La phonotaxe du russe dans la typologie des langues : focus sur la palatalisation

Ekaterina Biteeva Lecocq, Nathalie Vallée, Denis Faure-Vincent

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

La phonotaxe du russe dans la typologie des langues : focus sur la palatalisation

Ekaterina Biteeva Lecocq Nathalie Vallée Denis Faure-Vincent
Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
*Institute of Engineering Univ. Grenoble Alpes
ekaterina.lecocq@gipsa-lab.fr, nathalie.vallee@gipsa-lab.fr,
denis.faure-vincent@gipsa-lab.fr

RÉSUMÉ

Cet article présente un travail de description phonotactique du russe basé sur une analyse de 15 000 lemmes transcrits phonologiquement et syllabés. Un ensemble de données quantitatives relatives aux structures syllabiques a été examiné dans une perspective typologique. À partir d'une analyse distributionnelle des segments consonantiques ±PAL, des probabilités phonotactiques ont été estimées. Les résultats montrent que le russe suit globalement les tendances générales observées dans les langues de la base de données G-ULSID (Vallée, Rousset & Rossato, 2009) et mettent en évidence des asymétries de distribution des consonnes ±PAL à l'intérieur de la syllabe. Le fait que le système consonantique du russe présente une distinctivité ±PAL étendue à tous les lieux d'articulation, semble contraindre les coccurrences entre consonne et voyelle d'une même syllabe prédites par la théorie Frame/Content (MacNeilage, 1998) et trouvées dans de nombreuses langues.

ABSTRACT

This paper presents a phonotactic description of Russian based on an analysis of 15,000 phonologically transcribed and syllabified lemmas. A set of quantitative data relating to the syllabic structures of Russian has been examined in a typological perspective. From a distributional analysis of ±PAL consonant segments, phonotactic probabilities were estimated. Our results show that Russian broadly follows the general trends observed in the languages of the G-ULSID database (Vallée, Rousset & Rossato, 2009) and highlight asymmetries in the distribution of ±PAL consonants within-syllable units. The fact that Russian presents ±PAL distinctiveness extended to all its consonant places of articulation seems to constrain tautosyllabic consonant/vowel cooccurrences predicted by the Frame/Content Theory (MacNeilage, 1998) and overrepresented in lot of languages.

MOTS-CLÉS : russe, phonotaxe, syllabe, palatalisation, tendances universelles, G-ULSID
KEYWORDS: Russian phonotactics, syllable, palatalization, universal trends, G-ULSID

1 Introduction

La prise en compte des données phonotactiques dans la caractérisation des langues n'est plus à démontrer. Au niveau segmental, les langues ne se distinguent pas seulement dans l'inventaire de leurs systèmes phonologiques, les régularités distributionnelles des segments à l'intérieur des mots et des syllabes participent elles aussi à définir le processus structurel de formation des séquences phonologiques. Ainsi les langues possèdent des propriétés phonotactiques qui règlent les combinaisons de segments dans les syllabes et dans les enchainements de syllabes qui forment les

Volume 1 : Journées d'Études sur la Parole, pages 36–44.
HAL : hal-02798512.

séquences sonores. Les contraintes de positions syllabiques et celles sur la formation des mots, ainsi que les probabilités de cooccurrences entre segments déterminent la phonotaxe des langues (Jusczyk et al. 1994). Nombreuses sont les études qui ont montré un effet de la composante phonotactique de la langue sur le traitement du langage : la phonotaxe, avec ses patrons de régularité, conditionne la perception de la parole (Segui et al., 2002 ; Dehaene-Lambertz et al., 2000), fournit des indices de segmentation du flux de parole (Mattys & Jusczyk, 2001), agit sur les performances de rappel mnésique (Gathercole et al., 1999) et influence la production de la parole (Goldrick & Larson, 2008). La sensibilité aux contraintes phonotactiques de la langue maternelle se mettrait en place aux alentours de 9 mois (Jusczyk & Luce, 1994) et l'effet de probabilité sur le traitement lexical entre 7 et 10 ans (Storkel & Rogers, 2000). Les contraintes de la phonotaxe influencent aussi la phonologisation des emprunts (Kenstowicz, 2010) et on les retrouve également dans les patrons structurels des erreurs de production (Warker & Dell, 2006). Ces études montrent la nécessité de disposer de données de phonotaxe. D'autres travaux montrent également l'intérêt d'utiliser les indices phonotactiques dans le domaine de l'identification (ex. Najafian et al. 2016 ; Srivastava et al. 2017) ou du traitement automatique des langues (ex. Zhu & Adda-Decker, 2006) ou encore de l'apprentissage automatique (Eychenne, 2015). Ce dernier propose d'intégrer au cadre théorique formel Maximum Entropy de Hayes et Wilson (2008) une modélisation des phénomènes phonotactiques basée sur des contraintes pondérées de bonne formation afin de modéliser la grammaire sous forme probabiliste. On comprend alors pourquoi la recherche des régularités phonotactiques s'impose dans la description des langues.

