Academic Journal of Modern Philology

Patrice Pognan
Plidam, Inalco, Paris & Úfal, MFF UK, Praha

ISSN 2353–3218
Vol. 9 (2020)
Special Issue
s. 155–169

Analyse morphologique automatique du tchèque, mais où est le dictionnaire ?

Automatic Morphological Analysis of Czech: Where is the Dictionary?

Abstract

Presented here is a system for an automatic analysis of Czech morphology mainly based on pattern recognition of linguistic forms according to the linguistic Prague school's paradigm "form – value." Taking into account the text and the context allows for the processing of "big data." This system is based on the concept of calculability. Czech has a very high degree of calculability thanks to a very favorable phonological evolution. Furthermore, the correspondence between what is said and what is written is exact. The experience drawn from this work made it possible to design a grammar course taught for several decades to the satisfaction of the learners. However, the feasibility of this analysis would be demonstrated at best by making entries to a Czech general dictionary.

Keywords: Czech language, morphology, phonology, calculability, processing without dictionary

De tous les outils linguistiques, le dictionnaire sous toutes ses formes est certainement celui qui a l'existence la plus évidente, certainement bien avant les grammaires. Monolingue, il est symbole identitaire d'une communauté linguistique, ce qui est particulièrement sensible pour les langues ayant un rapport fort à l'oralité (berbère, langues d'Afrique noire). Il est toujours élément de référence à sa propre langue ou à une langue acquise à un haut niveau, mais aussi outil d'apprentissage. Sa notoriété publique est quasi évidente. Bilingue, il occupe chez les spécialistes de langues, enseignants, apprenants, traducteurs, interprètes, rédacteurs de toutes sortes,… une place prépondérante.

Les systèmes de traitement automatique des langues et des textes n'ont pas su s'abstraire de ce qui semble être une condition sine qua non, à savoir l'usage de dictionnaires exhaustifs et souvent de forme complexe. Nous présentons ci-dessous une illustration d'un tel dictionnaire où sont données, codées, toutes les formes fléchies. Il s'agit du dictionnaire Polimorf destiné au traitement automatique du polonais :

głoskowsko	głoskowski	adja
głoskowsku	głoskowski	adjp
głoskowskich	głoskowski	adj:pl:acc:m1.p1:pos
głoskowskie	głoskowski	adj:pl:acc:m2.m3.f.n1.n2.p2.p3:pos
głoskowskim	głoskowski	adj:pl:dat:m1.m2.m3.f.n1.n2.p1.p2.p3:pos
głoskowskich	głoskowski	adj:pl:gen:m1.m2.m3.f.n1.n2.p1.p2.p3:pos
głoskowskimi	głoskowski	adj:pl:inst:m1.m2.m3.f.n1.n2.p1.p2.p3:pos
głoskowskich	głoskowski	adj:pl:loc:m1.m2.m3.f.n1.n2.p1.p2.p3:pos
głoskowscy	głoskowski	adj:pl:nom.voc:m1.p1:pos
głoskowskie	głoskowski	adj:pl:nom.voc:m2.m3.f.n1.n2.p2.p3:pos
głoskowską	głoskowski	adj:sg:acc:f:pos
głoskowskiego	głoskowski	adj:sg:acc:m1.m2:pos

Fig. 1 : extrait du dictionnaire Polimorf

Il en résulte très souvent qu'une analyse morphologique automatique n'est qu'une simple consultation de dictionnaire, ce qui a pu faire dire, beaucoup trop rapidement, que le problème ne présentait plus de difficulté ni d'intérêt.

L'alternative courante actuelle est représentée par les techniques d'apprentissage sur des bases statistiques qui, à mon avis, nous privent des applications didactiques que l'on est en droit d'attendre de tels instruments d'analyse de la langue.

En tant que linguiste et enseignant de langues, je propose une autre approche de l'analyse automatique que l'on pourrait presque qualifier de grammaire automatique ou automatisée d'une langue. Il n'y a pas de dictionnaire, mais une parfaite illustration du paradigme de l'École de Prague: l'exploitation totale de la ***correspondance forme - valeur*** suffisante pour assurer l'intégralité de l'***analyse morphosyntaxique de la langue tchèque***. Il faut reconnaître que cette langue est un cas privilégié ayant bénéficié d'une évolution favorable comme je vais le montrer. Cependant, une langue qui a priori ne s'y prêtait pas, la forme kabyle du berbère, donne des résultats très intéressants (Pognan, Sadi & al. 2018).

