → ГИПОТЕНУЗНЫЙ
GIPOTENUZA GIPOTENUZNYJ
гипотенузная комната ‘chambre en hypoténuse’
gipotenuznaja komnata
ГИПОТЕНУЗА ‘hypoténuse’ → ГИПОТЕНУЗСКИЙ
GIPOTENUZA GIPOTENUZSKIJ
Под шёлком холя / прелесть их нерусскую, Укрыла Оля / грань ***гипотенузскую***. [М. А. Тарловский. «В то время с тем, кто у меня в начале...» [Веселый странник, 6] (1935)]
Pod šëlkom xolja / prelest' ix nerusskuju, Ukryla Olja / gran' ***gipotenuzskuju***
‘Sous la soie de la robe, leur beauté non russe, Olia a couvert / le bord **hypoténusien**.’ [M. A. Tarlovskij. «À cette époque avec celui qui était au début...» [Le voyageur joyeux, 6] (1935)]

d. ЯНТАРЬ ‘ambre’ → ЯНТАРСКИЙ
JANTAR' JANTARSKIJ
янтарское пиво ‘bière ambrée’
jantarskoe pivo
ЯНТАРЬ ‘ambre’ → ЯНТАРЁВЫЙ
JANTAR' JANTARËVYJ
Сок преценный, ***янтарёвый****, Дар души её суровой* [М. И. Цветаева. Встреча вторая [Царь-девица, 5] (1920)]
Sok precennyj, ***jantarëvyj****, Dar duši eë surovoj*
‘Le jus précieux, **ambré**, est le don de son âme sévère.’ [M. I. Cvetaeva. Deuxième rencontre [La princesse, 5] (1920)]

Dans le cas des doublets *-n-*/*-sk-*, où aucun des suffixes n'ajoute de syllabe supplémentaire, les contraintes de versification peuvent être moins restrictives. Par exemple, dans le cas de (218), l'adjectif ГНОМНЫЙ aurait pu être remplacé par ГНОМСКИЙ sans conséquences prosodiques significatives.

(218) ГНОМ ‘gnome’ → ГНОМСКИЙ
GNOM GNOMSKIJ
гномские руки ‘mains de gnome’
gnomskie ruki
ГНОМ ‘gnome’ → ГНОМНЫЙ
GNOM GNOMNYJ
Я тёмный дух, я ***гномный*** *царь, Минута не долга.* [К. Д. Бальмонт. Заклятие [Danses macabres] (1903)]
Ja tëmnyj dux, ja ***gnomnyj*** *car', Minuta ne dolga*
‘Je suis un esprit sombre, je suis le roi **des gnomes**, la minute n'est pas longue.’ [K. D. Bal'mont. Sortilège [Danses macabres] (1903)]

Il arrive parfois qu'un des adjectifs d'une paire de doublets développe un sens

qualificatif, en plus de son sens de relation initial (219).

(219) a. ТОРГАШ ‘commerçant’ → ТОРГАШСКИЙ
TORGAŠ TORGAŠSKIJ
торгашские мысли ‘pensées marchandes’
torgašskie mysli
ТОРГАШ ‘commerçant’ → ТОРГАШНЫЙ
TORGAŠ TORGAŠNYJ
торгашное российское правосудие ‘justice corrompue [lit. marchande] russe’
torgašnoe rossijskoe pravosudie

b. ТЕРМИНАТОР ‘Terminator’ → ТЕРМИНАТОРСКИЙ
TERMINATOR TERMINATORSKIJ
размеренный «терминаторский» шаг ‘pas mesuré «à la Terminator»’
razmerennyj «terminatorskij» šag
ТЕРМИНАТОР ‘Terminator’ → ТЕРМИНАТОРНЫЙ
TERMINATOR TERMINATORNYJ
терминаторная технология ‘technologie destructive [lit. terminator]’
terminatornaja texnologija

c. ГИППОПОТАМ ‘hippopotame’ → ГИППОПОТАМОВЫЙ
GIPPOPOTAM GIPPOPOTAMOVYJ
гиппопотамовая кожа ‘peau d’hippopotame’
gippopotamovaja koža
ГИППОПОТАМ ‘hippopotame’ → ГИППОПОТАМСКИЙ
GIPPOPOTAM GIPPOPOTAMSKIJ
гиппопотамские мерки ‘mesures exagérées [lit. d’hippopotame]’
gippopotamskie merki

Cependant, l’évolution vers un sens qualificatif n’est pas systématique. Les adjectifs doublets hapax peuvent être utilisés de manière interchangeable dans les deux contextes (220a), voire même apparaître dans le même contexte (220b)[6].

(220) a. ПЕНАЛЬТИ ‘tir au but’ → ПЕНАЛЬТИЕВЫЙ
PENAL’TI PENAL’TIEVYJ
пенальтиевый триумф ‘triomphe sur tirs au but’
penal’tievyj triumf
ПЕНАЛЬТИ ‘tir au but’ → ПЕНАЛЬТИЙНЫЙ
PENAL’TI PENAL’TIJNYJ
пенальтийные серии ‘séries de tirs au but’
penal’tijnye serii

[6]L’exemple en (220b) présente le seul cas dans la base de données de doublets où les adjectifs hapax sont employés avec le même nom recteur.

b. шпинель ‘spinelle’ → шпинелевый
ŠPINEL’ ŠPINELEVYJ
Справочники по минералогии утверждают: маггемит возникает при окислении магнетита, наследуя его ***шпинелевую*** *структуру.* [А. М. Портнов. Магнитная память о прошлых пожарах // «Химия и жизнь», 1986]
Spravočniki po mineralogii utveržḍajut: maggemit voznikaet pri okislenii magnetita, nasleḍuja ego ***špinelevuju*** *strukturu*
‘Les manuels de minéralogie affirment : la maghémite se forme par oxydation de la magnétite, héritant de sa structure **spinelle**. [A. M. Portnov. Mémoire magnétique des incendies passés // "Chimie et Vie", 1986]’
шпинель ‘spinelle’ → шпинельный
ŠPINEL’ ŠPINEL’NYJ
… что, вероятно, связано с образованием на поверхности катализатора соединений ***шпинельной*** *структуры…* [Раиса Кузьмина, Владимир Севостьянов. Каталитическая очистка газовых выбросов от оксидов азота и углерода // «Российский химический журнал», 2000]
… čto, verojatno, svjazano s obrazovaniem na poverxnosti katalizatora soedinenij ***špinel’noj*** *struktury…*
‘... ce qui est probablement lié à la formation de composés à structure **spinelle** à la surface du catalyseur... [Raisa Kuz’mina, Vladimir Sevost’janov. Nettoyage catalytique des émissions gazeuses d’oxydes d’azote et de carbone // "Journal chimique russe", 2000]’

9.2.3 Doublets occasionnels

Le tableau 9.8 illustre la distribution des doublets occasionnels, dont les fréquences relatives sont comprises entre 0.9 et 1 inclus. Il répertorie également les cas spécifiques où la fréquence absolue d’un des adjectifs est supérieure à 100, tandis que celle du second adjectif est équivalente à 1.

Type	*-n-/-Ov-*	*-n-/-sk-*	*-sk-/-Ov-*
$0.9 \leq \mathrm{R}$	193	122	24
$F_1 = 1$, $F_2 > 100$	49	25	4
$F_1 > 100$, $F_2 = 1$	31	30	3

Tableau 9.8: Doublets occasionnels

Parmi ces doublets occasionnels, les cas où l’un des adjectifs est très fréquent (plus de 100 occurrences) tandis que l’autre n’apparaît qu’une seule fois sont présents pour

tous les types de doublets[7]. La différence entre ces chiffres est particulièrement marqué pour les doublets *-n-*/*-Ov-*, avec 49 cas où l'adjectif en *-Ov-* apparaît une seule fois et l'adjectif en *-n-* plus de 100 fois, et 31 cas où l'adjectif en *-n-* apparaît une seule fois et l'adjectif en *-Ov-* plus de 100 fois. Pour les doublets *-n-*/*-sk-* et *-sk-*/*-Ov-*, les différences sont moins prononcées, mais elles sont tout de même présentes.

Comme dans le cas des doublets concurrents, un adjectif hapax (surtout formé avec le suffixe *-n-*) peut être utilisé pour traduire le sens qualificatif, contrairement à un adjectif plus fréquent et donc plus conventionnel qui est purement relationnel (avec notamment le suffixe *-sk-*). Ainsi, dans *шофёрные нервы* (*šofërnye nervy*) 'nerfs de chauffeur', le hapax ШОФЁРНЫЙ (ŠOFËRNYJ) est utilisé pour décrire des nerfs qui sont typiques ou caractéristiques d'un chauffeur, suggérant probablement une grande résistance au stress ou une capacité à rester calme dans des situations de conduite difficiles (cf. ШОФЁРСКИЙ (ŠOFËRSKIJ) (231)). La phrase *она осталась покинутой, с «кавалерным» ребёнком на руках* (*ona ostalas' pokinutoj, s «kavalernym» rebënkom na rukax*) 'elle est restée abandonnée, avec un 'enfant de son petit ami' dans les bras' contient un hapax КАВАЛЕРНЫЙ (KAVALERNYJ) qui semble être utilisé pour indiquer que l'enfant est d'une certaine manière associé à un petit ami, peut-être en suggérant que le père de l'enfant était le petit ami qui a quitté la mère (cf. КАВАЛЕРСКИЙ (KAVALERSKIJ) (110)). Le hapax БУХГАЛТЕРНЫЙ (BUXGALTERNYJ) dans *он держался в этом отношении бухгалтерного порядка* (*on deržalsja v ètom otnošenii buxgalternogo porjadka*) 'il a maintenu un ordre comptable à cet égard' est utilisé pour qualifier le type d'ordre qui est maintenu, suggérant une précision, une rigueur ou une attention aux détails qui sont typiques de la comptabilité (cf. БУХГАЛТЕРСКИЙ (BUXGALTERSKIJ) (1 580)).

Au contraire, l'adjectif hapax avec le suffixe *-sk-* peut être utilisé à la place d'un adjectif plus fréquent avec *-n-* pour renforcer une idée d'identité socio-institutionnelle. Une caractéristique du nom de base est ainsi attribuée au nom recteur non pas comme étant intrinsèquement inhérente, mais comme étant liée au nom recteur du point de vue de social, institutionnel, idéologique, politique, géographique ou ethnique (Hénault et Sakhno, 2015) (221).

(221) a. *Всю ночь дикаситься, лежит на кровати, бубнит чево-то, зубами скоргочет,* ***тюремские*** *песни поёт, свет зазря жгёт.* [Виктор Астафьев. Печальный детектив (1982-1985)]
Vsju noč' dikasitsja, ležit na krovati, bubnit čevo-to, zubami skorgočet, ***tjuremskie*** *pesni poët, svet zazrja žgët.*
'Toute la nuit, il se comporte comme un fou, allongé sur le lit, marmonnant quelque chose, grinçant des dents, chantant des chansons **de prison**, en gardant inutilement la lumière allumée.' [Victor Astaf'ev. Le détective triste (1982-1985)]

[7]Il est à noter que les exemples de hapax ci-dessous sont tirés des oeuvres anciens (19ème ou début du 20ème) ou celles qui relèvent du registre particulier.

b. *Посмотрите, с какою жадностью пожирают эту* ***литературскую*** *нечистоту все классы читателей!* [О. И. Сенковский. Похождения Чичикова, или мертвые души. Поэма Н. Гоголя (1842)]
Posmotrite, s kakoju žadnost'ju požirajut ètu ***literaturskuju*** *nečistotu vse klassy čitatelej!*
'Regardez avec quelle avidité tous les lecteurs dévorent cette saleté **littéraire** !' [O. I. Senkovskij. Les aventures de Čičikov, ou les âmes mortes. Poème de N. Gogol' (1842)]

c. *Годовая практика* ***колхозской*** *работы уже дала ряд ценных указаний на недостатки.* [М. Сысин. Основные мероприятия по укреплению строительства колхозов в Нижегородской губернии // «Нижегородский кооператор», 1928]
Godovaja praktika ***kolxozskoj*** *raboty uže dala rjad cennyx ukazanij na nedostatki.*
'Une année de pratique du travail **dans une ferme collective** a déjà fourni un certain nombre d'indications précieuses sur les lacunes.' [M. Sysin. Les principales mesures pour renforcer la construction des kolkhozes dans la province de Nižnij Novgorod // "Le coopérant de Nižnij Novgorod", 1928]

Comme pour les adjectifs doublets hapax, un nouvel adjectif peut être formé pour répondre à des exigences rythmiques et prosodiques, notamment lorsque l'adjectif existant ne convient pas à ces contraintes (222).

(222) a. *Толстый, жирный, поезд* ***пассажирный!*** [Алексей Иванов. Земля – Сортировочная (1990-1991)]
Tolstyj, žirnyj, poezd ***passažirnyj***
'Un train épais, gros, **de passagers** !' [Aleksej Ivanov. Terre - Triage (1990-1991)]

b. *Где-то там, в лесах, за Брянском, Запевают вдалеке На цыганском, на гигантском, На* ***гитарском*** *языке* [А. А. Штейнберг. Путь-дорога (1952)]
Gde-to tam, v lesax, za Brjanskom, Zapevajut vdaleke Na cyganskom, na gigantskom, Na ***gitarskom*** *jazyke*
'Quelque part là-bas, dans les forêts, au-delà de Brjansk, on commence à chanter au loin, en langue tsigane, en langue géante, en langue **de guitare**.' [A. A. Štejnberg. La route (1952)]

c. *Голодному городу всунута в рот, Зернится для кладки готовая На* ***тротуаровый*** *бутерброд — Асфальта икра осетровая.* [М. А. Зенкевич. «Подавившись обрубком дубового пня...» (1928-1929)]
Golodnomu gorodu vsunuta v rot, Zernitsja dlja kladki gotovaja Na ***trotuarovyj*** *buterbrod — Asfal'ta ikra osetrovaja.*
'Une ville affamée s'est vue mettre dans sa bouche Le caviar **d'esturgeon**

de l'asphalte – Du grain prêt à être posé pour le pavage Sur un sandwich **de trottoir**.' [M. A. Zenkevič. "En étouffant sur un bout de souche de chêne..." (1928-1929)]

On peut également observer la présence des hapax d'auteur dans le style poétique, c'est-à-dire d'adjectifs créés par un auteur particulier et utilisés uniquement dans son œuvre (223)

(223) a. *Гони буржуй / на рыбий пир — у океана в яме. Корабль / **буржуевый** / топи рабочими рублями!* [В. В. Маяковский. Подводный комсомолец (1930)]
*Goni buržuj / na rybij pir — u okeana v jame. Korabl' / **buržuevyj** / topi rabočimi rubljami!*
'Chasse les bourgeois / vers un festin de poisson – dans un trou d'océan. Coule / le navire **bourgeois** / avec avec des roubles des travailleurs !' [V. V. Maïakovski. Le sous-marinier du Komsomol (1930)]

b. *Лопушиный, **ромашный** Дом — так мало домашний! С тем особенным взглядом Душ — тяжёлого весу.* [М. И. Цветаева. Певица (1935)]
*Lopušinyj, **romašnyj** Dom — tak malo domašnij! S tem osobennym vzgljadom Duš — tjažëlogo vesu.*
'**Pelucheuse**, de couleur camomille, La maison – si peu accueillante ! Avec ce regard particulier Des âmes – lourds de poids.' [M. I. Cvetaeva. La chanteuse (1935)]

c. *Окна, светло-алы, Вступают, как фламинго, в **лампный** океан.* [Ф. К. Сологуб. Майская ночь (1923.12.21)]
*Okna, svetlo-aly, Vstupajut, kak flamingo, v **lampnyj** okean.*
'Les fenêtres, d'un rouge clair, plongent, comme des flamants roses, dans l'océan **éclairé**.' [F. K. Sologub. Nuit de mai (21/12/1923)]

Les hapax peuvent également être créés à des fins stylistiques, en particulier pour un usage métaphorique (224a-224b), ou encore à des fins humoristiques (224c), par exemple pour imiter les fautes grammaticales commises par des personnes non russophones.