Nous proposons ici une étude de la phonotaxe du russe basée sur une analyse de 15 000 lemmes phonologisés et syllabés de première main. Elle nous permet de présenter un ensemble de généralisations pour le russe que nous abordons dans une perspective typologique en adressant aux contraintes phonotactiques la question de la marque. Notre étude propose également un focus sur les consonnes palatalisées pour lesquelles des désaccords persistent entre les deux écoles phonologiques de Saint-Pétersbourg (ex. Zinder, 1979) et de Moscou (ex. Avanesov, 1956). En russe, contrairement à d'autres langues slaves (Iskarous & Kavitskaya, 2018), la palatalisation connaît un fort rendement phonologique puisque la plupart des consonnes possèdent leur équivalent palatalisé. La palatalisation peut être également l'output d'un processus d'assimilation régressive lorsque $C_{[-PAL]} \rightarrow C_{[+PAL]} / \_ C_{[+PAL]}$. La phonologisation de la palatalisation devant les voyelles coexiste ainsi avec la palatalisation comme résultat d'une assimilation régressive lorsque les consonnes acquièrent une articulation palatale secondaire au contact de $C_{[+PAL]}$. Ce processus allophonique ainsi que les six phonèmes palatals /ʧ ʃ ʃʲ: ʒ (ʒʲ:) j/ rendent le système phonologique du russe très intéressant pour observer, grâce à l'approche typologique, des régularités distributionnelles des consonnes palatalisées considérées comme articulatoirement complexes. À notre connaissance, les travaux antérieurs de description phonotactique du russe ont été établis à partir de deux études (Peshkovskii, 1925 ; Pirogova, 2018) basées sur des types de textes de divers genre et longueur. Au-delà de la recherche d'indices phonotactiques, la comparaison de nos résultats avec les données de ces études nous permet de discuter la corrélation entre les régularités observées dans les entrées lemmatiques d'un lexique référant à une prononciation standardisée et la fréquence des mots en corpus.

2 Matériel et méthode

Le lexique du russe sur lequel s'appuie notre étude contient 15 000 lemmes extraits d'un dictionnaire d'environ 35 000 mots. Ce dictionnaire de fréquences du russe était un des projets de l'Institut russe de recherche en intelligence artificielle conduit d'abord par Sharov (http://www.artint.ru/projects/frqlist.php) puis poursuivi par Lyashevskaya & Sharov (2009). Il a été élaboré sur une collection de textes du corpus national de la langue russe

(http://www.ruscorpora.ru/new/), représentant la langue de la période 1950-2007. Une version électronique du dictionnaire est publiée sur le site de l'Institut de la Langue Russe de V.V. Vinogradov de l'Académie des Sciences (http://dict.ruslang.ru). La liste des items lexicaux est représentative du russe moderne. Il comprend une sélection de prose, de mémoires politiques, de journaux et de littérature scientifique populaire (environ 40 millions de mots). Tous les textes du corpus ont été écrits en russe entre 1970 et 2002.

Pour notre étude, les traitements effectués sur le lexique relèvent du protocole mis en place pour le projet G-ULSID (*Grenoble & UCLA – Lexical and Syllabic Inventory Database*). Ce projet vise à constituer une base de données relationnelles (MYSQL) de lexiques phonologisés et syllabés pour la recherche de régularités dans la phonotaxe des langues du monde qui prennent en compte la structure de la syllabe et les niveaux infra-et supra-syllabiques (Vallée et al., 2009). Chaque entrée lexicale est phonologiquement transcrite et découpée en syllabe(s), et chaque syllabe est décomposée en sous constituants (attaque et rime décomposée en noyau et coda). Dans la continuité de Maddieson & Precoda (1992), seuls les lemmes sont pris en compte et les emprunts récents sont écartés. Par conséquent, les prénoms et patronymes, noms propres, abréviations, interjections et emprunts récents n'ont pas été pris en compte pour le russe. La transcription ne tient pas compte des allophonies et n'ont donc pas été retenus : (1) les cas de palatalisation par anticipation de la première consonne dans les clusters si suivie d'une consonne palatalisée ; (2) le dévoisement final ; (3) les processus de réduction vocalique ; (4) la simplification des groupes de consonnes systématiquement observée à l'oral dans les combinaisons -stn, -zdn, -stl, -ntsk, -stsk, -vstv, -lnts.