La reconnaissance automatique de formes linguistiques, le recours au texte et au contexte sont, ensemble, des techniques adéquates au traitement de grandes quantités de données textuelles.

En fonctionnant sans dictionnaire, un tel système appliqué à de grandes quantités de textes permet de construire les entrées d'un dictionnaire de la langue, accompagnées de la majorité des étiquettes morphosyntaxiques associées. Les résultats peuvent être organisés de diverses manières débouchant sur une structure de base de données ou même directement en une édition Word dont la mise en page précise peut être réalisée à l'aide du langage de programmation intégré VBA[1].

Le présent article se veut avant tout un plaidoyer pour le ***concept de calculabilité*** d'une langue. Il est possible de faire l'analogie avec l'arithmétique qui longtemps a été réalisée à l'aide d'abaques et non point calculée. Ce qui me semble important, c'est de montrer que l'on peut quitter les abaques (le fameux « dictionary look-up ») pour un vrai calcul sur la langue. J'ai appuyé l'élaboration des méthodes de calcul

1 Pratiquement personne ne prend conscience, ni même ne sait que Word, et de manière générale Microsoft Office, dispose d'un environnement de programmation orienté objet très puissant. L'inconvénient est que la documentation afférente est difficilement accessible.

sur des études du système linguistique des langues slaves de l'Ouest, sur la grammaire historique du groupe et sur la calculabilité induite dans les différentes langues actuelles.

1. D'où vient la calculabilité du tchèque ?

La calculabilité du tchèque est issue de la conjonction de ***trois phénomènes*** qui ont un effet convergent.

1.1. Ecriture (évolution de la graphie)

Le premier d'entre eux est l'évolution de la graphie, puis de l'orthographe que l'on peut diviser en plusieurs périodes.

La première est celle de l'Empire de Grande Moravie (limite des 8ème et 9ème siècles jusqu'en 907) pendant laquelle Constantin «le Philosophe», fin lettré forgea l'écriture glagolitique. Il la constitua à partir des symboles de l'onciale grecque qui permettaient la transcription des phonèmes slaves. Il la compléta, pour les phonèmes propres au slave, par des signes inspirés des alphabets orientaux qu'il connaissait (en particulier, des régions au Nord du Caucase) et non pas en combinant des caractères préexistants. Le trait de génie de Constantin est d'avoir créé une écriture dans laquelle un graphème unique renvoyait à un phonème et inversement. L'écriture cyrillique a conservé le principe de correspondance phonie - graphie ainsi que l'héritage des caractères glagolitiques, simplifiés («ч», «ш», «ж»,...).

La seconde période, environ 550 ans après l'oeuvre de Constantin, débute avec l'introduction par Jan Hus du principe de l'écriture diacritée. Jan Hus marque la longueur des voyelles par un accent aigu sur le graphème représentant la voyelle et les consonnes molles par un point qui deviendra plus tard le « háček ». Hus, connu surtout comme théologien, était un excellent linguiste et connaissait l'écriture glagolitique. Il a retenu le principe d'univocité entre la graphie et la phonie à l'exception près du digraphe « ch », certainement à cause de l'habitude de l'allemand.

L'écriture diacritée du tchèque a été largement reprise par les Slaves écrivant en caractères latins, en premier lieu les Slovaques, mais aussi les Sorabes, les Slovènes et les Croates. Les Polonais ont peu de caractères diacrités en complément de groupes de caractères tels que « cz », « sz », « rz », « ść » ou « szcz ». Les Baltes (Lituaniens, Lettons) ont repris le système tchèque de manière explicite.

De l'histoire de la construction de l'écriture, le tchèque conserve des caractéristiques importantes pour la calculabilité. En premier lieu, l'écriture actuelle maintient la bi-univocité écrit – oral / graphème - phonème. Ensuite, et c'est extrêmement important et fait souvent la différence avec d'autres langues slaves écrites en caractères latins, l'écriture tchèque est phonologique et prend en compte une opposition de sonorité qui est gérée par deux règles phonétique simples, l'assourdissement en fin de mot et l'assimilation régressive de sonorité.

Cette écriture renvoie à une certaine étymologie comme dans le cas du maintien des deux « i » (i/í - y/ý) alors qu'ils sont prononcés de manière identique en tchèque contrairement au polonais ou au russe.