(224) a. *Бабушка Аня включила **бульоновое** радио. Радио забулькало и из него полился бульон* [Бульоновое радио // «Трамвай», 1990]
*Babuška Anja vključila **bul'onovoe** radio. Radio zabul'kalo i iz nego polilsja bul'on.*
'La grand-mère Anna a allumé la radio **bouillon**. La radio s'est mise à grésiller et le bouillon en a coulé.' [Un radio bouillon // «Tramway», 1990]

b. *[...] чулки **ажурового** цвета с цветочками* [М. Л. Гаспаров. Записи и выписки (2001)]
*[...] čulki **ažurovogo** cveta s cvetočkami*

'[...] des bas de couleur **dentelle** avec des petites fleurs.' [M. L. Gasparov. Notes et extraits (2001)]

c. *Сладкий зазывный голос, напевно предлагающий отведать «**миндальский** халва».* [С. Д. Кржижановский. Салыр-Гюль (1933)]
Sladkij zazyvnyj golos, napevno predlagajuščij otvedat' «mindal'skij xalva»
'Une voix douce et séduisante, qui propose en chantant de goûter à «l'halva **d'amande**».' [S. D. Kržižanovskij. Salyr-Gul (1933)]

Dans certains cas, des doublets sont créés exprès lorsque la forme existante est généralement utilisée dans un contexte spécifique. Par exemple, l'adjectif ГЕРКУЛЕСОВЫЙ (GERKULESOVYJ) 'Hercule' est associé au nom КАША (KAŠA) 'porridge' pour former *геркулесовая каша* (*gerkulesovaja kaša*) 'porridge d'avoine'. L'adjectif ГЕРКУЛЕСНЫЙ (GERKULESNYJ) peut être alors nécessaire pour faire référence à une personne très forte (225).

(225) *[...] блистательное представление малобариста **геркулесного** жонглёра эквилибриста «Бруно фон Солерно»* [В. А. Никифоров-Волгин. Кануны Великого поста (1923-1938)]
*[...] blistatel'noe predstavlenie malobarista **gerkulesnogo** žonglëra èkvilibrista «Bruno fon Solerno»*
'[...] une performance éblouissante du jongleur équilibriste **herculéen**, "Bruno von Solerno"' [V. A. Nikiforov-Volgin. La veille du Grand Carême (1923-1938)]

La majorité des créations lexicales ont cependant la même capacité référentielle que les adjectifs attestés et fréquents (226).

(226) a. *спонсорная поддержка* 'soutien du sponsor'
sponsornaja podderžka
∼ *спонсорская поддержка*
sponsorskaja podderžka

b. *престижное лицейное образование* 'éducation prestigieuse au lycée'
prestižnoe licejnoe obrazovanie
∼ *престижное лицейское образование*
prestižnoe licejskoe obrazovanie

c. *брендный товар* 'produit de marque'
brendnyj tovar
∼ *брендовый товар*
brendovyj tovar

d. *брезентная палатка* 'tente en bâche'
brezentnaja palatka
∼ *брезентовая палатка*
brezentovaja palatka

e. *свекловые плантации* ‘plantations de betteraves’
sveklovye plantacii
∼ *свекольные плантации*
svekol'nye plantacii

9.3 Similarité contextuelle

Après avoir développé une série de commentaires qualitatifs sur la nature des doublets, en particulier des hapax, basés sur leurs fréquences relatives, nous conclurons ce chapitre avec l'analyse quantitative de leurs contextes. Comme évoqué dans la section 9.1.1, les doublets ne présentent pas une combinabilité uniforme. L'objectif de cette section sera de compléter l'analyse des fréquences relatives entre les deux adjectifs dans un couple de doublets, présentée dans la section 9.1.2, sans pour autant engager une analyse détaillée de la concurrence sémantique. Par ailleurs, nous nous concentrerons uniquement sur l'analyse de similarité contextuelle entre deux adjectifs d'un doublet, en nous appuyant sur leurs distributions, spécifiquement le nombre de noms recteurs partagés[8]. Pour ce faire, nous utiliserons l'intégralité de notre base de données de doublets, excluant les faux doublets et les doublets concurrents hapax (fréquences relatives : 0.5, fréquences absolues : 1). Comme nous l'avons précisé, parmi les doublets concurrents hapax, il n'y a qu'un seul cas où les deux adjectifs ont le même nom recteur, comme dans *шпинельная/шпинелевая структура* (*špinel'naja/špinelevaja struktura*). Par conséquent, l'inclusion de ce type de doublets ne présente pas d'intérêt pour l'analyse de la similarité contextuelle, et notamment pour l'analyse de dépendance.

L'analyse de dépendance (Mitkov, 2022, p.240) est une méthode utilisée en linguistique informatique pour analyser la structure grammaticale d'une phrase. Elle repose sur l'idée que les mots d'une phrase sont interconnectés et dépendent les uns des autres. Dans cette approche, chaque mot dépend d'un seul autre mot, sauf le verbe principal de la phrase qui est la racine de la structure. La relation qui nous intéresse plus particulièrement est celle entre le nom et l'adjectif qui modifie ce nom. Afin de pouvoir repérer les noms recteurs, nous avons extrait les contextes pour les adjectifs de la base de données des doublets à partir de RusCorpora. Pour automatiser la tâche, nous avons utilisé la bibliothèque `SpaCy` (Honnibal et Montani, 2017) en Python qui fournit une suite complète d'outils pour l'analyse de texte, y compris l'analyse de dépendance, et notamment le modèle pré-entraîné `ru_core_news_sm`.

Après avoir extrait automatiquement les noms recteurs pour les adjectifs doublets notre objectif est de comparer le nombre de noms recteurs que les adjectifs de doublets ont en commun. Pour cela, nous avons utilisé l'indice de Jaccard (Jaccard, 1901).

L'indice de Jaccard, soit le coefficient de similarité de Jaccard, est une statistique utilisée pour mesurer la similarité entre des ensembles d'échantillons.

[8]Pour les mesures de différenciation contextuelle, incluant notamment le nombre de contextes non partagés par les mots concurrents, voir Gries (2001, 2003).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

où :

$|A \cap B|$ = la taille de l'intersection des ensembles A et B,
$|A \cup B|$ = la taille de l'union des ensembles A et B

Cette métrique est exprimée comme un nombre entre 0 et 1. Un indice de Jaccard de 1 signifie que les deux ensembles sont identiques, avec tous les éléments en commun. À l'inverse, un indice de Jaccard de 0 signifie que les deux ensembles n'ont aucun élément en commun.

Le tableau 9.9 présente les moyennes des indices de Jaccard calculés en fonction des fréquences de tokens (commençant par l'ensemble de données sauf les hapax, et finissant par les couples de doublets où chaque adjectif à une fréquence supérieure à 100).

	-n-/-Ov-	*-n-/-sk-*	*-sk-/-Ov-*
$F_1 > 1$, $F_2 > 1$	0.036	0.017	0.020
$F_1 > 10$, $F_2 > 10$	0.046	0.022	0.029
$F_1 > 100$, $F_2 > 100$	0.064	0.037	0.046

Tableau 9.9: Indices de Jaccard ; doublets

Dans la mesure où les fréquences absolues des adjectifs doublets augmentent, le nombre de contextes qu'ils partagent augmente également, et vice versa. Les tendances sont les mêmes pour chaque niveau de fréquence : les doublets *-n-/-Ov-* partagent le plus grand nombre de noms recteurs (indices de Jaccard les plus élevés), tandis que les doublets *-n-/-sk-* ont le nombre de noms recteurs le plus faible. Cela signifie que les doublets *-n-/-Ov-* s'emploient dans les mêmes contextes plus souvent que les doublets avec les autres suffixes ; les doublets *-n-/-Ov-* sont ainsi plus susceptibles d'être employés d'une manière interchangeable.

Une analyse plus approfondie de l'ensemble de données des doublets sauf les hapax (ligne F_1 `> 1,` F_2 `> 1` dans le tableau 9.9) révèle que les indices de Jaccard les plus élevés pour les doublets *-n-/-Ov-* sont supérieurs à 0.1. Par contre, dans les données *-n-/-sk-* et *-sk-/-Ov-* les indices de Jaccard sont systématiquement inférieurs à 0.1.

Les données présentées dans le tableau 9.10 suggèrent que pour les doublets *-n-/-Ov-* qui partagent le plus grand nombre de noms recteurs, les fréquences absolues peuvent varier considérablement, allant de fréquences relativement faibles, comme dans le cas de КАРМИНОВЫЙ (KARMINOVYJ) à des fréquences extrêmement élevées, comme dans le cas de СНЕЖНЫЙ (SNEŽNYJ).

Cependant, l'examen des fréquences relatives révèle une tendance distincte : malgré un grand nombre de contextes partagés, un adjectif est généralement beaucoup plus utilisé que l'autre, avec une fréquence relative comprise entre 0.7 et 1.0. De plus, dans

Nom	A1	F1	R1	A2	F2	R2	J
графит *grafit* 'graphite'	*графитный* *grafitnyj*	57	0.2	*графитовый* *grafitovyj*	239	0.8	0.103
жасмин *žasmin* 'jasmin'	*жасминный* *žasminnyj*	55	0.3	*жасминовый* *žasminovyj*	140	0.7	0.141
кармин *karmin* 'carmin'	*карминный* *karminnyj*	85	0.6	*карминовый* *karminovyj*	49	0.4	0.100
озон *ozon* 'ozone'	*озонный* *ozonnyj*	57	0.1	*озоновый* *ozonovyj*	457	0.9	0.121
парусина *parusina* 'toile à voile'	*парусинный* *parusinnyj*	102	0.1	*парусиновый* *parusinovyj*	716	0.9	0.102
пурпур *purpur* 'pourpre$_N$'	*пурпурный* *purpurnyj*	1 370	0.7	*пурпуровый* *purpurovyj*	512	0.3	0.120
сафьян *saf'jan* 'safran (cuir)'	*сафьянный* *saf'jannyj*	226	0.4	*сафьяновый* *saf'janovyj*	351	0.6	0.177
смородина *smorodina* 'groseille'	*смородинный* *smorodinnyj*	91	0.3	*смородиновый* *smorodinovyj*	175	0.7	0.143
снег *sneg* 'neige'	*снежный* *snežnyj*	14 944	1.0	*снеговой* *snegovoj*	716	0.0	0.128
яблоня *jablonja* 'pommier'	*яблонный* *jablonnyj*	53	0.1	*яблоневый* *jablonevyj*	518	0.9	0.112

Tableau 9.10: Indices de Jaccard > 0.1 ; *-n-*/*-Ov-*

la plupart de ces cas, l'adjectif le plus fréquent est celui qui est formé avec le suffixe *-Ov-*.

Le tableau 9.11 présente les indices de Jaccard les plus élevés pour les doublets *-n-/-sk-*. Comme nous l'avons mentionné, ces indices ne dépassent pas 0.01. Du point de vue des fréquences relatives, les adjectifs avec *-n-* sont plus fréquents. Comme dans le cas des doublets *-n-/-Ov-*, on observe une variation très élevée des fréquences absolues.

Nom	A1	F1	R1	A2	F2	R2	J
интеллигент *intelligent* 'intellectuel$_N$'	*интеллигентный* *intelligentnyj*	8 077	0.8	*интеллигентский* *intelligentskij*	1 769	0.2	0.086
абонент *abonent* 'abonné$_N$'	*абонентный* *abonentnyj*	74	0.0	*абонентский* *abonentskij*	2 444	1.0	0.060
чародей *čarodej* 'sorcier'	*чародейный* *čarodejnyj*	113	0.6	*чародейский* *čarodejskij*	68	0.4	0.044
лилипут *liliput* 'lilliputien'	*лилипутный* *liliputnyj*	5	0.1	*лилипутский* *liliputskij*	53	0.9	0.043
абонемент *abonement* 'abonnement'	*абонементный* *abonementnyj*	217	1.0	*абонементский* *abonementskij*	6	0.0	0.039

Tableau 9.11: Indices de Jaccard les plus élevés ; *-n-/-sk-*

Finalement, le tableau 9.12 référence les indices de Jaccard les plus élevés pour les doublets *-sk-/-Ov-*. Ces indices sont les plus faibles, les adjectifs doublets sont alors moins susceptibles de s'employer d'une manière interchangeable. Les fréquences relatives sont les plus extrêmes (de 0.9 à 1.0), en faveur du suffixe *-Ov-* pour les indices les plus élevés (0.068).

Conclusion

Dans ce chapitre, nous avons exploré les doublets sous divers angles. L'analyse de la distribution des propriétés des noms de base a montré que lorsqu'un nom autorise la formation d'adjectifs avec deux suffixes différents, ce nom combine généralement les propriétés des noms qui ne sont associés qu'à un seul suffixe. Bien que les spectres de fréquence des doublets soient assez similaires, c'est le suffixe *-Ov-* qui construit le plus de hapax dans les paires *-n-/-Ov-* et *-sk-/-Ov-*. En ce qui concerne les fréquences relatives des adjectifs et de leurs noms de base, tous les adjectifs doublets suivent les mêmes tendances que celles observées dans le corpus RuDenom, mais ces tendances

Nom	A1	F1	R1	A2	F2	R2	J
государь	*государский*	527	0.1	*государевый*	3 219	0.9	0.068
gosudar'	*gosudarskij*			*gosudarevyj*			
'souverain'							
кашемир	*кашемирский*	29	0.1	*кашемировый*	295	0.9	0.068
kašemir	*kašemirskij*			*kašemirovyj*			
'cachemire'							
ад	*адский*	4 421	1.0	*адовый*	129	0.0	0.028
ad	*adskij*			*adovyj*			
'enfer'							
ангора	*ангорский*	118	1.0	*ангоровый*	5	0.0	0.025
angora	*angorskij*			*angorovyj*			
'angora'							
гжель	*гжельский*	86	1.0	*гжелевый*	4	0.0	0.024
gžel'	*gžel'skij*			*gželevyj*			
'gjel (technique)'							

Tableau 9.12: Indices de Jaccard les plus élevés ; *-sk-*/*-Ov-*

sont encore plus prononcées : les adjectifs doublets peuvent être considérés comme davantage décomposables que les adjectifs simples. Enfin, l'examen des fréquences relatives entre les adjectifs au sein de chaque couple de doublets a mis en évidence le fait qu'il existe peu de doublets en variation libre (dont les fréquences sont assez proches) dans les trois paires, mais une préférence marquée pour l'utilisation d'un adjectif par rapport à l'autre au sein de ces paires.