La base de donnée est consultable depuis des pages web (via le serveur APACHE et les langages de programmation PHP, HTML et Javascript). Le lexique du russe a été intégré d'abord par une conversion des graphèmes cyrilliques vers les symboles de l'API, traitement effectué par un programme PHP et des requêtes SQL. Une autre application web a permis à un locuteur natif de contrôler le lexique et d'effectuer la syllabation, vérifiée ensuite par deux autres locuteurs natifs. Enfin, la plupart des statistiques et des graphiques de cet article sont issus de l'application web qui analyse les données du russe provenant de la base de données MYSQL.

3 Résultats

Dans une première partie est présenté un ensemble de données quantitatives relatives aux structures syllabiques et lexicales du russe, et replacé dans le contexte de l'étude typologique de Vallée & Rousset (2004) réalisée à partir d'un corpus constitué d'une quinzaine de langues de G-ULSID. Une seconde partie est consacrée à une analyse distributionnelle des segments consonantiques palatalisés (C^j). Elle contient une analyse des occurrences et cooccurrences des consonnes et des voyelles dans les syllabes de types CV, CVC, C^jV, C^jVC, VC, CVC, VC^j et CVC^j à partir desquelles sont estimées des probabilités phonotactiques.

3.1 Données typologiques

Les lemmes de 3 syllabes sont trouvés majoritaires en russe (près de 37 % des entrées), les unités dissyllabiques arrivant au deuxième rang avec 27.36 %. Avec une moyenne de 3.06 syllabes par lemme, le russe se place dans le Type 4 de la typologie des langues de G-ULSID proposée par Vallée & Rousset (2004), à savoir, parmi les langues présentant dans le même temps peu de lemmes monosyllabiques et de plus de 3 syllabes (Figure 1). À noter qu'aucune langue de la base de données ne comporte un mode statistique supérieur à 3.

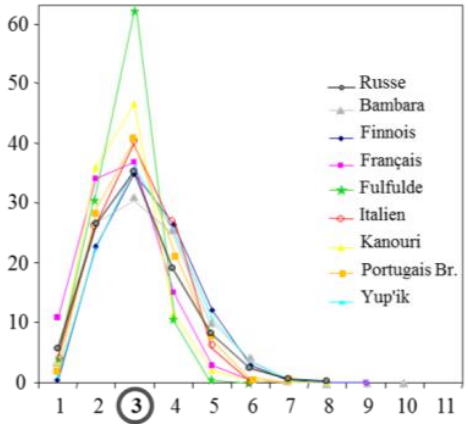

FIGURE 1: Répartition des unités lexicales en fonction du nombre de syllabes qu'elles contiennent. Le russe est ici représenté avec les autres langues de g-ulsid qui présentent un mode de distribution à 3 syllabes. Il s'agit du Type 4 regroupant les langues ayant une majorité d'unités lexicales trisyllabiques (adapté à partir de Vallée, 2017).

La syllabe en russe est majoritairement constituée de 2 segments comme c'est le cas pour la plupart des langues de la base de données, suivie de près par les syllabes à 3 segments (47 % et 42 % respectivement). Les syllabes de plus de 3 phonèmes sont sous-représentées (moins de 8 % du nombre total) ce qui correspond aux tendances générales remarquées pour l'ensemble des langues. Quant au nombre moyen de phonèmes par lemme, celui-ci s'élève à 7.82 (min=5.09, max=7.91), valeurs cohérentes avec celles présentées par les langues des types 3 et 4 de G-ULSID (le type 3 regroupe les langues dont le mode statistique de la distribution des unités lexicales par nombre de syllabes est 2). La loi de Menzerath (1954) décrit le lien qui existe entre le nombre de phonèmes par syllabe et le nombre de syllabes par unité lexicale : le premier a tendance à diminuer lorsque le deuxième augmente. Cette tendance ne se vérifie pas en russe où le nombre de phonèmes par syllabe reste stable et oscille autour de 2,5 quelle que soit la longueur de l'unité lexicale. Le rendement syllabique calculé pour un lexique (ratio entre le nombre total de syllabes et le nombre de syllabes différentes) permet d'évaluer l'efficacité d'une langue dans la combinaison de ses syllabes. Pour le russe, le ratio calculé est égal à 12.81. Cette valeur est cohérente avec celles des langues du type 3 qui ont un rendement entre 5 et 14 (Vallée & Rousset, 2004) soutenant l'idée que plus la proportion d'unités lexicales longues est importante, plus les langues réutilisent des syllabes limitant ainsi la taille de leur inventaire syllabique. Les gabarits lexicaux les plus fréquents en russe sont CV.CVC et CV.CV.CVC avec 828 et 814 entrées. En revanche, le gabarit dissyllabique le plus fréquent CV.CV qui arrive au 1er rang pour les langues du type 3 ou 4 ne représente en russe que 1.45 % des lemmes (sur 15 000).

Les structures syllabiques CV et CVC totalisent respectivement 44 % et 36 % des syllabes et sont de loin les plus fréquentes en russe. Ces deux structures sont également dominantes dans la plupart des langues de G-ULSID. Le russe comporte plus de syllabes à attaque pleine et satisfait ainsi au principe du MOP *Maximal Onset Principal*. En revanche, les structures à attaque et/ou coda branchante(s) sont très peu fréquentes (6 % pour CCV et < 4 % pour les autres). Cela est cohérent avec la relation inverse universelle entre la fréquence d'une structure syllabique donnée et son degré de complexité.

3.2 Consonnes palatalisées et phonotaxe

Dans le lexique analysé, le nombre d'occurrences des consonnes palatalisées (+PAL) est inférieur à celui des non palatalisées (-PAL), modes et lieux d'articulation confondus. Les consonnes +PAL totalisent 31 % (18 271) d'occurrences sur l'ensemble des consonnes -PAL et +PAL et sont moitié moins représentées que les -PAL avec 69 % (41 395) du nombre global d'occurrences (59 666). Plus précisément, l'analyse de leur distribution dans la syllabe indique que les -PAL apparaissent plus souvent dans des attaques et codas simples (respectivement AS et CS) ou complexes (respectivement AC et CC) que leurs contreparties +PAL (Table 2). On note cependant que dans les structures syllabiques avec attaque et/ou coda complexe(s) la proportion de consonnes -PAL et +PAL diminue quels que soient les lieu et mode d'articulation : ex. la différence est de 12 511 occurrences entre AS et AC et de 7 330 entre CS et CC pour les -PAL. Pour les +PAL, la différence est égale à 9 321 entre AS et AC et à 3 818 entre CS et CC.

	AS	AC	CS	CC
-PAL	64 % (20 499)	79 % (7 988)	70 % (10 119)	86 % (2 789)
+PAL	36 % (11 441)	21 % (2 120)	30 % (4 264)	14 % (446)
Total	31 940	10 108	14 383	3 235

TABLE 2 : Distribution des C ±PAL en fonction de la position dans la syllabe

La distribution des consonnes -PAL *vs* +PAL varie également en fonction du lieu articulatoire et une interaction lieu*position à l'intérieur de la syllabe est relevée. Premièrement, les +PAL les plus favorisées sont labiales ou coronales : le ratio -PAL / +PAL est de 3 pour les bilabiales *Bi* et les labiodentales *LDe*, 2 pour les coronales *Co* et 6 pour les vélaires *Ve*. Parmi les +PAL la coronale /tʲ/ est surreprésentée et totalise presque 28 % (5 152) d'occurrences. Au 2ème rang on trouve /lʲ/ suivie des deux autres sonantes coronales /nʲ rʲ/ avec respectivement 14 % (2 544), 12 % (2 238) et 12 % (2 227) du nombre global de +PAL (18 271). Les pourcentages suivants sont basés sur le nombre total d'occurrences pour chacune des catégories : 8 448 occurrences pour *Bi*, 5 254 pour *LDe*, 39 240 pour *Co* et enfin, 6 724 occurrences relevées pour *Ve*.

Bi	*Bi Pal*	*LDe*	*LDe Pal*	*Co*	*Co Pal*	*Ve*	*Ve Pal*
75,44 %	*24,55 %*	*74,34 %*	*25,65 %*	*64,35 %*	*35,64 %*	*87,2 %*	*12,79 %*

Deuxièmement, les labiales *La* (*Bi*+*LDe*) et les *Co* présentent un certain nombre d'asymétries distributionnelles. La Table 3 présente la répartition des consonnes ±PAL en russe en fonction du lieu d'articulation et de la position dans la syllabe.

		AS (%)	AC (%)	CS (%)	CC (%)	Total
Bi	-PAL	41,35	20,71	12,11	1,28	8 448
	+PAL	22,49	1,69	0,36	0,01	
LDe	-PAL	50,00	12,30	10,93	1,12	5 254
	+PAL	21,53	3,86	0,23	0,04	
Co	-PAL	28,52	11,63	18,02	6,18	39 240
	+PAL	19,64	4,12	10,76	1,13	
Ve	-PAL	47,41	15,27	21,58	2,94	6 724
	+PAL	10,47	2,32	0	0	

TABLE 3 : Répartition des C ±PAL selon leur lieu d'articulation et leur position dans la syllabe