Notre étude s'est concentrée sur les doublets dont le sens référentiel est proche. Nous avons donc écarté les doublets ayant des sens distincts. Lors de l'analyse des vrais doublets, nous avons constaté que le choix entre les deux formes peut être conditionné par les spécificités du style poétique, ainsi que par des nécessités stylistiques (dans les métaphores ou dans un but humoristique). De plus, un adjectif moins conventionnel peut être éventuellement créé pour mettre en évidence une propriété inhérente d'un nom de base avec le suffixe *-n-*, ou le caractère d'identité avec le suffixe *-sk-*. Cependant, dans de nombreux cas, l'existence de doublets n'est pas conditionnée par des facteurs externes.

L'analyse de similarité contextuelle a révélé que ce sont les doublets *-n-*/*-Ov-* qui partagent le plus grand nombre de contextes, notamment de noms recteurs, pour les cas où les fréquences relatives sont plutôt extrêmes, mais indépendamment des fréquences absolues des adjectifs. Ces doublets sont donc plus susceptibles d'être utilisés de manière interchangeable que les doublets *-n-*/*-sk-* ou *-sk-*/*-Ov-*.

Conclusion générale et perspectives

L'objectif principal de cette thèse a été d'étudier la concurrence entre les suffixes *-n-*, *-sk-*, et *-Ov-* utilisés pour former des adjectifs dénominaux. Afin de parvenir à une analyse approfondie, nous avons adopté une méthodologie quantitative tripartite. En premier lieu, nous avons examiné comment les propriétés du nom de base (phonologiques, morphologiques, sémantiques ou étymologiques) influencent le choix du suffixe. Ensuite, nous avons analysé les fréquences absolues et relatives des adjectifs par rapport à leurs noms de base, tout en évaluant la productivité des trois suffixes selon divers critères établis dans la littérature. Enfin, nous avons porté notre attention sur les adjectifs doublets, prenant en compte les propriétés des noms de base, leurs fréquences relatives et les noms recteurs qui leur sont associés.

Notre recherche s'est basée sur une base de données exhaustive RuDénom qui contient plus de 12 000 adjectifs issus des données de RusCorpora. Comme nous nous sommes limité à l'analyse des adjectifs construits à partir des noms, ce corpus contient uniquement les adjectifs, leurs noms de base, ainsi que leurs fréquences respectives. Nous avons exploité l'intégralité de cette base pour l'étude des fréquences et de la productivité. Pour l'analyse spécifique des adjectifs doublets, un sous-corpus dédié a été élaboré. Par ailleurs, notre exploration des propriétés des noms de base s'est concentrée sur les adjectifs de haute et de basse fréquence (notamment des hapax). La distinction entre ces deux sous-corpus s'est imposée en raison de notre désir de mettre en évidence les éventuelles différences dans les mécanismes de construction du lexique établi et des nouvelles créations lexicales.

Comme nous l'avons remarqué plus haut, les propriétés des noms de base se sont articulés autour de quatre axes linguistiques : phonologique, morphologique, sémantique et étymologique. La plupart des propriétés que nous avons annotées sont catégorielles et correspondent à des conventions bien définies (par exemple, le nombre de syllabes équivaut au nombre de voyelles ; le genre se rapporte à l'un des trois genres en russe – masculin, féminin, neutre ; les allomorphies vocaliques correspondent à la présence ou l'absence de la voyelle mobile dans le paradigme nominal, etc.). Cependant, l'application d'une approche catégorielle à la sémantique et à l'étymologie se révèle plus

délicate. Malgré l'introduction de la classification catégorielle pour ces deux propriétés, nous avons proposé deux approches pour les quantifier. La quantification sémantique a été réalisée à l'aide de modèles vectoriels ; quant à l'étymologie, nous avons calculé les scores étymologiques sur la base de bigrammes. Contrairement à la dernière méthode qui a réussi à éliminer complètement les biais liés à l'encodage manuel, l'approche quantitative à la sémantique a permis de réduire significativement ce même biais.

Toutes les propriétés des noms de base, qu'elles soient catégorielles ou non, ont été examinées pour déterminer leur impact sur le choix entre les trois suffixes dans une analyse multivariée basées sur les arbres. L'interaction des propriétés sémantiques, du dernier phonème des radicaux et de la longueur des noms de base en syllabes s'est révélée être un discriminant clé pour les trois suffixes dans les données de haute fréquence. La sémantique est un prédicteur significatif pour le suffixe -*sk*-, tandis que les suffixes -*n*- et -*Ov*- sont distingués par les phonèmes finaux et le nombre de syllabes. En ce qui concerne les données de basse fréquence, les mêmes propriétés sont pertinentes, avec l'ajout des allomorphies consonantiques, qui contribuent notamment à distinguer le suffixe -*n*-.

Les modèles basées sur les arbres ont démontré une excellente performance sur les données de haute fréquence ; toutefois, leur efficacité est moindre sur les données de basse fréquence. Ce constat confirme l'hypothèse d'une plus grande variabilité et de moins de contraintes dans la création de néologismes et d'occasionalismes. En outre, notre étude a révélé que la représentation sémantique vectorielle est plus pertinente pour le choix des suffixes que la sémantique catégorielle. En ce qui concerne l'étymologie, bien que l'étymologie non catégorielle ait une influence plus significative sur le choix du suffixe adjectival, elle ne permet pas de discriminer efficacement entre les trois suffixes examinés.

Nous avons ensuite mené une analyse de la productivité des suffixes. Les différentes méthodes de mesure de la productivité ont donné des résultats similaires lorsqu'elles étaient appliquées aux affixes en anglais et en russe, avec une hypothèse initiale qu'un affixe était plus productif que l'autre. Toutefois, dans le contexte de notre étude, les trois suffixes sont considérés comme productifs en synchronie, les diverses approches d'évaluation de la productivité ont ainsi conduit à des résultats contradictoires. Néanmoins, la mesure de productivité la plus robuste – celle de la productivité potentielle proposée par Baayen (2009) – classe la productivité des suffixes comme suit : -*sk*- plus productif que -*Ov*-, -*Ov*- plus productif que -*n*-. Cependant, la base de données RuDenom comprend des adjectifs formés à partir de noms propres et de noms communs ; lors de l'analyse multivariée nous avons observé que le suffixe -*sk*- est largement préféré si la base nominale est un nom propre. Nous avons donc évalué la productivité des suffixes uniquement sur la base des adjectifs dérivés de noms communs. Cette analyse a conduit à un classement différent et a fait ressortir le suffixe -*Ov*- comme le plus productif, et -*sk*- comme le moins productif. Ainsi, la productivité du suffixe -*sk*- semble se restreindre principalement au domaine des noms propres, tandis que le suffixe -*Ov*- est le plus productif dans le contexte des noms communs.

Notre étude finale s'est centrée sur les doublets. Nous en avons conclu que les noms constituant les doublets partagent des propriétés communes, typiques pour ces deux suffixes. La majorité des doublets est formée avec les suffixes -*n*- et -*Ov*-, la plus faible quantité de doublets a été observée -*sk*- et -*Ov*-. En ce qui concerne les fréquences, dans les trois cas, un des adjectifs doublets est nettement plus fréquent que l'autre, indiquant une préférence marquée pour une forme spécifique, la variation libre entre les deux étant moins courante. L'analyse des noms recteurs a révélé que ce sont principalement les doublets formés avec les suffixes -*n*- et -*Ov*- qui partagent le plus grand nombre de noms recteurs en communs et peuvent donc être utilisés d'une manière interchangeable.

Notre étude aboutit à la conclusion que la concurrence entre les suffixes -*n*- et -*Ov*- s'avère la plus complexe, tant au niveau des propriétés des noms de base qu'en termes de nombre de doublets et du volume de leurs noms recteurs partagés. Il semble donc exister une corrélation positive entre l'intensité de la concurrence de deux schémas dérivationnels et le nombre de doublets qu'ils construisent, ainsi que le degré d'interchangeabilité de ces derniers. Le suffixe -*sk*-, quant à lui, se distingue nettement des deux autres, avec une préférence marquée pour les noms propres et une productivité essentiellement réservée à ces derniers. Il forme moins de doublets avec les autres suffixes et, dans les cas où ces doublets existent, il partage moins de noms recteurs en commun, particulièrement avec le suffixe -*Ov*-. Les adjectifs doublets formés avec -*sk*- se distinguent donc davantage sur le plan sémantique des adjectifs formés avec les deux autres suffixes.

La principale contribution de notre recherche a été la quantification des différents aspects de la concurrence entre les suffixes -*n*-, -*sk*- et -*Ov*-. Les éléments discutés dans cette thèse ouvrent, cependant, plusieurs voies pour de futures recherches.

Premièrement, comme la sémantique des noms de base constitue un élément clé pour la distinction des suffixes, notamment le suffixe -*sk*-, il serait pertinent d'expérimenter nos modèles avec un spectre plus large de classes de la sémantique catégorielle (donc explorer des catégories plus fines), en recourant par exemple à la liste des relations sémantiques dans WordNet (Balkova *et al.*, 2004). Les mêmes données de WordNet pourraient potentiellement être utilisées pour une autre méthode de quantification de la sémantique des noms, et notamment pour calculer les voisins en fonction du nombre de nœuds qu'ils partagent dans le graphe. De plus, l'extension de la méthode distributionnelle à l'ensemble des adjectifs et de leurs noms de base, représente une autre voie de recherche pour explorer la concurrence entre les suffixes du point de vue des distances entre les noms de base et les adjectifs construits (Bonami et Guzmán Naranjo, 2023).

Notre étude a également démontré que l'introduction de variables non catégorielles, illustrée par les exemples de la sémantique et de l'étymologie, conduit à de meilleures performances et permet de traiter de manière plus précise un plus grand nombre de noms. Il serait donc pertinent d'étendre l'application de cette méthodologie non catégorielle à d'autres propriétés afin de comparer leur impact. Par exemple, les

propriétés phonologiques pourraient être représentées par des scores de similarité, grâce au calcul des distances de Levenstein (Strnadová, 2014) ou à l'utilisation de modèles analogiques (Chapman et Skousen, 2005).

Par ailleurs, notre approche pourrait bénéficier de l'apport des études psycholinguistiques. En premier lieu, une enquête sur le degré d'acceptabilité des adjectifs hapax par les locuteurs natifs pourrait s'avérer enrichissante. En outre, l'examen des fréquences relatives des adjectifs et de leurs noms de base, ainsi que la décomposabilité des adjectifs pourrait être complété par des expériences où les sujets sont demandés à évaluer la complexité des formes construites.

De manière plus générale, notre analyse de la concurrence s'est limitée uniquement à trois suffixes : *-n-*, *-sk-* et *-Ov-*. Toutefois, il serait enrichissant d'explorer la concurrence entre, par exemple, *-sk-* et *-Ovsk-*, ou entre *-ičn-* et *-ičesk-*. Ces derniers sont aussi très compétitifs et sont également capables de construire des adjectifs doublets.

Finalement, notre étude n'a fait que quelques explorations préliminaires dans l'analyse des adjectifs doublets, dont l'examen en profondeur s'impose toujours. La différence entre doublets étant minime, en plus des propriétés des noms de base et des fréquences absolues et relatives il conviendrait d'étudier la date de première attestation des deux formes. Une comparaison plus approfondie des sens des noms de base qui sont mobilisés lors de la construction des adjectifs doublets serait nécessaire. Une analyse plus détaillée de la sémantique des noms modifiés par les adjectifs doublets viendrait compléter notre recherche, en particulier la différenciation contextuelle de ces derniers. De plus, une exploration plus exhaustive de l'inventaire des relations sémantiques et pragmatiques véhiculées par les adjectifs doublets est également requise.

Annexes

A1 Structure de la base de données RuDénom

Colonne	Description et types de données
`Nom`	Nom de base
`FreqN`	Fréquence totale du nom de base (corpus général, des journaux, multimédia, oral, poétique)
`SourceN`	Étymologie du nom de base (`autre` = origine étrangère ; `slave` = origine slave)
`BigramN`	Bigrammes du nom de base
`Score_etym`	Score étymologique du nom de base
`ClSem_p`	Classe sémantique des noms propres (`1` = propre ; `0` = commun)
`ClSem_h`	Classe sémantique des noms humains/animés (`1` = humain/animé ; `0` = non-humain/non-animé)
`ClSem_c`	Classe sémantique des noms concrets (`1` = concret ; `0` = non-concret)
`ClSem_a`	Classe sémantique des noms abstraits (`1` = abstrait ; `0` = non-abstrait)
`Score_p`	Distance moyenne des 5 voisins les plus proches de la classe `ClSem_p`
`Score_h`	Distance moyenne des 5 voisins les plus proches de la classe `ClSem_h`
`Score_c`	Distance moyenne des 5 voisins les plus proches de la classe `ClSem_c`
`Score_a`	Distance moyenne des 5 voisins les plus proches de la classe `ClSem_a`
`SyllN`	Longueur des noms de base en syllabes (`1` = une syllabe, `2` = deux syllabes, etc. jusqu'à `5+` = cinq syllabes et plus)
`AccSyllN`	Position de l'accent dans la forme du NOM.SG. (`d` = dernière syllabe, `ad` = avant dernière syllabe ; `aad+` avant avant dernière syllabe et plus loin)
`AccZal`	Types accentuels d'après Zaliznjak (`0` = nom indéclinables ; de `a` à `f` : pour les noms déclinables)

`ClFlex`	Classes flexionnelles canoniques (`I` = première ; `II` = deuxième ; `III` = troisième)
`ClFlexZal`	Classes flexionnelles d'après Zalizniak (`0` = nom indéclinables ; de `1` à `8` : pour les noms déclinables)
`Genre`	Genres du nom de base (`m` = masculin ; `f` = féminin ; `n` = neutre)
`AllomV`	Allomorphie vocalique dans les thèmes du nom de base (`1` = présente ; `0` = absente)
`AllomC`	Allomorphies consonantiques dans les thèmes du nom de base (`0` = absentes ; `mouil` : mouillure ; `palat` = palatalisation)
`Segment`	Allomorphies segmentales des thèmes du nom de base (`0` = absentes ; `ajout` = ajout ; `suppr` = suppression ; `inf` = interférence ; `rempl` = modification)
`Adjectif`	Adjectif dénominal
`Contexte`	Contexte pour les hapax
`FreqA`	Fréquence totale des adjectifs (corpus général, des journaux, multimédia, oral, poétique)
`Radical`	Radical des adjectifs
`DPhoR`	Dernier phonème des radicaux (`cDent` = consonne dentale ; `cAlv` = consonne alvéolaire ; `cLab` = consonne labiale ; `cVel` = consonne vélaire ; `Vow` = voyelle)
`SuffDerA`	Suffixe dérivationnel des adjectifs (`-n-`, `-sk-`, `-Ov-`)

Tableau 9.13: Structure de la base de données des adjectifs dénominaux

Il convient de noter que nous utilisons également les conventions `ClSem` et `Score_sem` pour faire référence à l'ensemble des cinq classes et des cinq scores sémantiques.

La base de données est disponible à cette adresse : https://github.com/eliswind/PhD.