Les *La* +PAL occupent assez rarement les positions de CS ou CC contrairement à leurs équivalents -PAL : on retrouve seulement une occurrence de /bʲ/ en CC et une trentaine de *Bi* +PAL en CS sur 8 448 occurrences de bilabiales relevées dans le lexique. Les labiales rencontrées le plus souvent en attaque, simple ou complexe, sont –PAL : 3 493 *vs* 1 900 pour +PAL en AS ; cet écart est plus important encore en AC où les +PAL totalisent seulement 143 occurrences contre 1 750 pour les -PAL. Les *LDe* suivent exactement les mêmes tendances. On note une très faible présence de /f/ et /fʲ/ dans le lexique avec un rendement nul pour /fʲ/ en CS et CC. Les *Co* +PAL sont généralement moins représentées dans le lexique que leurs contreparties -PAL mais on relève cependant quelques exceptions : /tʲ/ est surreprésentée en CS par rapport à /t/: 3 242 contre 941 ; le nombre d'occurrences de /lʲ/ est supérieur à /l/ sauf en AC. Les *Ve* +PAL sont souvent limitées à la position initiale de syllabe devant voyelle et sont plus rares en AC : 153 occurrences sur 602 pour /kʲ/ et seulement 2 (sur 190) et 1 (sur 68) respectivement pour /gʲ/ et /xʲ/. Enfin, nous n'avons pas relevé de *Ve* +PAL en coda, simple ou complexe, dans le lexique. Nous avons calculé les matrices de cooccurrences entre consonnes et voyelles tautosyllabiques pour les structures CV, CVC et VC afin d'estimer la probabilité des segments en séquence en fonction de la présence ou non d'une consonne palatalisée. Ainsi les matrices ont été calculées pour les 7 types de syllabes suivants : CV, CʲV, CVC, CʲVC, CVCʲ, VC et VC ʲ. Des études antérieures ayant montré l'absence de tendance au niveau des modes articulatoires dans

ces types de syllabe (Rousset, 2004 ; Vallée et al., 2009), seuls les lieux d'articulation ont retenu notre intérêt. Les préférences phonotactiques ont donc été observées en regroupant les C ±PAL par lieu d'articulation (La /p b m f v/, Co /t d s z l n r/ ou Ve /k g x/). Ce regroupement a permis de comparer les résultats obtenus pour le russe, à ceux trouvés pour d'autres langues dans des études antérieures. Les résultats présentés Table 4 regroupent les syllabes CVC avec CV, et C^{j}VC avec C^{j}V, car aucun effet de la coda n'a été relevé sur les préférences entre consonne en attaque et noyau vocalique.

	I	A	U	E	O
La	0.98	0.66	0.92	0	1.29
Co	1.29	1.07	0.99	0	0.84
Ve	0	1.35	1.17	0	1.07

	I	A	U	E	O
La	0.71	0.93	0.27	1.35	0.46
Co	1.01	1.15	1.41	0.94	1.34
Ve	1.91	0	0.11	0.27	0

TABLE 4 : Ratio R entre syllabes attestées et attendues : cvc avec cv (à gauche) et c^{j}vc avec c^{j}v (à droite). Les attaques sont en colonnes. R < 1 signifie que les combinaisons ne sont pas favorisées ; R = 1 signifie qu'il n'est pas possible de faire une prédiction et R > 1 indique que la combinaison est favorisée. Les cellules en grisé correspondent aux cooccurrences favorisées en russe.

Un nombre élevé d'occurrences de /s^{j}/ en position d'attaque simple et de syllabe /s^{j}a/ a été relevé dans le lexique. Ce phénomène est lié à la présence des verbes réfléchis formés en ajoutant le postfixe *-s^{j}a* à la base verbale. Le rendement important de cette syllabe (1 241 occurrences) renvoyant à une fonction grammaticale est susceptible d'apporter un biais dans les préférences phonotactiques. Nous avons choisi de l'extraire des calculs sans l'évacuer de notre discussion. Les résultats présentés tiennent compte de cette soustraction. Les résultats sur les préférences entre voyelle et consonne en attaque dans les syllabes de structure CV et CVC montrent que la voyelle centrale est moins favorisée après les attaques labiales qui privilégient d'une part un noyau postérieur et d'autre part /o/ à /u/. Les consonnes *Ve* sont plutôt suivies des voyelles postérieures dans CVC et des postérieures et centrales dans CV. Les combinaisons entre *Co* et voyelles centrales sont privilégiées. Cette tendance est à relier à la centralisation de /i/ réalisé [ɨ] en russe après les consonnes -PAL. En revanche, la tendance coronale-antérieure est portée essentiellement par les syllabes de type C^{j}V et C^{j}VC suggérant que l'assimilation palatale dans un système avec une distinctivité ±PAL étendue à la plupart des unités consonantiques qui le constituent, efface la tendance observée dans les langues du monde. En effet, la matrice montre que lorsque l'attaque est une consonne +PAL, le noyau antérieur est retrouvé dans les combinaisons quel que soit le lieu d'articulation consonantique. L'examen des cooccurrences entre constituants de la rime dans VC, CVC, VC^{j} et CVC^{j}, montre que le nombre de combinaisons possibles entre voyelle et consonne est plus restreint dans une structure VC^{j} que dans CVC^{j}. Cependant, les combinaisons préférées pour les deux structures sont /a/+/l^{j}/ et /i/+/t^{j}/ alors que CVC^{j} favorise /a/+/t^{j}/ et /u/+/t^{j}/. Les rimes avec /a/, /i/ ou /u/ suivi de la coda *Co* +PAL /t^{j}/ relèvent pour l'essentiel des formes verbales infinitives. La structure VC présente quatre séquences saillantes /i/+/s/, /i/+/z/, /o/+/b/, /o/+/t/ attribuables aux suffixes perfectifs.