A2 Guide d'annotation sémantique

Ce guide répertorie les étiquettes sémantiques et fournit des exemples non exhaustifs pour chacune d'entre elles.

Noms propres

Cette étiquette est utilisée pour des noms propres non humains (toponymes et des objets célestes, des organisations) et les noms propres humains (noms des personnes, des animaux, des idéalités, des familles ou des dynasties).

Les toponymes et les objets célestes : Антарктида 'Antarctique', Бали 'Bali', Бриансон 'Brianson', Гольфстрим 'Gulf Stream', Дарданеллы 'Dardanelles', Евфрат 'Euphrates', Луна 'Lune', Марс 'Mars', Монмартр 'Montmartre'.

Les organisations spécifiques : Динамо 'Dynamo', Зингер 'Zinger', Икея 'Ikéa', Любэ 'Ljubè (groupe musical)', Майкрософт 'Microsoft', Норникель 'Nornickel', Рено 'Renault'.

Les noms propres des humains, réels ou des personnages de fiction : Артур 'Arthur', Жизель 'Giselle', Икар 'Icare', Кастанеда 'Castaneda', Ксенофан 'Xénophane', Ландау 'Landau', Маргарита 'Marguerite', Маугли 'Maugli', Торичелли 'Torricelli'.

Les noms propres faisant référence à des familles ou des dynasties : Альморавид 'Almoravide', Альмохад 'Almohade', Кеянид 'Kayanide', Сун 'Songhai'.

Les noms propres d'animaux : Баюн 'Bajun (chat)', Пегас 'Pegasus', Рыжик 'Ryžik (chien)'.

Les idéalités peuvent également être identifiées en tant que noms propres : Валькирья 'Valkyrie', Пандора 'Pandore', Перун 'Péroun'.

Noms d'êtres humains/animés

Les noms d'êtres humains ou non humains animés incluent des noms de peuples et des gentilés, des métiers et des titres, des liens de parenté, des qualificateurs, des animaux et des personnages fictifs.

Les noms des peuples et des gentilés : аморей 'Amoréen', ацтек 'Aztèque', ашкеназ 'Achéen', бритт 'Breton', конголезец 'Congolais', мелькит 'Melkite', мосараб 'Mozarabe', пермяк 'Permien', халкидонит 'Halkidonite', хорват 'Croate'.

Les noms de métiers ou des titres : адвокат 'avocat', атаман 'ataman', воевода 'voïvode', друид 'druide', император 'empereur', кадет 'cadet', парикмахер 'coiffeur', сторож 'gardien'.

Noms désignant des liens de parentés : брат 'frère', прадед 'grand-père'.

Les qualificateurs : вольнодум 'libéral', обжора 'glouton', ретроград 'rétrograde', ублюдок 'vaurien', эстет 'esthète'.

Les noms d'animaux : БАБОЧКА 'papillon', ВОРОБЕЙ 'moineau', ГУСЕНИЦА 'chenille', КОНЬ 'cheval', ЛОСОСЬ 'saumon', СКОРПИОН 'scorpion'.

Les noms désignant des personnages fictifs, humanoïdes ou animés : АНГЕЛ 'ange', БОГ 'dieu', ДРАКОН 'dragon', ДЬЯВОЛ 'diable', ОБОРОТЕНЬ 'loup-garou'.

La première méthode permettant de déterminer si un nom est animé consiste à comparer les formes de l'accusatif singulier, du génitif singulier et du nominatif singulier.[9] :

- Si ACC.SG. est égal à GEN.SG., alors le nom est animé
  ДРАКОН 'dragon' : *дракон-а*$_{\text{ACC.SG}}$ = *дракон-а*$_{\text{GEN.SG}}$;
- Si ACC.SG. est égal à NOM.SG., alors le nom est non animé
  ВИРУС 'virus' : *вирус*$_{\text{ACC.SG}}$ = *вирус*$_{\text{NOM.SG}}$.

La deuxième méthode consiste en vérification du choix de l'interrogatif :

- ***Кто*** *это?* 'Qui est-ce?' – *Это дракон* 'C'est un dragon'
- ***Что*** *это?* 'Qu'est-ce que c'est?' – *Это вирус* 'C'est un virus'

Noms concrets

La méthode principale utilisée pour déterminer si un nom est concret consiste à le placer dans un contexte particulier (Haas *et al.*, 2022).

- Ce contexte consiste en l'utilisation du verbe *находиться* 'se trouver' et d'un complément de localisation (par exemple : ***стол*** *находится в саду* 'la **table** se trouve dans le jardin')
  De plus :
  - Pour les noms concrets comptables : le contexte implique un nom dénotant une unité de mesure (par exemple : ***стол*** *шириной два метра* 'une **table** de deux mètres de large').
  - Pour les noms massifs : le contexte approprié implique l'utilisation d'un nom dénotant une unité de mesure combiné avec le nom en question (par exemple : *два килограмма муки* 'deux kilos de **farine**').
  - Enfin, pour les noms collectifs : le contexte inclut soit un nom au génitif (***стая*** *волков* 'une **meute** de loups'), soit une unité de mesure (*семья из пяти человек* 'une **famille** de cinq personnes').

[9] L'inconvénient de cette méthode est qu'elle n'est applicable qu'aux noms masculins et aux noms féminins au génitif pluriel. Cf. plus de détails dans la section 5.2.4.

Les noms d'objets naturels qui dénotent les entités situées dans l'espace : АПАТИТ 'apatite', ГОРА 'montagne', ОЗЕРО 'lac', РАВНИНА 'plaine', РИФ 'récif', РУБИН 'rubis', ТАЙГА 'taïga', ЯНТАРЬ 'ambre'.

Les noms d'artefacts : АРКА 'arche', БАРРИКАДА 'barricade', БАХЧА 'bacche', БИБЛИОТЕКА 'bibliothèque', ВИНО 'vin', ВИСКОЗА 'viscose', ГАРАЖ 'garage', ГИТАРА 'guitare', КАБИНЕТ 'bureau', КОНЬЯК 'cognac', КРЕЙСЕР 'croiseur', КУПОЛ 'dôme', МАЙОНЕЗ 'mayonnaise', ПАТРОН 'cartouche', САМОВАР 'samovar', СИГАРА 'cigare', ТРОН 'trône', ТЮЛЬ 'taffetas'.

Les noms de substances munies d'une extension dans l'espace : АТМОСФЕРА 'atmosphère', ВОЗДУХ 'air', ДЫМ 'fumée', КУНЖУТ 'sésame', МОЛОКО 'lait', ПЛАЗМА 'plasma', СНЕГ 'neige', ТАЛЬК 'talc', ЧАЙ 'thé', ЭЛАСТАН 'élastique'.

Les noms collectifs qui dénotent les individus constitués d'une pluralité interne : ЭСКАДРОН 'escadron', БАТАЛЬОН 'bataillon', ТОЛПА 'foule', СТАДО 'troupeau', КАРАВАН 'caravane', БРИГАДА 'brigade'.

Cependant, certains noms collectifs ne répondent pas aux critères requis et ne peuvent pas être utilisées dans des constructions telles que 'un nom collectif + une unité de mesure + un autre nom[10]' : ЗНАТЬ 'noblesse', ЧЕЛЯДЬ 'domestiques', ЭЛЕКТОРАТ 'électorat'. Ces noms sont considérés comme abstraits.

Les noms démunis d'autonomie référentielle, noms localisateurs spaciaux ou temporels : БЕРЕГ 'bord', ВЕРХУШКА 'sommet', КОНЕЦ 'fin', НАЧАЛО 'début', ПЕРИФЕРИЯ 'périphérie', СЕРЕДИНА 'milieu', ЮГ 'sud'.

Les noms d'idéalités concrètes dénombrables qui concernent la langue, la musique, la mathématique, la géométrie. Ce sont des noms non sensibles, car en absence d'exécution, ces idéalités concrètes ne sont pas accessibles aux sens[11] : АККОРД 'accord', ВЕКТОР 'vecteur', ВЕРТИКАЛЬ 'vertical', ГИПОТЕНУЗА 'hypoténuse', ГЛАГОЛ 'verbe', КВАДРАТ 'carré', ЛОЗУНГ 'slogan', ОКСЮМОРОН 'oxymore', ПУНКТИР 'pointillé', СКАЗКА 'conte', СКАЛЯР 'scalaire', ЭПИЛОГ 'épilogue', ТОККАТА 'toccata', ФРАЗЕМА 'phrasème'.

Les nes noms de microorganismes (entités biologiques non doués d'intentionnalité) sont exclus de la catégorie 'humain/animé' ; ils font partie de la catégorie 'concret' : АДИПОЦИТ 'adipocyte', ВИРУС 'virus'.

Noms abstraits

Cette catégorie d'étiquettes englobe l'ensemble des noms qui ne sont pas considérés comme des noms propres, qui ne sont pas animés ou inanimés, et qui ne sont pas concrets. Elle comprend notamment les idéalités indénombrables, les qualités, les

[10]Contrairement aux noms collectifs qui dénotent les individus constitués d'une pluralité interne, ils ne s'emploient pas au pluriel, n'introduisent pas des limites, ne visent pas des totalités closes, n'admettent pas les déterminants quantitatifs.

[11]Ces noms ne font pas partie de la catégorie 'concret' selon Haas *et al.* (2022) ; ils font partie d'une classe sémantique à part d'*objet cognitif*. Nous incluons ces noms dans la catégorie 'concret' suivant Flaux et Van de Velde (2000).

sentiments, les facultés, les manières, les états, les actions, les activités, les quantités, les périodes ainsi que les objets financiers.

Les idéalités indénombrables : АРТХАУС ‘cinéma d’auteur’, ГРИЗАЙЛЬ ‘grisaille’, ИКОНОПИСЬ ‘icônographie’, КАПОЭЙРА ‘capoeira’, НАУКА ‘science’, РОКОКО ‘rococo’, ТРАГЕДИЯ ‘tragédie’[12].

Les noms des facultés physiques et mentales : АНАЛИЗ ‘analyse’, ВОЛЯ ‘volonté’, ГОЛОС ‘voix’, ПАМЯТЬ ‘mémoire’, СЛУХ ‘ouïe’, УМ ‘intelligence’.

Les noms d’état, d’action et d’activité : АНОНС ‘annonce’, АПАРТЕИД ‘aparté’, БОБСЛЕЙ ‘bobsleigh’, ВЫГОВОР ‘sermon’, ВЫДАЧА ‘distribution’, ГУЛЯНКА ‘fête’, ДЕФИЛЕ ‘défilé’, ИГРА ‘jeu’, МАРШ ‘marche’, ПЕНАЛЬТИ ‘penalty’, РАЛЛИ ‘rallye’, СКУКА ‘ennui’, ЦИГУН ‘tai-chi’.

Les quantités exactes, approximatives et métaphoriques : ГРАММ ‘gramme’, ДЮЙМ ‘pouce’, ЕДИНИЦА ‘unité’, КЛАСС ‘classe’, КУСОК ‘morceau’, КУЧА ‘tas’, ЛИТР ‘litre’, МАСШТАБ ‘échelle’, МЕТР ‘mètre’, МИНУТА ‘minute’, ПАРА ‘paire’, ПИНТА ‘pinte’, ТРИМЕСТР ‘trimestre’, ФРАГМЕНТ ‘fragment’.

Les périodes : АРХЕЙ ‘Archéen’, ВЕСНА ‘printemps’, ДЕНЬ ‘jour’, ГОД ‘année’, ИЮЛЬ ‘juillet’, ПОНЕДЕЛЬНИК ‘lundi’.

Les objets financiers : ДРАХМА ‘drachme’, РУБЛЬ ‘rouble’, СТЕРЛИНГ ‘sterling’, ТЕНГЕ ‘téngé’, ТУГРИК ‘togriq’, ФРАНК ‘franc’.

Les langues et cultures : ГРАВЕТТ ‘gravettien’, ЗУЛУ ‘zoulou’, ИПИУТАК ‘ipiutak’, ФАРСИ ‘persan’, САНСКРИТ ‘sanskrit’.

Catégories hétérogènes

Outre les noms collectifs précédemment cités, d’autres catégories peuvent présenter des hétérogénéités dans l’étiquetage concret et abstrait. C’est notamment le cas des noms désignant des maladies et les pathologies. Dans le cas de ГЕМАТОМА ‘hématome’ ou КАРЦИНОИД ‘carcinome’, il s’agit des pathologies localisables et mesurables (donc étiquetées comme concrètes), tandis que dans le cas de ВЕТРЯНКА ‘varicelle’ ou КОНЪЮНКТИВИТ ‘conjonctivite’, il s’agit de maladies dont les symptômes sont localisables et mesurables, mais pas la maladie en elle-même (étiquetées comme abstraites).

De plus, certains mots polysémiques peuvent être étiquetés comme concrets ou abstraits selon le contexte[13]. Dans le cas des hapax, le contexte permet la désambiguïsation facile et l’étiquetage unique d’un mot. Dans d’autres cas, les entrées ont été doublées et chacune d’entre elles a été annotée différemment au niveau sémantique (КОРЕНЬ ‘racine (de l’arbre, de problème)’, СЛОВАРЬ ‘dictionnaire / lexique’, ОБЛАСТЬ ‘région / domaine’).

[12]Flaux et Van de Velde (2000) les classifient par défaut dans la catégorie des noms concrets indénombrables, bien qu’elle souligne que la distinction entre concret et abstrait dépend du contexte. Nous choisissons de les étiqueter comme abstraits, car ces noms ne satisfont pas les critères permettant de les identifier comme concrets.

[13]Cf. la section 5.2.4.

A3 Guide d'annotation étymologique

L'annotation étymologique permet de faire la distinction entre les termes ayant une origine slave et ceux provenant d'autres langues. Toutefois, dans ce contexte, l'origine slave est définie de manière restrictive et se limite aux noms issus des langues du groupe slave orientale – russe, biélorusse, ukrainien – ayant pour origine le vieux slavon (227a) ainsi qu'aux termes d'origine indoeuropéenne (227b).

(227) a. ДАЛЬ 'étendue' < slavon **dalъ*,
DAL'
МОЛНИЯ 'éclair' < slavon **mъlni*,
MOLNIJA
ПИЩА 'nourriture' < slavon *pitia̯*
PIŠČA

b. ГРИВНА 'hryvnia (monnaie)' < indoeur. *grīva* 'nuque'
GRIVNA

L'annotation repose sur des sources lexicographiques telles que les dictionnaires en ligne (Semënov, 2003 ; Šanskij, 2004 ; Fasmer, 2006 ; Krylov, 2008) ainsi que le Wiktionnaire russe[14].

Les noms qui étaient déjà attestés dans le vieux slave mais qui ont été empruntés dans cette période sont annotés comme étrangers (228).

(228) САПФИР 'saphir' < lat. *sapphires*
SAPFIR

Les toponymes situés à l'intérieur de la Fédération de Russie qui désignent des régions habitées par des peuples non-slaves et dont l'origine n'est pas non plus slave sont étiquetés comme étrangers (229a). La même logique a été appliquée pour les noms désignant les objets de la vie quotidienne et l'héritage culturel et des peuples non slaves (229b).

(229) a. Зея 'Zeïa (Russie)' < évène
Zeja
Кимры 'Kimry (Russie)' < finno-ougrien
Kimry
Майкоп 'Maïkop (Russie)' < adyguéen
Majkop
Стерлитамак 'Sterlitamak (Russie)'< bachkir
Sterlitamak

b. ТАЗ 'bassine' < tatar de Crimée
TAZ

[14] Accessible via l'URL `https://ru.wiktionary.org/`.

ТАЙГА ‘taïga’ < altaïque
TAJGA
ШАНЬГА ‘changa (plat)’ < komi
ŠAN’GA

Les mots dont l’origine étymologique est incertaine ou controversée ont été marqués comme étant d’origine slave (230).

(230) ДУРМАН ‘datura’ < slavon/turque
DURMAN
КОЗЫРЬ ‘atout’ < slavon/polonais/turque
KOZYR’

Les mots dérivés, dans lesquels le lexème de base provient d’une langue étrangère et l’affixe est d’origine slave ont été annotés en fonction de l’origine de leur lexème de base (231).

(231) БЕЙСБОЛКА ‘casquette’ < anglais
BEJSBOLKA
БУРЖУЙКА ‘poêle’ < français)
BURŽUJKA

Les noms composés ont été étiquetés en fonction de l’origine de l’élément le plus long. Ainsi, МЕТЕО**ЧУВСТВИТЕЛЬНОСТЬ** (METEOČUVSTVITEL’NOST’) ‘météosensibilité’ est annoté comme slave, ТЕХ**ОСМОТР** (TEXOSMOTR) ‘contrôle technique’ – slave aussi. En cas d’égalité de longueur entre les deux lexèmes de base, le nom composé a été étiqueté en fonction de l’origine du premier élément : БИОТОК (BIOTOK) ‘flux biotique’ est considéré comme étranger.

Bibliographie

Adams, V. (2014). *Complex Words in English*. London/New York : Routledge.

Alekseeva, E. (2011). *Адъективные новообразования в современном русском языке [Nouveaux adjectifs dérivés dans la langue russe contemporaine]*. Thèse de doctorat, Sankt-Peterburgskij gosudarstvennyj universitet.

Alekseeva, E. V. (2007). «Образование новых прилагательных: наблюдения и размышления [La formation de nouveaux adjectifs : observations et réflexions]». *XXXVI Meždunarodnaja filologičeskaja konferencija*, 19, 13–17.

Allassonnière-Tang, M., Brown, D. et Fedden, S. (2021). «Testing semantic dominance in Mian gender: three machine learning models». *Oceanic Linguistics*, 60(2), 302–334.

Anderson, S. R. (1992). *A-Morphous Morphology*. Cambridge : Cambridge University Press.

Antic, E. (2012). «Relative frequency effects in Russian morphology». In Gries, S. T. et Divjak, D. (eds.), *Frequency Effects in Language Learning and Processing*, 83–107. Berlin : De Gruyter Mouton.

Antipina, O. P. (2012a). *Сопоставительный анализ паронимов русского и английского языков [Analyse comparative des paronymes en russe et en anglais]*. Thèse de doctorat, Baškirskij gosudarstvennyj universitet.

Antipina, O. P. (2012b). «Структурно-словообразовательная характеристика адъективных паронимов русского и английского языков [Caractéristique structurelle et morphologique des adjectifs paronymes en russe et en anglais]». *Vestnik Čeljabinskogo gosudarstvennogo universiteta*, 23, 14–17.

Apresjan, Y. (1974). *Лексическая семантика: Синонимические средства языка [Sémantique lexicale : Les moyens synonymiques de la langue]*. Moskva : Nauka.

Arapov, M., Efimova, E. et Šrejder, J. (1975). «О смысле ранговых распределений [Sur le sens des distributions de rangs]». *Naučno-texničeskaja informacija*, 2, 9–20.

Archaimbault, S. (1992). «L'adjectif dans la tradition grammaticale russe». *Histoire, Épistémologie, Langage*, 14(1), 211–221.

Arndt-Lappe, S. (2014). «Analogy in suffix rivalry: the case of English *-ity* and *-ness*». *English Language & Linguistics*, 18(3), 497–548.

Aronoff, M. (1976). *Word Formation in Generative Grammar*. Cambridge : MIT Press.

Aronoff, M. (2007). «In the beginning was the word». *Language*, 83(4), 803–830.

Aronoff, M. (2016). «Competition and the lexicon». In Elia, A., Iacobini, C. et Voghera, M. (eds.), *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della Società Linguistica Italiana*, 39–52. Roma : Bulzoni Editore srl.

Aronoff, M. (2019). «Competitors and alternants in linguistic morphology». In Rainer, F., Gardani, F., Dressler, W. U. et Luschützky, H. C. (eds.), *Competition in Inflection and Word-Formation*, 39–66. Berlin : Springer.

Aronoff, M. et Anshen, F. (2017). «Morphology and the lexicon: lexicalization and productivity». In Spencer, A. et Zwicky, A. M. (eds.), *The Handbook of Morphology*, 237–247. Hoboken : John Wiley & Sons, Inc.

Aronoff, M. et Cho, S. (2001). «The semantics of *-ship* suffixation». *Linguistic Inquiry*, 167–173.

Aronoff, M. et Fuhrhop, N. (2002). «Restricting suffix combinations in German and English: closing suffixes and the monosuffix constraint». *Natural Language & Linguistic Theory*, 451–490.

Aronoff, M. et Lindsay, M. (2014). «Productivity, blocking and lexicalization». In Lieber, R. et Štekauer, P. (eds.), *The Oxford Handbook of Derivational Morphology*, 67–83. Oxford : Oxford Handbooks.

Arsenijevic, B., Boleda Torrent, G., Gehrke, B. et McNally, L. (2010). «Ethnic adjectives are proper adjectives». In Baglini, R., Grinsell, T., Keane, J., Singerman, A. R. et Thomas, J. (eds.), *46th Annual Meeting of the Chicago Linguistic Society*, 46, 17–30. Chicago : University of Chicago.

Arutjunova, N. (1961). *Очерки по словообразованию в современном испанском языке [Esquisses sur la dérivation en espagnole contemporain]*. Moskva : AN SSSR.

Arutjunova, N. (1999). *Язык и мир человека [La langue et le monde de l'homme]*. Moskva : Jazyki russkoj kul'tury.

Aslanoff, S. (1986). *Manuel typographique du russiste*. Paris : IES.

Azarx, J. (1987). «О синонимии однокореневых слов [Sur la synonymie des mots dérivés d'une même racine]». *Derivacija i istorija jazyka*, 64–73.

Baayen, R. H. (1992). «Quantitative aspects of morphological productivity». In Booij, G. et Marle, J. (eds.), *Yearbook of Morphology 1991*, 109–149. Berlin : Springer.

Baayen, R. H. (1993). «On frequency, transparency and productivity». In Booij, G. et Marle, J. (eds.), *Yearbook of Morphology 1992*, 181–208. Berlin : Springer.

Baayen, R. H. (1994). «Derivational productivity and text typology». *Journal of Quantitative Linguistics*, 1(1), 16–34.

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht : Kluwer Academic Publishers.

Baayen, R. H. (2008). *Analyzing Linguistic Data : a Practical Introduction to Statistics Using R*. New York : Cambridge University Press.

Baayen, R. H. (2009). «Corpus linguistics in morphology: morphological productivity». In Lüdeling, A. et Merja, K. (eds.), *Corpus Linguistics. An International Handbook*, 900–919. Berlin : De Gruyter Mouton.

Baayen, R. H. et del Prado Martín, F. M. (2005). «Semantic density and past-tense formation in three Germanic languages». *Language*, 72(1), 666–698.

Baayen, R. H., Endresen, A., Janda, L. A., Makarova, A. et Nesset, T. (2013). «Making choices in Russian: pros and cons of statistical methods for rival forms». *Russian Linguistics*, 37(3), 253–291.

Baayen, R. H. et Lieber, R. (1991). *Productivity and English Derivation: a Corpus-Based Study*. Berlin/New York : Walter de Gruyter.

Baayen, R. H. et Renouf, A. (1996). «Chronicling the times: productive lexical innovations in an English newspaper». *Language*, 72(1), 69–96.

Baeskow, H. (2012). «-*ness* and -*ity*: phonological exponents of *n* or meaningful nominalizers of different adjectival domains?». *Journal of English Linguistics*, 40(1), 6–40.

Balkova, V., Sukhonogov, A. et Yablonsky, S. (2004). «Russian WordNet». In Sojka, P., Pala, K., Smrž, P., Fellbaum, C. et Vossen, P. (eds.), *Proceedings of the Second Global Wordnet Conference*, 100, 31–38. Brno : Masaryk University.

Bally, C. (1944). *Linguistique générale et linguistique française*. Bern : A. Frankle.

Bartning, I. et Noailly, M. (1993). «Du relationnel au qualificatif : flux et reflux». *L'information grammaticale*, 58(1), 27–32.

Bauer, L. (1983). *English Word-Formation*. Cambridge : Cambridge University Press.

Bauer, L. (2001). *Morphological Productivity*. Cambridge : Cambridge University Press.

Bauer, L., Lieber, R. et Plag, I. (2015). *The Oxford Reference Guide to English Morphology*. Oxford : Oxford University Press.

Beliakov, V. (2014). *Introduction à la lexicologie et à la sémantique lexicale russes*. Toulouse : Presses universitaires du Mirail.

Benninger, C. (2001). «Une meute de loups/une brassée de questions : collection, quantification et métaphore». *Langue française*, 129, 21–34.

Bloomfield, L. (1933). *Language*. London : Allen & Unwin.

Bobkova, N. (2022a). «Interpreting statistical models for denominal adjective formation in Russian». *The Prague Bulletin of Mathematical Linguistics*, 118, 5–23.

Bobkova, N. (2022b). «Statistical modelization of suffixal rivalry in Russian: adjectival formations in *-sk-* and *-n-*». *Corpus*, 23.

Bobkova, N. et Montermini, F. (2019). «Suffix rivalry in Russian: what low frequency words tell us». In Audring, J., Koutsoukos, N. et Manouilidou, C. (eds.), *Mediterranean Morphology Meetings*, 12, 1–17. Patras : University of Patras.

Bobkova, N. et Montermini, F. (2021). «Suffixal variation in Russian denominal adjectives of Slavic and foreign origin» [communication orale]. *Internationalisms in Slavic as a Window into the Architecture of Grammar*, Université de Graz, 24–26 février 2021.

Bobkova, N. et Montermini, F. (2023). «A quantitative approach to doublets in Russian denominal adjective construction». *Word Structure*, 16(1), 63–86.

Bojanowski, P., Grave, E., Joulin, A. et Mikolov, T. (2017). «Enriching word vectors with subword information». *Transactions of the Association for Computational Linguistics*, 5, 135–146.

Boleda, G. (2020). «Distributional semantics and linguistic theory». *Annual Review of Linguistics*, 6, 213–234.

Bonami, O. et Boyé, G. (2003). «Supplétion et classes flexionnelles». *Langages*, 152, 102–126.

Bonami, O. et Boyé, G. (2005). «Construire le paradigme d'un adjectif». *Recherches linguistiques de Vincennes*, 34, 77–98.

Bonami, O., Boyé, G. et Kerleroux, F. (2009). «L'allomorphie radicale et la relation flexion-construction». In Fradin, B., Kerleroux, F. et Plénat, M. (eds.), *Aperçus de morphologie du français*, 103–125. Saint-Denis : Presses universitaires de Vincennes.

Bonami, O. et Guzmán Naranjo, M. (2023). «Distributional evidence for derivational paradigms». In Kotowski, S. et Plag, I. (eds.), *The Semantics of Derivational Morphology: Theory, Methods, Evidence*, 219–258. Berlin/New York : Walter de Gruyter.

Bonami, O. et Pellegrini, M. (2022). «Derivation predicting inflection: a quantitative study of the relation between derivational history and inflectional behavior in Latin». *Studies in Language*, 46(4), 753–792.

Bonami, O. et Thuilier, J. (2019). «A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*». *Word Structure*, 12(1), 4–41.

Bonami, O. et Tribout, D. (2021). «Échantinom: a hand-annotated morphological lexicon of French nouns». In Namer, F., Hathout, N., Lignon, S., Ševčíková, M. et Žabokrtskỳ, Z. (eds.), *International Workshop on Resources and Tools for Derivational Morphology*, 42–51. Nancy : Université de Lorraine.

Booij, G. (2010). *Construction Morphology*. Oxford : Oxford University Press.

Booij, G. (2012). *The Grammar of Words: an Introduction to Linguistic Morphology*. Oxford : Oxford University Press.

Boyé, G. et Plénat, M. (2015). «L'Allomorphie radicale dans les lexèmes adjectivaux en français : le cas des adverbes en *-ment*». In Alsina, E. B., Lloret, M.-R. et Altimiras, J. M. (eds.), *Understanding Allomorphy: Perspectives from Optimality Theory*, 70–106. London : Equinox Publishing.

Bréal, M. (1904). *Essai de sémantique (science des significations)*. Paris : Hachette.

Breiman, L. (2001). «Random forests». *Machine Learning*, 45, 5–32.

Breuillard, J. et Viellard, S. (2015). *Histoire de la langue russe : des origines au XVIIIe siècle*. Paris : IES.

Brezina, V. (2021). «Classical monofactorial (parametric and non-parametric) tests». In Paquot, M. et Gries, S. T. (eds.), *A Practical Handbook of Corpus Linguistics*, 473–503. Berlin : Springer.

Butterworth, B. (1983). «Lexical representation». *Language Production*, 2, 257–294.

Bybee, J. L. (1985). *Morphology: a Study of the Relation between Meaning and Form*. Amsterdam : John Benjamins.

Bybee, J. L. (1995). «Diachronic and typological properties of morphology and their implications for representation». In Feldman, L. B. (ed.), *Morphological Aspects of Language Processing*, 225–246. New Jersey : Lawrence Erlbaum Associates.

Bybee, J. L. (2002). «Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change». *Language Variation and Change*, 14(3), 261–290.

Chapman, D. et Skousen, R. (2005). «Analogical modeling and morphological change: the case of the adjectival negative prefix in English». *English Language & Linguistics*, 9(2), 333–357.

Chen, T. et Guestrin, C. (2016). «XGBoost: a scalable tree boosting system». In Chen, T. et Guestrin, C. (eds.), *Proceedings of the 22 International Conference on Knowledge Discovery and Data Mining*, 785–794. New York : ACM.

Chitashvili, R. et Baayen, R. H. (1993). «Word frequency distributions». In Altmann, L. et Hřebíček, L. (eds.), *Quantitative Linguistics*, 54–135. Trier : Wissenschaftlicher Verlag.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge : MIT Press.

Chovanová, I. (2011). *Morphologie constructionnelle du slovaque et éléments de comparaison avec le français : les adjectifs dénominaux construits par composition et dérivation*. Thèse de doctorat, Université Nancy 2.

Comtet, R. (1995). «L'école phonologique de Léningrad et l'école phonologique de Moscou». *Histoire, épistémologie, langage*, 17(2), 183–209.

Corbett, G. et Comrie, B. (2003). *The Slavonic Languages*. London/New York : Routledge.

Corbett, G. G. (1987). «The morphology/syntax interface: evidence from possessive adjectives in Slavonic». *Language*, 63(2), 299–345.