4 Discussion

Cet article présente des données sur les tendances dans les structures syllabiques et lexicales du russe analysées dans une perspective typologique. Les résultats montrent que le russe suit globalement les tendances générales observées dans les langues de la base de données G-ULSID par Vallée & Rousset (2004), Rousset (2004). Parmi celles-ci citons les syllabes constituées de 2 segments, les unités lexicales formées de 3 syllabes, cependant sans suivre la relation tendancielle entre le nombre de syllabes dans un mot et le nombre moyen de segments qui la compose. Rousset

(2004) confronte la typologie basée sur la répartition du nombre de syllabes par unité lexicale au nombre de phonèmes par syllabe et obtient une deuxième typologie présentant trois patrons différents d'organisation lexicale A, B et C. D'après cette typologie, les langues semblent favoriser les syllabes constituées de 2 segments (type B, le plus représenté) de la même manière qu'elles favorisent les unités lexicales de 2 syllabes. Avec deux langues dissyllabiques, le sora et le suédois, et une langue trissyllabique, le yup'ik, le russe appartient au type C qui regroupe les langues dont le mode statistique de la distribution des syllabes par nombre de phonème est 2, avec une proportion très proche des syllabes de 3 segments, les deux totalisant près de 90 % des syllabes relevées sur l'ensemble du lexique. Le russe est une langue à structure syllabique majoritairement CV (structure canonique universelle), et CVC au 2e rang des fréquences. Les structures se raréfient dans le lexique en fonction du degré de complexité des attaques et/ou des codas (complexité estimée par le nombre de segments qui les composent). Le cadre ou gabarit lexical le plus fréquent en russe, bien que dissyllabique, diffère du gabarit CV.CV le plus fréquemment trouvé dans les langues de G-ULSID par le fait que la deuxième syllabe est composée avec une coda simple. Les résultats obtenus mettent en évidence des asymétries de distribution des consonnes ±PAL à l'intérieur de la syllabe. Ils montrent que quels que soient le lieu d'articulation et la position occupée par ces consonnes dans une structure syllabique donnée, la proportion des +PAL est toujours inférieure à celle des -PAL avec quelques exceptions au niveau des coronales /tʲ lʲ nʲ rʲ/, et des occurrences sous-représentées voire nulles dans le cas des vélaires et des labiales. Le classement de ces dernières par ordre de fréquence décroissant dans le lexique est /k/>/g/>/x/ et /v/>/p/>/m/>/b/> /f/ pour les ±PAL. Il est identique à celui proposé par Pirogova (2018) et à celui de Peshkovskij (1925). Notons un faible nombre d'occurrences voire une absence de /gʲ/ et /xʲ/ et de /fʲ/ en attaque complexe et en coda. La sous-représentation de /f/ et /fʲ/ dans le lexique a une raison historique. D'abord présentes dans les emprunts, elles apparaissent en vieux slave au XIe siècle lorsque /v/, après la chute des deux voyelles faibles antérieure et postérieure, se retrouve en fin de mot ou devant une consonne sourde. Dans les langues slaves et dans certains dialectes du russe, il existe toujours une tendance à éviter [f] en le remplaçant par [p] ou [xv] (Remneva, 2012). La distribution des vélaires trouvée dans notre lexique est aussi un héritage du proto-slave qui n'autorisait pas aux consonnes k, g et x de se trouver devant les voyelles antérieures. Cela reste le cas en russe qui présente pour CV les combinaisons favorisées $Ve_{[+PAL]}$-An et $Ve_{[-PAL]}$-Po. Selon l'école phonologique de Moscou (Avanesov, 1956), [kʲ], [gʲ] et [xʲ] n'ont pas de statut phonémique car elles ne se trouvent jamais en position finale de mot alors que [k], [g] et [x] y sont trouvées. Les oppositions /k/~/kʲ/, /g/ ~/gʲ/, /x/~/xʲ/ sont également marginales devant voyelle (Remneva, 2012). Elles correspondent effectivement dans notre étude aux occurrences les plus faibles en position d'attaque syllabique. Nos résultats confirment ceux de Kochetov (2002) selon lesquels les palatalisations labiales et coronales sont plus fréquentes dans les langues slaves que les palatalisations vélaires. Quant aux coronales, l'ordre de fréquences des segments ±PAL relevé par Pirogova et Peshkovskij diffère de celui trouvé dans notre lexique. Pirogova propose /n/>/t/>/s/>/l/>/r/>/d/>/z/ et idem pour les +PAL. Nous avons obtenu pour les -PAL /n/>/s/>/r/>/t/>/l/>/d/>/z/ et pour les +PAL /tʲ/>/lʲ/>/nʲ/>/rʲ/>/sʲ/>/dʲ/>/zʲ/. La surreprésentation de /tʲ/ dans nos données s'explique par la présence des verbes à l'infinitif en /tʲ/ car notre lexique est composé de lemmes alors que Pirogova et Peshkovskij ont travaillé à partir de textes et donc de formes verbales fléchies. La forte présence de /lʲ/ a également une origine morphologique puisque cette consonne est un élément des suffixes *-/lʲiv/* et *-/alʲ/* qui servent à former des adjectifs et des adverbes. L'explication d'un fort rendement des coronales +PAL se trouve peut-être dans l'évolution du proto-slave. Selon Shevelov (1965), la palatalisation des consonnes peut être reconstruite vers le proto-slave du Ve au VIIIe siècle après JC. À cette époque s'est produite la coalescence des consonnes avec -j. Les combinaisons La+[j] sont devenues La+[lʲ] par ajout d'une liquide épenthétique : pj>plʲ, vj>vlʲ etc. Les autres séquences C+[j] ont fusionné en consonnes palatalisées Cj>Cʲ. Ainsi, les contrastes /l/~/lʲ/, /n/~/nʲ/, /r/~/rʲ/, /t/~/tʲ/ étaient déjà présents dans tous les dialectes du proto-slave. D'autre part, /t/ et /tʲ/ en finale de mot jouent le rôle d'un indice morphologique, en