Corbett, G. G. (1991). *Gender*. Cambridge : Cambridge University Press.

Corbett, G. G. (1995). «Slavonic's closest approach to Suffix Copying: the possessive adjective». *Double Case: Agreement by Suffixaufnahme*, 265–282.

Corbett, G. G. (2004). «The Russian adjective: a pervasive yet elusive category». *Adjective Classes: A Cross-Linguistic Typology*, 1, 199–222.

Corbin, D. (2012). *Morphologie dérivationnelle et structuration du lexique*. Berlin : De Gruyter Mouton.

Dal, G., Fradin, B., Grabar, N., Namer, F., Lignon, S., Plancq, C., Zweigenbaum, P. et Yvon, F. (2008). «Quelques préalables au calcul de la productivité des règles constructionnelles et premiers résultats.». In Durand, J., Habert, B. et Laks, B. (eds.), *Congrès Mondial de Linguistique Française - CMLF'08*, 1525–1538. Paris : Institut de Linguistique Française.

Dal, G., Hathout, N., Lignon, S., Namer, F. et Tanguy, L. (2018). «Toile versus dictionnaires : les nominalisations du français en -*age* et en -*ment*». In Neveu, F., Harmegnies, B., Hriba, L. et Prévost, S. (eds.), *Congrès Mondial de Linguistique Française - CMLF'18*, 46. Paris : Institut de Linguistique Française.

Dal, G., Hathout, N. et Namer, F. (2004). «Morphologie constructionnelle et traitement automatique des langues : le projet MorTAL». *Lexique 16/La formation des mots: horizons actuels*, 16, 199–231.

Dal, G. et Namer, F. (2012). «Faut-il brûler les dictionnaires? Ou comment les ressources numériques ont révolutionné les recherches en morphologie». In Neveu, F., Muni Toke, V., Blumenthal, P., Klingler, T., Ligas, P., Prévost, S. et Teston-Bonnard, S. (eds.), *Congrès Mondial de Linguistique Française - CMLF'12*, 1, 1261–1276. Paris : Institut de Linguistique Française.

Dal, G. et Namer, F. (2016). «À propos des occasionnalismes». In Neveu, F., Bergounioux, G., Côté, M.-H., Fournier, J.-M., Hriba, L. et Prévost, S. (eds.), *Congrès Mondial de Linguistique Française - CMLF'16*, 27. Paris : Institut de Linguistique Française.

Dement'ev, A. (1974). «О так называемых 'интерфиксах' в русском языке [Sur les dits 'interfixes' en russe]». *Voprosy jazykoznanija*, 4, 116–120.

Dič, N. (2003). «О текстах XIX века в Национальном корпусе русского языка [Sur les textes du XIXe siècle dans le Le corpus national de la langue russe]». *Nacional'nyj korpus russkogo jazyka*, 89–93.

Divjak, D. et Arppe, A. (2013). «Extracting prototypes from exemplars. What can corpus data tell us about concept representation?». *Cognitive Linguistics*, 24(2), 221–274.

Dressler, W. U., Merlini Barbaresi, L., Schwaiger, S., Ransmayr, J., Sommer-Lolei, S. et Korecky-Kröll, K. (2019). «Rivalry and lack of blocking among Italian and German diminutives in adult and child language». In Rainer, F., Gardani, F., Dressler, W. U. et Luschützky, H. C. (eds.), *Competition in Inflection and Word-Formation*, 123–143. Berlin : Springer.

Endresen, A., Janda, L. A., Kuznetsova, J., Lyashevskaya, O., Makarova, A., Nesset, T. et Sokolova, S. (2012). «Russian 'purely aspectual' prefixes: not so 'empty' after all?». *Scando-Slavica*, 58(2), 231–291.

Evert, S. (2004). «A simple LNRE model for random character sequences». In Purnelle, G., Fairon, C. et Dister, A. (eds.), *Proceedings of JADT*, 1, 411–422. Louvain : PUL.

Evert, S. et Baroni, M. (2007). «zipfR: word frequency distributions in R». In Gupta, S., Nenkova, A. et Jurafsky, D. (eds.), *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 29–32. Stroudsburg : Association for Computational Linguistics.

Fabre, C. et Lenci, A. (2015). «Distributional semantics today. Introduction to the special issue». *Revue TAL*, 56(2), 7–20.

Fasmer, M. (2006). *Этимологический словарь русского языка [Dictionnaire étymologique du russe]*. Moskva : Progress.

Fedden, S., Guzmán Naranjo, M. et Corbett, G. G. (2021). «Typological richness of the German gender system revealed by data mining» [communication orale]. *Third International Symposium of Morphology (ISMo 2021)*, Université de Toulouse 2 Jean Jaurès, 22–24 septembre 2021.

Firth, J. R. (1957). «A synopsis of linguistic theory, 1930-1955». In Firth, J. R. (ed.), *Studies in Linguistic Analysis*, 10–32. Oxford : Basil Blackwell.

Flaux, N. et Van de Velde, D. (2000). *Les noms en français : esquisse de classement*. Paris : Ophrys.

Fradin, B. (2003). *Nouvelles approches en morphologie*. Paris : PUF.

Fradin, B. (2008). «Les adjectifs relationnels et la morphologie». In Fradin, B. (ed.), *La raison morphologique. Hommage à Danielle Corbin*, 69–92. Amsterdam : John Benjamins Publishing.

Fradin, B. (2014). «La variante et le double». In Villoing, F., Leroy, S. et David, S. (eds.), *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, 109–147. Paris : Presses Universitaires de Paris Ouest.

Fradin, B. (2016). «L'interprétation des nominalisations en N-*age* et N-*ment* en français». In Buchi, E., Chauveau, J.-P. et Pierrel, J.-M. (eds.), *Actes du XXVIIe congrès international de linguistique et philologie romanes*, 3, 53–66. Strasbourg : ELiPhi.

Fradin, B. (2017). «The multifaceted nature of denominal adjectives». *Word Structure*, 10(1), 27–53.

Fradin, B. (2019). «Competition in derivation: what can we learn from French doublets in *-age* and *-ment*?». In Rainer, F., Gardani, F., Dressler, W. U. et Luschützky, H. C. (eds.), *Competition in Inflection and Word-Formation*, 67–93. Berlin : Springer.

Fraser, N. M. et Corbett, G. G. (1995). «Gender, animacy, and declensional class assignment: a unified account for Russian». In Booij, G. et Marle, J. (eds.), *Yearbook of Morphology 1994*, 123–150. Berlin : Springer.

Friedman, J. H. (2001). «Greedy function approximation: a gradient boosting machine». *Annals of Statistics*, 1189–1232.

Gaeta, L. et Ricca, D. (2006). «Productivity in Italian word formation: a variable-corpus approach». *Linguistics*, 44(1), 57–89.

Gardani, F., Rainer, F. et Luschützky, H. C. (2019). «Competition in morphology: a historical outline». In Rainer, F., Gardani, F., Dressler, W. U. et Luschützky, H. C. (eds.), *Competition in Inflection and Word-Formation*, 3–36. Berlin : Springer.

Garde, P. (1998). *Grammaire russe. Phonologie et morphologie*. Paris : IES.

Gause, G. F. (1934). *The Struggle for Existence*. Baltimore : Williams and Wilkins.

Gehrke, B. et McNally, L. (2015). «Distributional modification: the case of frequency adjectives». *Language*, 91(4), 837–870.

Goes, J. (1999). *L'adjectif entre verbe et nom*. Bruxelles : Duculot.

Graudina, L., Ickovič, V. et Katalinskaja, L. (2001). *Грамматическая правильность русской речи: Стилистический словарь вариантов [La précision grammaticale de la langue russe : Dictionnaire stylistique des variantes]*. Moskva : Nauka.

Graščenkov, P. (2022). *Грамматика прилагательного. Типология адъективности и атрибутивности [Grammaire de l'adjectif. Typologie de l'adjectivité et de l'attributivité]*. Moskva : LitRes.

Gries, S. T. (2001). «A corpus-linguistic analysis of English *-ic* vs *-ical* adjectives». *Icame Journal*, 25(1), 65–108.

Gries, S. T. (2003). «Testing the sub-test: an analysis of English *-ic* and *-ical* adjectives». *International Journal of Corpus Linguistics*, 8(1), 31–61.

Gries, S. T. (2013). *Statistics for Linguistics with R*. Berlin/Boston : Walter de Gruyter.

Gries, S. T. (2016). *Quantitative Corpus Linguistics with R: a Practical Introduction*. London/New York : Routledge.

Gries, S. T. (2019). «On classification trees and random forests in corpus linguistics: some words of caution and suggestions for improvement». *Corpus Linguistics and Linguistic Theory*, 16(3), 617–647.

Grišina, E. (2003). «Устная речь в Национальном корпусе русского языка [La langue orale dans le Corpus national de la langue russe]». *Nacional'nyj korpus russkogo jazyka*, 94–110.

Grišina, E. (2005). «Два новых проекта для Национального корпуса: мультимедийный корпус и подкорпус названий [Deux nouveaux projets pour le Corpus national : le corpus multimédia et le sous-corpus des appelations]». *Nacional'nyj korpus russkogo jazyka: 2003-2005. Rezul'taty i perspektivy*, 233–250.

Grišina, E. (2009). «Мультимедийный русский корпус (МУРКО): проблемы и аннотации [Un corpus multimédia russe (MURKO) : problèmes et annotations]». *Nacional'nyj korpus russkogo jazyka 2006-2008: novye rezul'taty i perspektivy*, 175–213.

Guiraud-Weber, M. (2004). *Le verbe russe : temps et aspect*. Publications de l'Université de Provence.

Guzmán Naranjo, M. (2019). *Analogical Classification in Formal Grammar*. Berlin : Language Science Press.

Guzmán Naranjo, M. (2020). «Analogy, complexity and predictability in the Russian nominal inflection system». *Morphology*, 30(3), 219–262.

Guzmán Naranjo, M. et Bonami, O. (2021). «Overabundance and inflectional classification: quantitative evidence from Czech». *Glossa: a Journal of General Linguistics*, 6(1).

Haas, P., Barque, L., Huyghe, R. et Tribout, D. (2022). «Pour une classification sémantique des noms en français appuyée sur des tests linguistiques». *Journal of French Language Studies*, 1–30.

Halle, M. (1973). «Prolegomena to a theory of word formation». *Linguistic Inquiry*, 4, 3–16.

Harris, Z. S. (1954). «Distributional structure». *Word*, 10(2-3), 146–162.

Harwood, F. W. et Wright, A. M. (1956). «Statistical study of English word formation». *Language*, 32(2), 260–273.

Haspelmath, M. et Sims, A. (2013). *Understanding Morphology*. London/New York : Routledge.

Hathout, N. (2009). *Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie*. Habilitation à diriger des recherches, Université Toulouse le Mirail-Toulouse II.

Hathout, N. (2011). «Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en *anti-*». In Roché, M., Boyé, G., Hathout, N., Lignon, S. et Plénat, M. (eds.), *Des unités morphologiques au lexique*, 251–318. Paris : Lavoisier.

Hathout, N., Montermini, F. et Tanguy, L. (2008). «Extensive data for morphology: using the World Wide Web». *Journal of French Language Studies*, 18(1), 67–85.

Hathout, N. et Tanguy, L. (2002). «Webaffix: discovering morphological links on the WWW». In González Rodríguez, M. et Suarez Araujo, C. P. (eds.), *Language Resources and Evaluation Conference 2002*, 1799–1804. Las Palmas de Gran Canaria : ELRA.

Hay, J. (2001). «Lexical frequency in morphology: is everything relative?». *Linguistics*, 39(6), 1041–1070.

Hay, J. (2004). *Causes and Consequences of Word Structure*. London/New York : Routledge.

Hay, J. et Baayen, R. H. (2002). «Parsing and productivity». In Booij, G. et Van Marle, J. (eds.), *Yearbook of Morphology 2001*, 203–235. Berlin : Springer.

Hénault, C. et Sakhno, S. (2015). «Чем супермаркет-*н*-ый лучше супермаркет-*ск*-ого? Словообразовательная синонимия в русских адъективных неологизмах по данным Интернета [Pourquoi *supermarket-n-yj* est-il meilleur que *supermarket-sk-ij* ? La synonymie dérivationnelle dans les néologismes adjectivaux russes sur la base de données d'Internet]». In Tošovic, B. et Wonisch, A. (eds.), *Wortbildung und Internet*, 107–124. Graz : Institut für Slawistik.

Hilpert, M. et Blasi, D. E. (2021). «Fixed-effects regression modeling». In Paquot, M. et Gries, S. T. (eds.), *A Practical Handbook of Corpus Linguistics*, 505–533. Berlin : Springer.

Honnibal, M. et Montani, I. (2017). «spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing», `https://spacy.io/`.

Huguin, M. (2021). «The MoNoPoli database». In Namer, F., Hathout, N., Lignon, S., Ševčíková, M. et Žabokrtskỳ, Z. (eds.), *International Workshop on Resources and Tools for Derivational Morphology*, 76–85. Nancy : Université de Lorraine.

Huyghe, R. et Varvara, R. (2023). «Affix rivalry: theoretical and methodological challenges». *Word Structure*, 16(1), 1–23.

Huyghe, R. et Wauquier, M. (2021). «Distributional semantics insights on agentive suffix rivalry in French». *Word Structure*, 14(3), 354–391.

Jaccard, P. (1901). «Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines». *Bulletin de la Société vaudoise des sciences naturelles*, 37, 241–272.

Jackendoff, R. (1975). «Morphological and semantic regularities in the lexicon». *Language*, 51, 639–671.

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Berlin : Springer.

Janda, L. A. et Lyashevskaya, O. (2011). «Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian». *Cognitive Linguistics*, 22(4), 719–763.

Kapatsinski, V. (2010). «Velar palatalization in Russian and artificial grammar: constraints on models of morphophonology». *Laboratory Phonology*, 1(2), 361–393.

Karpacheva, O. (2000). «Secondary stress in Russian compound words: evidence from poetic metrics». *The Linguistic Review*, 27, 387–478.

Khan, M. Y., Qayoom, A., Nizami, M. S., Siddiqui, M. S., Wasi, S. *et al.* (2021). «Automated prediction of Good Dictionary EXamples (GDEX): a comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques». *Complexity*, 2021, 1–18.

King, D., Sims, A. et Elsner, M. (2020). «Interpreting sequence-to-sequence models for Russian inflectional morphology». In Ettinger, A., Jarosz, G. et Pater, J. (eds.), *Proceedings of the Society for Computation in Linguistics 2020*, 409–418. New York : Association for Computational Linguistics.

Kleiber, G. (1981). *Problèmes de référence : descriptions définies et noms propres*. Metz : CAS.

Kopotev, M., Lyashevskaya, O. et Mustajoki, A. (2017). «Russian challenges for quantitative research». In Kopotev, M., Lyashevskaya, O. et Mustajoki, A. (eds.), *Quantitative Approaches to the Russian Language*, 3–29. London/New York : Routledge.

Krylov, G. (2008). *Этимологический словарь русского языка [Dictionnaire étymologique du russe]*. Saint-Pétersbourg : Viktorija pljus.

Krysin, L. (1965). «О причинах лексического заимствования [Sur les raisons de l'emprunt lexical]». *Russkij jazyk v škole*, 3, 11–15.