marquant soit les formes verbales du participe passé passif ou de l'infinitif ; /tʲ/ est aussi un des marqueurs des noms féminins de la 3e déclinaison. Les matrices de cooccurrences obtenues révèlent que les *La* et les *Co* +PAL présentent un certain nombre d'asymétries distributionnelles, tandis que les *Ve* +PAL semblent assez limitées à la position prévocalique de début de syllabe. Selon Kochetov (2002), cette dernière observation est valable pour les autres langues slaves. Le fait que le système consonantique du russe présente une distinctivité ±PAL, pourrait contraindre d'autres cooccurrences entre consonne et voyelle d'une même syllabe que celles prédites par la théorie Frame/Content (MacNeilage, 1998) et trouvées dans des inventaires de langues par MacNeilage & Davis, 2000 ; Vallée et al. 2009) : *La-Ce*, *Co-An* et *Ve-Po*. Ainsi, en russe, nous trouvons /e i/ privilégiées après les consonnes +PAL indépendamment de leur lieu d'articulation. Les *La* et les *Ve* -PAL favorisent les combinaisons avec les noyaux postérieurs. Ce résultat confirme ceux de Kochetov (2002) et Pirogova (2018). Les *Co* ±PAL sont trouvées favorisées dans plusieurs combinaisons avec des noyaux vocaliques comme ce que pointait Pirogova (2018).

Notre corpus de lemmes renvoie aux tendances des autres études basées sur d'autres types de corpus qui permettent de considérer les mots en contexte, ce qui indique que les propriétés phonotactiques du lexique sont proches de celles des mots en contexte. Une transcription qui intègrerait des phénomènes d'allophonie fréquents pourrait sans doute permettre d'affiner les tendances observées à ce jour. Une étude plus complète est en cours avec la recherche de régularités phonotactiques dans les clusters et entre syllabes consécutives. Au-delà d'une description de la phonotaxe du russe, notre étude s'inscrit dans une perspective typologique des propriétés phonotactiques universelles. Dans le cadre de la linguistique russe, les résultats obtenus sont utilisables dans des paradigmes de phonotactique prédictive pour le TAL ou de tests psycholinguistiques.

Remerciements

Ce travail a bénéficié du soutien financier de l'IRS IDEX ComUE UGA projet PALGEST.

Références

AVANESOV R. (1956). *Fonetika sovremennogo russkogo literaturnogo jazyka*. Moscow: MUP.

AVANESOV R. (1972). *Russkoe literaturnoe proiznoshenie*. Moscow: Prosveshchenie.

DEHAENE-LAMBERTZ G., DUPOUX E. & GOUT A. (2000). Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of Cognitive Neuroscience*, 12(4), 635-647.