Krysin, L. (2008). «Структурные и функциональные свойства иноязычных неологизмов [Propriétés structurelles et fonctionnelles des néologismes étrangers]». In Bondarko, A., Kustova, G. et Rozina, R. (eds.), *Dinamičeskie modeli : slovo, predloženie, tekst : Sbornik statej v čest' E.V. Padučevoj*, 482–488. Moskva : Jazyki slavjanskix kul'tur.

Kustova, G., O, L., Padučeva, E. et Raxilina, E. (2005). «Семантическая разметка лексики в национальном корпусе русского языка: принципы, проблемы, перспективы [Le marquage sémantique du lexique dans le Corpus national de la langue russe : principes, problèmes, perspectives]». *Nacional'nyj korpus russkogo jazyka: 2003-2005. Rezul'taty i perspektivy*, 155–174.

Kustova, G. I. (2018). «Прилагательные [Adjectifs]». *Materialy k korpusnoj grammatike russkogo jazyka*, 3, 40–107.

Kutuzov, A. et Kuzmenko, E. (2017). «WebVectors: a toolkit for building web interfaces for vector semantic models». In Ignatov, D. I., Khachay, M. Y., Labunets, V. G., Loukachevitch, N., Nikolenko, S. I., Panchenko, A., Savchenko, A. V. et Vorontsov, K. (eds.), *Analysis of Images, Social Networks and Texts: 5th International Conference 2016*, 155–161. Berlin : Springer.

Kuznetsova, J. (2015). *Linguistic Profiles: Going from Form to Meaning via Statistics*. Berlin, München, Boston : De Gruyter Mouton.

Lehrer, A. (2000). «Are affixes signs? The semantic relationships of English derivational affixes». *Amsterdam Studies in the Theory and History of Linguistic Science*, 4, 143–154.

Lenci, A. (2018). «Distributional models of word meaning». *Annual Review of Linguistics*, 4, 151–171.

Levshina, N. (2015). *How to Do Linguistics with R. Data Exploration and Statistical Analysis*. Amsterdam : Benjamins.

Levshina, N. (2021). «Conditional inference trees and random forests». In Paquot, M. et Gries, S. T. (eds.), *A Practical Handbook of Corpus Linguistics*, 611–643. Berlin : Springer.

Lignon, S. (2013). «-*iser* and -*ifier* suffixations in French: verify data to verize hypotheses?. In Hathout, N., Montermini, F. et Tseng, J. (eds.), *Morphology in Toulouse. Selected Proceedings of Décembrettes 7*, 119–132. München : LINCOM Europa.

Lignon, S. et Plénat, M. (2009). «Échangisme suffixal et contraintes phonologiques». In Fradin, B., Kerleroux, F. et Plénat, M. (eds.), *Aperçus de morphologie du français*, 65–81. Saint-Denis : Presses Universitaires de Vincennes.

Lindsay, M. et Aronoff, M. (2013). «Natural selection in self-organizing morphological systems». In Hathout, N., Montermini, F. et Tseng, J. (eds.), *Morphology in Toulouse. Selected Proceedings of Décembrettes 7*, 133–153. München : LINCOM Europa.

Ljaševskaja, O., Plungjan, V. et Sičinava, D. (2003). «О морфологическом стандарте национального корпуса русского языка [Sur la norme morphologique du Corpus national de la langue russe]». *Nacional'nyj korpus russogo jazyka*, 111–135.

Ljaševskaja, O. et Šarov, S. (2009). *Частотный словарь современного русского языка: на материалах Национального корпуса русского языка [Dictionnaire de fréquence du russe contemporain : basé sur les matériaux du Corpus national de la langue russe]*. Azbukovnik.

Lopatin, V. (1977). *Русская словообразовательная морфемика [Morphématique de la dérivation en russe]*. Moskva : Nauka.

Lotka, A. J. (1925). *Elements of Physical Biology*. Baltimore : Williams and Wilkins.

Marchand, H. (1960). *The Categories and Types of Present-Day English Word-Formation*. Wiesbaden : Otto Harrassowitz.

Matthews, M. (1974). *An Introduction to the Theory of Word Structure*. Cambridge : Cambridge University Press.

Maurel, J.-P. (1993). «Des adjectifs de relation en latin». *L'information grammaticale*, 58(1), 23–26.

McNally, L. et Boleda, G. (2004). «Relational adjectives as properties of kinds». In Bonami, O. et Cabredo Hofherr, P. (eds.), *Empirical Issues in Syntax and Semantics*, 5, 179–196.

Mel'čuk, I. (1993). *Cours de morphologie générale, volume 1*. Montréal/Paris : Presses de l'Université de Montréal/CNRS Editions.

Mezhevich, I. (2002). «English compounds and Russian relational adjectives». In Morrison, G. S. et Zsoldos, L. (eds.), *Proceedings of the Northwest Linguistic Conference 2002*, 95–114. Burnaby : Simon Fraser University Linguistics Graduate Student Association.

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013). «Efficient estimation of word representations in vector space [workshop poster]». In Bengio, Y. et LeCun, Y. (eds.), *International Conference on Learning Representations 2013*.

Miller, G. A. et Charles, W. G. (1991). «Contextual correlates of semantic similarity». *Language and Cognitive Processes*, 6(1), 1–28.

Missud, A. et Villoing, F. (2020). «The morphology of rival *-ion*, *-age* and *-ment* selected verbal bases». *Lexique*, 26, 29–52.

Missud, A. et Villoing, F. (2021). «Investigating the distributional properties of rival *-age* suffixation and verb to noun conversion in French». *Verbum : revue de linguistique*, 43(1), 41–68.

Mitkov, R. (2022). *The Oxford Handbook of Computational Linguistics*. Oxford : Oxford University Press.

Montermini, F. (2019). Later generative grammar and beyond: lexicalism. In Audring, J. et Masini, F. (eds.), *The Oxford Handbook of Morphological Theory*, 122–142. Oxford : Oxford University Press.

Naccarato, C. (2019). «Agentive (para) synthetic compounds in Russian: a quantitative study of rival constructions». *Morphology*, 29(1), 1–30.

Nemčenko, V. (1976). *Словообразовательная структура отсубстантивных имён прилагательных в современном русском языке [Structure dérivationnelle des adjectifs dénominaux en russe contemporain]*. Thèse de doctorat, Gor'kovskij gosudarstvennyj universitet.

Noailly, M. (1999). *L'adjectif en français*. Paris : Ophrys.

Oskol'skaja, S. (2006). «Корпус письменных текстов XIX века: сферы использования и жанровое разнообразие [Corpus de textes écrits du XIXe siècle : domaines d'utilisation et diversité des genres]». *Nacional'nyj korpus russkogo jazyka*, 46–51.

Parker, J. et Sims, A. (2020). «Irregularity, paradigmatic layers, and the complexity of inflection class systems: a study of Russian nouns». In Arkadiev, P. et Gardani, F. (eds.), *The Complexities of Morphology*, 23–51. Oxford : Oxford University Press.

Partee, B. H. (2009). «Formal semantics, lexical semantics, and compositionality: the puzzle of privative adjectives». *Philologia*, 7(1), 1–24.

Paul, H. (1897). *Über die Aufgaben der Wortbildungslehre*. München : Buchdr. von F. Straub.

Pèrènlèj, U. (2006). *Языковая концептуализация социального пространства (на материале отсубстантивных префиксально-суффиксальных прилагательных русского языка) [La conceptualisation linguistique de l'espace social (sur l'exemple des adjectifs dénominaux préfixés et suffixés de la langue russe)]*. Thèse de doctorat, Irkutskij gosudarstvennyj universitet.

Plag, I. (2006). «Productivity». In Aarts, B., McMahon, A. M. et Hinrichs, L. (eds.), *Handbook of English Linguistics*, 537–556. Hoboken : Blackwell Publishing.

Plag, I. (2018). *Word-Formation in English*. Cambridge : Cambridge University Press.

Plag, I. et Baayen, R. H. (2009). «Suffix ordering and morphological processing». *Language*, 85(1), 109–152.

Plénat, M. (2011). «Enquête sur divers effets des contraintes dissimilatives en français». In Roché, M., Boyé, G., Hathout, N., Lignon, S. et Plénat, M. (eds.), *Des unités morphologiques au lexique*, 145–190. Paris : Lavoisier.

Plonsky, L., Egbert, J. et Laflair, G. T. (2015). «Bootstrapping in applied linguistics: assessing its potential using shared data». *Applied Linguistics*, 36(5), 591–610.

Plungjan, V. (2005). «Зачем нужен Национальный корпус русского языка? Неформальное введение [Pourquoi a-t-on besoin du Corpus national de la langue russe ? Une introduction informelle]». *Otečestvennye zapiski. Obščestvo v zerkale jazyka*, 23(2), 296–308.

Plungjan, V. (2008). «Корпус как инструмент и идеология: о некоторых уроках современной корпусной лингвистики [Le corpus en tant qu'outil et idéologie : à propos de certaines leçons de la linguistique de corpus contemporaine]». *Russkij jazyk v naučnom osveščenii*, 2, 7–20.

Plungjan, V., Grišina, E., K, K. et Sičinava, D. (2009). «Поэтический корпус в рамках Национального корпуса русского языка: общая структура и перспективы использования [Corpus poétique dans le cadre du Corpus national de la langue russe : structure générale et perspectives d'utilisation]». *Nacional'nyj korpus russkogo jazyka 2006-2008: novye rezul'taty i perspektivy*, 71–113.

Plungjan, V., Reznikova, T. et Sičinava, D. (2005). «Национальный корпус русского языка: общая характеристика [Corpus national de la langue russe : caractéristiques générales]». *Naučno-texničeskaja informacija. Serija 2: Informacionnye processy i sistemy*, 3, 9–13.

Poljakov, A. (2003). «Технология подготовки информации в Национальном корпусе русского языка [Technologie de préparation des informations dans le corpus national de la langue russe]». *Nacional'nyj korpus russkogo jazyka*, 2005, 175–192.

Postal, P. M. (1969). «Anaphoric islands». In Pires, A. et Taylor, H. L. (eds.), *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 5, 205–239. Chicago : University of Chicago.

Prince, A. et Smolensky, P. (2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Hoboken : John Wiley & Sons.

Pustejovsky, J. (1998). *The Generative Lexicon*. Boston : MIT press.

Rainer, F. (2012). «Morphological metaphysics: virtual, potential, and actual words». *Word Structure*, 5(2), 165–182.

Rainer, F. (2013). «Can relational adjectives really express any relation? An onomasiological perspective.». *SKASE Journal of Theoretical Linguistics*, 10(1), 12–40.

Raxilina, E., Kobricov, B., Kustova, G., Ljaševskaja, O. et Šemanaeva, O. (2006). «Многозначность как прикладная проблема: лексико-семантическая разметка в национальном корпусе русского языка [La polysémie en tant que problème appliqué : l'annotation lexicosémantique dans le Corpus national de la langue russe]». In Laufer, N., Narinjani, A. et Selegej, V. (eds.), *Komp'juternaja lingvistika i intellektual'nye texnologii. Trudy meždunarodnoj konferencii 'Dialog'*, 2006, 445–451. Moskva : Izdatel'skij centr RGGU.

Raxilina, E., Kustova, G., Ljaševskaja, O., Reznikova, T. et Šemanaeva, O. (2009). «Задачи и принципы семантической разметки лексики в НКРЯ [Objectifs et principes d'annotation sémantique du lexique dans le Corpus national de la langue russe]». *Nacional'nyj korous russkogo jazyka: 2006-2008. Novye rezul'taty i perspektivy*, 215–240.

Reformatskij, A. (1967). *Введение в языковедение [Introduction à la linguistique]*. Moskva : Prosveščenie.

Resnik, P., Elkiss, A., Lau, E. et Taylor, H. (2005). «The Web in theoretical linguistics research: two case studies using the Linguist's Search Engine». In Jaeger, T. F., Wasow, T., Cover, R. T. et Kim, Y. (eds.), *Proceedings of the 31st Meeting of the Berkeley Linguistics Society*, 265–276. Berkely : Berkeley Linguistics Society.

Roché, M. (1997). «*Briard*, *bougeoir* et *camionneur* : dérivés aberrants, dérivés possibles». In Corbin, D., Fradin, B., Habert, B., Kerleroux, F. et Plénat, M. (eds.), *Mots possibles et mots existants. 1ères rencontres du forum de morphologie*, 1, 241–250. Villeneuve d'Ascq : SILEX, Université de Lille III.

Roché, M. (2006). «Comment les adjectifs sont sémantiquement construits». *Cahiers de grammaire*, 30, 373–387.

Roché, M. (2008). «Structuration du lexique et principe d'économie : le cas des ethniques». In Durand, J., Habert, B. et Laks, B. (eds.), *Congrès mondial de linguistique française*, 1571–1585. Les Ulis : EDP Sciences.

Roché, M. (2010). «Base, thème, radical». *Recherches linguistiques de Vincennes*, 39, 95–134.

Roché, M. et Plénat, M. (2014). «Le jeu des contraintes dans la sélection du thème présuffixal». In Neveu, F., Blumenthal, P., Hriba, L., Gerstenberger, A., Meinschäfer, J. et Prévost, S. (eds.), *Congrès Mondial de Linguistique Française - CMLF'16*, 1863–1878. Paris : Institut de Linguistique Française.

Romaine, S. (1983). «On the productivity of word formation rules and limits of variability in the lexicon». *Australian Journal of Linguistics*, 3(2), 177–200.

Roon, K. (2006). «Stress in Russian compound nouns: head dominance or anti-faithfulness? In Lavine, J. E., Franks, S., Tasseva-Kurktchieva, M. et Filip, H. (eds.), *Proceedings of FASL*, 14, 319–330. Michigan : Michigan Slavic Publications.

Sahlgren, M. (2008). «The distributional hypothesis». *Italian Journal of Disability Studies*, 20, 33–53.

Säily, T. (2011). «Variation in morphological productivity in the BNC: sociolinguistic and methodological considerations». *Corpus Linguistics and Linguistic Theory*, (1), 119–141.

Sakhno, S. (2011). «Les députés du Parlement russe pensent-ils? Rapport entre la synchronie et la diachronie dans l'analyse de certains termes de langues européennes liés au concept de 'parlement'». In Briu, J.-J. (ed.), *Terminologie (I) : analyser des termes et des concepts*, 153–190. Bern : P. Lang.

Sakhno, S. (2015). «Les mots russes de la culture matérielle : les ambiguïtés du dualisme 'mot d'origine slave'/'mot occidental'». *La Revue russe*, 41, 85–99.

Sakhno, S. (2022). «Проблема семантической модели для описания дистрибуции адъективных суффиксов *-н-*, *-ов-* в русской лексике: почему *Россия снеж-н-ая*, а не *снег-ов-ая*? [Le problème du modèle sémantique pour décrire la distribution des suffixes adjectivaux *-n-*, *-ov-* dans le lexique russe : pourquoi la Russie est-elle définie *snež-n-aja*, et non *sneg-ov-aja* ?]». In Chelbaeva, T. et Lehmann-Carli, G. (eds.), *Verbunden mit den Slaven. Festschrift für Swetlana Mengel*, 45, 217–229. Berlin : Frank & Timme.