EYCHENNE J. (2015). De l'émergence des contraintes phonotactiques en français. *Langages*, 2, 73-90.

GATHERCOLE S. E., FRANKISH C. R., PICKERING S. J. & PEAKER S. (1999). Phonotactic influences on short-term memory. *J. of Experimental Psychology: Learning, Memory, and Cognition*, 25(1), 84-95.

GOLDRICK M. & LARSON M. (2008). Phonotactic probability influences speech production. *Cognition*, 107(3), 1155-1164.

HAYES B. & WILSON C. (2008), "A maximum entropy model of phonotactics and phonotactic learning", *Linguistic Inquiry* 39, 379-440.

ISKAROUS K. & KAVITSKAYA D. (2018). Sound change and the structure of synchronic variability: Phonetic and phonological factors in Slavic palatalization. *Language*, 94, 1-41.

JUSCZYK P. W., LUCE P. A. & CHARLES-LUCE J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5), 630–645.

KENSTOWICZ M. (2010). Loan phonology and Enhancement. In Kang Y.-S. et al., Eds., *Proceedings of the Seoul International Conference on Linguistics, Universal Grammar and Particular*

Languages, p. 104–112, Seoul, South Korea: Linguistic Society of Korea.

KOCHETOV A. (2002). *Production, perception and emergent phonotactic patterns: A case of contrastive palatalization*. New York: Routledge.

LIASHEVSKAIA O. N. & SHAROV S. A. (2009). *Tchastotnyi slovar sovremennogo russkogo iazyka (na materialakh Natsionalnogo korpusa russkogo yazika)*. Moscow: Azbukovnik.

MACNEILAGE P. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21(04), 499-511.

MACNEILAGE P. & DAVIS B. (2000). On the origin of internal structure of word forms. *Sciences*, 288, 527-531.

MADDIESON I. & PRECODA K. (1992). Syllable structure and phonetic models. *Phonology*, 9, 45-60.

MATTYS S. L. & JUSCZYK P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78, 91–121.

NAJAFIAN M., SAFAVI S., WEBER P. & RUSSELL M. J. (2016). Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems. In *Odyssey 2016,* p. 132-139, Bilbao, Spain.

PESHKOVSKII A. M. (1925). Sbornik statei: Metodika rodnogo iazyka, lingvistika, stilistika, poetika. Moscow: Gosizdat.

PIROGOVA N. K. (2018). *Konsonantizm russkogo jazyka*. Moscow: MAKS Press.

REMNEVA M. L. (2012). *Staroslavianskii iazyk*. Moscow: Izdatelstvo Moskovskogo universiteta.

ROUSSET I. (2004). *Structures syllabiques et lexicales des langues du monde. Données, typologies, tendances universelles et contraintes substantielles*. Thèse de doctorat, Université Grenoble III - Stendhal, Grenoble.

SHEVELOV G. Y. (1965). *A prehistory of Slavic: The historical phonology of Common Slavic*. New York: Columbia University Press.

SEGUI J., FRAUENFELDER U. H. & HALLE P. (2002). Les contraintes phonotactiques conditionnent la perception de la parole : implications pour les traitements lexicaux et sous-lexicaux. In E. Dupoux, *Les langages du cerveau*, p. 199-211, Paris : O. Jacob.

SRIVASTAVA B. M. L., VYDANA H., VUPPALA A. K. & SHRIVASTAVA M. (2017). Significance of neural phonotactic models for large-scale spoken language identification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, p. 2144-2151, Anchorage, Alaska, USA: IEEE.

STORKEL H. L. & ROGERS M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition. C*linical linguistics & phonetics*, 14(6), 407-425.

VALLEE N. & ROUSSET I. (2004). Indices en typologie des structures lexicales et syllabiques pour la discrimination et l'identification des langues. In *Actes du colloque Identification des langues et des variétés dialectales par les humains et par les machines*, p. 37-42, Paris : ENST.

VALLÉE N., ROSSATO S. & ROUSSET I. (2009). Favoured syllabic patterns in the world's languages and sensorimotor constraints. In Pellegrino F. et al., Eds., *Approaches to Phonological Complexity*, p. 111-139, Berlin : Mouton de Gruyter.

VALLEE N. (2017). *Phonologie et capacités sensorimotrices : de la syllabe au phonème*. Habilitation à diriger des recherches, Université Lumières Lyon 2, Lyon.

WARKER J. A. & DELL G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 387–398.

ZHU D. & ADDA-DECKER M. (2006). Language identification using lattice-based phonotactic and syllabotactic approaches. In *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*, p. 1-4, San Juan (Puerto Rico): IEEE.

ZINDER L. R. (1979). *Obshchaia fonetika*. Moscow: Vysshaia shkola.