Šanskij, N. (2004). *Этимологический словарь русского языка [Dictionnaire étymologique du russe]*. Moskva : Drofa.

Šanskij, N. et Tixonov, A. (1987). *Современный русский язык [La langue russe contemporaine]*. Moskva : Prosveščenie.

Savčuk, O. (2009). «Корпус текстов первой половины XX века [Corpus de textes de la première moitié du XXe siècle]». *Nacional'nyj korpus russkogo jazyka 2006-2008: novye rezul'taty i perspektivy*, 27–45.

Savčuk, S. (2005). «Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции [Le marquage métatextuel dans le Corpus national de la langue russe : principes fondamentaux et fonctions principales]». *Nacional'nyj korpus russkogo jazyka: 2003-2005. Rezul'taty i perspectivy*, 62–88.

Savčuk, S. (2011). «Национальный корпус русского языка: перспективы использования в лингвистических исследованиях и преподавании [Le Corpus national de la langue russe : perspectives d'utilisation dans les recherches linguistiques et l'enseignement]». *Vestnik Aziatsko-Tixookeanskoj associacii prepodavatelej russkogo jazyka i literatury*, 2-3, 62–67.

Savčuk, S. et Sičinava, D. (2006). «Корпус русских текстов XVIII века в составе НКРЯ: проблемы и перспективы [Le corpus des textes russes du XVIIIe siècle dans le cadre du Corpus national de la langue russe : problèmes et perspectives]». *Nacional'nyj korpus russkogo jazyka*, 52–70.

Scalise, S. (2011). *Generative Morphology*. Berlin : De Gruyter Mouton.

Ščerba, L. (1958). *Избранные работы по языкознанию и фонетике [Œuvres choisies de linguistique et de phonétique]*. Léningrad : Izdatel'stvo Leningradskogo universiteta.

Schmid, H.-J. (2012). *English Abstract Nouns as Conceptual Shells*. Berlin : De Gruyter Mouton.

Schulte, M. (2015). «Polysemy and synonymy in derivational affixation – a case study of the English suffixes -*age* and -*ery*». *Morphology*, 25(4), 371–390.

Šemanaeva, O., Kustova, G., Ljaševskaja, O. et Raxilina, E. (2007). «Семантические фильтры для разрешения многозначности в национальном корпусе русского языка: прилагательные [Les filtres sémantiques pour la résolution de la polysémie dans le Corpus national de la langue russe : adjectifs]». In Iomdin, L., Laufer, N., Narinjani, A. et Selegej, V. (eds.), *Komp'juternaja lingvistika i intellektual'nye texnologii. Trudy meždunarodnoj konferencii 'Dialog'*, 2007, 582–587. Moskva : Izdatel'skij centr RGGU.

Semënov, A. (2003). *Этимологический словарь русского языка [Dictionnaire étymologique du russe]*. Moskva : Junves.

Sims, A. D. (2006). *Minding the Gaps: Inflectional Defectiveness in a Paradigmatic Theory*. Thèse de doctorat, State University of Ohio.

Sims, A. D. (2017). «Slavic morphology: recent approaches to classic problems, illustrated with Russian». *Journal of Slavic Linguistics*, 25(2), 489–524.

Sims, A. D. et Parker, J. (2015). «Lexical processing and affix ordering: cross-linguistic predictions». *Morphology*, 25(2), 143–182.

Šmelëva, T. (2008). «Притяжательные прилагательные: почему не сбывается виноградовский прогноз? [Les adjectifs d'appartenance : pourquoi la prédiction de Vinogradov ne se réalise-t-elle pas ?]». *Slavica Helsingiensia*, 34, 358–371.

Sokolova, S., Lyashevskaya, O. et Janda, L. A. (2012). «The locative alternation and the Russian 'empty' prefixes: a case study of the verb *gruzit'* 'load'». In Divjak, D. et Gries, S. T. (eds.), *Frequency Effects in Language Representation*, 51–85. Berlin : De Gruyter Mouton.

Sorokina, E. (1984). *Прилагательные-неологизмы современного русского языка [Les adjectifs néologiques de la langue russe contemporaine]*. Thèse de doctorat, Moskovskij gosudarstvennyj universitet.

Stemberger, J. P. et MacWhinney, B. (2004). «Are inflected forms stored in the lexicon?». *Morphology: Critical Concepts in Linguistics*, 6, 107–122.

Strnadová, J. (2014). *Les réseaux adjectivaux. Sur la grammaire des adjectifs dénominaux en français*. Thèse de doctorat, Université Paris Diderot (Paris 7) Sorbonne Paris Cité ; Univerzita Karlova.

Švedova, N. (1980). *Русская грамматика [Grammaire russe]*, 1. Moskva : Nauka.

Taft, M. (1985). «The decoding of words in lexical access: a review of the morphographic approach». *Reading Research: Advances in Theory and Practice*, 5, 83–123.

Tagliamonte, S. A. et Baayen, R. H. (2012). «Models, forests, and trees of York English: was/were variation as a case study for statistical practice». *Language Variation and Change*, 24(2), 135–178.

Therneau, T., Atkinson, B., Ripley, B. et Ripley, M. B. (2015). «Package 'rpart'», `https://cran.r-project.org/web/packages/rpart/index.html`.

Timberlake, A. (2004). *A Reference Grammar of Russian*. Cambridge : Cambridge University Press.

Toldova, S., Kustova, G. et Ljaševskaja, O. (2008). «Семантические фильтры для разрешения многозначности в национальном корпусе русского языка: глаголы [Les filtres sémantiques pour la résolution du problème de la polysémie dans le Corpus national de la langue russe : verbes]». In Kibrik, A., Belikov, V., Dobrov, B., Dobrovol'skij, D. *et al.* (eds.), *Komp'juternaja lingvistika*

i intellektual'nye texnologii. Po materialam ežegodnoj meždunarodnoj konferencii 'Dialog', 2008, 522–529. Moskva : Izdatel'skij centr RGGU.

Townsend, C. E. (1975). *Russian Word-Formation*. Bloomington : Slavica Publishers.

Tribout, D. (2010). *Les conversions de nom à verbe et de verbe à nom en français*. Thèse de doctorat, Université Paris Diderot (Paris 7).

Trubeckoj, N. S. et Jakobson, R. O. (2004). *Письма и заметки Н. С. Трубецкого [Lettres et notes de N. S. Trubeckoj]*. Moskva : Jazyki slavjanskix kultur.

Uluxanov, I. (1977). *Словообразовательная семантика в русском языке и принципы её описания [La sémantique dérivationnelle en russe et les principes de sa description]*. Moskva : Nauka.

Uluxanov, I. (2005). *Мотивация в словообразовательной системе русского языка [Motivation dans le système dérivationnel du russe]*. Moskva : Azbukovnik.

Uth, M. (2010). «The rivalry of French -*ment* and -*age* from a diachronic perspective». In Rathert, M. et Alexiadou, A. (eds.), *The Semantics of Nominalizations across Languages and Frameworks*, 215–244. Berlin/New York : De Gruyter Mouton.

Varvara, R. (2017). *Verbs as Nouns: Empirical Investigations on Event-Denoting Nominalizations*. Thèse de doctorat, University of Trento.

Varvara, R. (2019). «Misurare la produttività morfologica: i nomi d'azione nell'italiano del ventunesimo secolo». In Berruto, G., Bernini, G., Bianconi, S., D'Achille, P., Ferrari, A., La Fauci, N., Loporcaro, M., Moretti, B., Natale, S., Pandolfi, E. M. *et al.* (eds.), *Le tendenze dell'italiano contemporaneo rivisitate. Atti del LII Congresso Internazionale di Studi della Società di Linguistica Italiana*, 187–201. Milan : Officinaventuno.

Varvara, R. (2020). «Constraints on nominalizations: investigating the productivity domain of Italian -*mento* and -*zione*». *Zeitschrift für Wortbildung/Journal of Word Formation*, 4(2), 78–99.

Varvara, R., Lapesa, G. et Padó, S. (2021). «Grounding semantic transparency in context». *Morphology*, 311, 213–234.

Vinogradov, V. (1952). *Русский язык [La langue russe]*. Moskva : Izdatel'stvo Moskovskogo universiteta.

Vinogradov, V. (1977). *Избранные труды. Лексикология и лексикография [Œuvres choisies. Lexicologie et lexicographie]*. Moskva : Nauka.

Vinogradov, V. et Švedova, N. (1964). *Очерки по исторической грамматике русского литературного языка XIX века [Essais sur la grammaire historique de la langue russe littéraire du XIXe siècle]*. Moskva : Nauka.

Vinokur, G. (1959). «Заметки по русскому словообразованию [Notes sur la formation des mots en russe]». In Barxudarov, S. et Vinokur, G. (eds.), *Izbrannye raboty po russkomu jazyku*, 419–442. Moskva : AN SSSR.

Virtanen, P., Gommers, R., Oliphant, T. E. *et al.* (2020). «SciPy 1.0: fundamental algorithms for scientific computing in python». *Nature Methods*, 17(3), 261–272.

Volterra, V. (1926). «Fluctuations in the abundance of a species considered mathematically». *Nature*, 118, 558–560.

Wauquier, M. (2020). *Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels*. Thèse de doctorat, Université Toulouse II Jean Jaurès.

Wauquier, M., Hathout, N. et Fabre, C. (2020). «Semantic discrimination of technicality in French nominalizations». *Zeitschrift für Wortbildung/Journal of Word Formation*, 4(2), 100–119.

Zaliznjak, A. (2002). *'Русское именное словоизменение' с приложением избранных работ по современному русскому языку и общему языкознанию ['La déclinaison nominale en russe' avec des œuvres choisies sur la langue russe moderne et la linguistique générale]*. Moskva : Jazyki slavjanskoj kul'tury.

Zaliznjak, A. (2003). *Грамматический словарь русского языка [Dictionnaire grammatical de la langue russe]*. Moskva : Russkie slovari.

Zemskaja, E. (1965). «О некоторых факторах развития словообразовательной системы современного русского языка [Sur quelques facteurs du développement du système dérivationnel du russe contemporain]». In Xrapčenko, M. et Vinogradov, V. (eds.), *Problemy sovremennoj filologii. Sbornik statej k 70-letiju akademika V. V. Vinogradova.*, 142–148. Moskva : Nauka.

Zemskaja, E. (1991). «Относительные прилагательные как конструктивный элемент номинативной системы современного языка. [Les adjectifs relationnels en tant qu'élément constructif du système nominal de la langue contemporaine]». In Il'ina, N. et Voroncova, V. (eds.), *Grammatičeskie issledovanija: funkcional'no stilističeskij aspekt. Morfologija. Slovoobrazovanie. Sintaksis*, 132–165. Moskva : Nauka.

Zemskaja, E. (2000). *Русский язык конца XX столетия (1985-1995) [La langue russe à la fin du XXe siècle (1985-1995)]*. Moskva : Jazyki russkoj kul'tury.

Zemskaja, E. A. (2011). *Современный русский язык. Словообразование. [La langue russe contemporaine. Formation des mots]*. Moskva : Flinta.

Zemskaja, E. A. (2015). *Язык как деятельность. Морфема, слово, речь. [La langue en tant qu'activité. Morphème, mot, parole.]*. Moskva : Flinta.

Zhang, Z., Mayer, G., Dauvilliers, Y., Plazzi, G., Pizza, F. *et al.* (2018). «Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from european narcolepsy network database with machine learning». *Scientific Reports*, 8(1), 1–11.

Zipf, G. K. (1935). *The Psycho-Biology of Language: an Introduction to Dynamic Philology.* Boston : Houghton Mifflin.

Zolotarëva, T. (2003). *Семантические особенности английских абстрактных существительных, влияющие на использование артикля. [Les particularités sémantiques des noms abstraits anglais qui influencent l'utilisation de l'article].* Thèse de doctorat, Moskovskij pedagogičeskij gosudarstvennyj universitet.

Zverkovskaja, N. (1986). *Суффиксальное словообразование русских прилагательных, XI-XVII вв. [La dérivation suffixale des adjectifs russes aux XIe-XVIIe siècles.].* Moskva : Nauka.

Table des matières

Résumé

La présente étude se focalise sur la concurrence entre trois suffixes, -n-, -sk- et -ov-, qui servent à construire des adjectifs dénominaux en russe. L'objectif consiste à analyser cette concurrence dans une perspective quantitative. Nous commençons par la présentation de la base de données RuDénom, que nous avons constituée à partir du Corpus National de la Langue Russe et qui se compose exclusivement d'adjectifs dérivés d'une base nominale. Ce corpus répertorie les adjectifs, les noms de base et leurs fréquences respectives. Par la suite, nous exposons les annotations des propriétés des noms susceptibles d'influencer le choix des trois suffixes étudiés, à savoir les propriétés phonologiques, morphologiques, sémantiques et étymologiques. Afin de réduire les biais d'encodage manuel de la sémantique et de l'étymologie, nous proposons deux méthodes quantitatives pour ces deux propriétés. L'examen de la concurrence est envisagé sous trois aspects. En premier lieu, nous analysons les propriétés des noms de base et proposons des modèles numériques décrivant l'interaction de ces propriétés dans les données de haute fréquence et dans les hapax. Cette analyse nous permet d'identifier les propriétés des noms ayant le plus d'impact sur le choix des suffixes dans les deux sous-corpus, ainsi que la comparaison des tendances entre le lexique courant et les créations lexicales. Dans un deuxième temps, nous étudions la productivité des suffixes, en nous appuyant principalement sur les mesures de fréquences des adjectifs et des noms. Bien que -n-, -sk- et -ov- soient très productifs en russe contemporain, nous établissons un classement de productivité pour ces trois suffixes. Enfin, notre recherche se porte sur les adjectifs doublets. Nous examinons les tendances dans leurs fréquences et les propriétés des noms de base, puis nous complétons cette étude par un analyse combinatoire des adjectifs doublets et de leurs noms recteurs. Cette analyse nous permet d'estimer la similarité fonctionnelle et sémantique des adjectifs doublets.

Abstract

The present study focuses on the competition between three suffixes, -n-, -sk-, and -ov-, used to form denominal adjectives in Russian. The objective is to analyze this competition using a quantitative approach. We begin with an introduction to the RuDénom database, which we compiled from the Russian National Corpus and which is composed exclusively of adjectives derived from a nominal base. This corpus contains the adjectives, their base nouns, and their respective frequencies. Subsequently, we detail the annotations of the properties of base nouns that may potentially influence the choice of the three suffixes, namely phonological, morphological, semantic, and etymological. To minimize manual encoding bias in semantics and etymology, we propose two quantitative methods for these properties. The investigation of competition is considered from three perspectives. First, we analyze the properties of base nouns and propose numerical models describing the interaction of these properties in high-frequency data and hapaxes. This analysis enables us to identify the noun properties that most influence the choice of suffixes in both sub-corpora, as well as to compare trends between common lexicon and lexical creations. Next, we study the productivity of the suffixes, primarily relying on frequency measures of adjectives and nouns. Even though -n-, -sk-, and -ov- are highly productive in contemporary Russian, we establish a productivity ranking for these suffixes. Finally, our research extends to doublets. We study trends in their frequencies and the properties of base nouns, supplementing this analysis with a combinatorial examination of the head nouns of the doublet adjectives. This enables us to estimate the functional and semantic similarity of doublets.