Non moins important est le maintien dans l'écriture, grâce à des signes diacritiques, d'oppositions fondamentales pour la calculabilité : l'opposition de mouillure marquée par le « háček » et l'opposition de longueur marquée par la « délka ».

1.2. Aménagements lexicaux

Le second est le résultat de ce que l'on pourrait appeler avant l'heure des « aménagements linguistiques »: la réorganisation du lexique et notamment des morphèmes suffixaux à l'époque du « Renouveau national » qui va de la fin du 18ème siècle jusqu'à l'avènement de la République tchécoslovaque après la première guerre mondiale. On y assiste à une série de transformations importantes de la langue tchèque sous l'influence de nombreux savants de Dobrovský jusqu'à Gebauer, en particulier au niveau d'une ***mise en ordre morpho-sémantique de la suffixation***. La reprise de suffixations délaissées, voire la création de nouveaux suffixes font de cette période une étape importante pour la calculabilité de la langue.

De nombreux suffixes réfèrent actuellement (avec très peu d'exceptions) à un champ sémantique déterminé:

- « -iště », à deux exceptions près renvoie à un lieu ouvert: « hřiště » est un lieu ouvert où l'on peut « hr- » jouer = terrain de jeu,
- « -ovna » à un lieu fermé: « knihovna » la bibliothèque, « -árna » et « -írna » à des lieux où l'on vend, fabrique quelque chose, où l'on se livre à une activité en général: « lékárna » la pharmacie, « tavírna » la fonderie, …

Bareš, dans son travail de thèse sur les noms en « -dlo » (Bareš 1970), montre de manière explicite le travail de « nettoyage » et d'organisation qui a été conduit sur les suffixations comme on peut l'observer dans le schéma ci-dessous pour le suffixe -dlo:

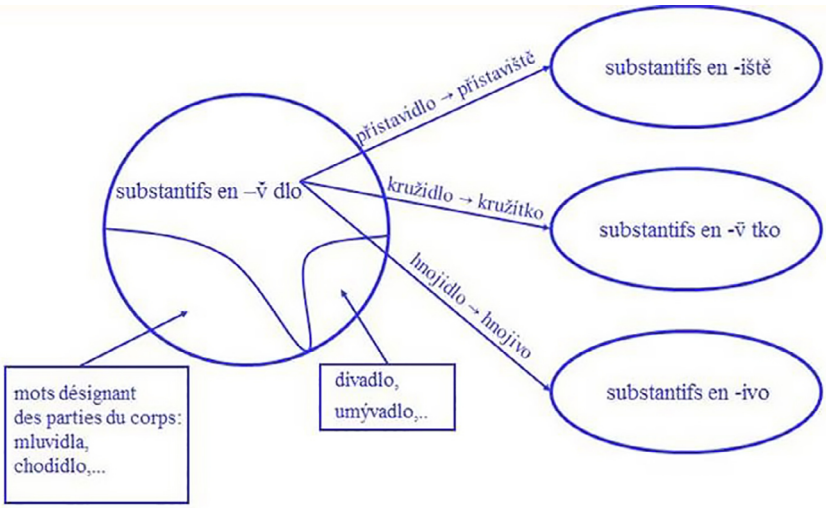

Fig. 2 : aménagement des suffixations - exemple du suffixe -dlo

C'est ce travail « d'aménagement linguistique » qui permet d'avoir aujourd'hui une relation bi-univoque importante entre certains suffixes et une valeur sémantique déterminée. Comme on peut le voir sur le schéma ci-dessus, les valeurs sémantiques principales des suffixes ont été renforcées par des substitutions de suffixes :

- « přistavidlo » (*embarcadère*) représente un lieu ouvert et va devenir « přistaviště » (cf. supra suffixe « -iště »).

- « kružidlo » (*compas*), petit instrument va devenir « kružítko ».
- « hnojidlo » (*engrais*) devient « hnojivo » avec le suffixe « -ivo » qui désigne des substances qui s'écoulent généralement en liquide ou en granulé.

Le suffixe « voyelle brève-dlo », en dehors de quelques exceptions :

- « divadlo » (*théâtre*), « umývadlo » (*lavabo*) ou
- une dizaine de termes désignant des organes ou des parties du corps tels que « mluvidla » (*organes phonatoires*), « rodidla » (*organes génitaux*), « kusadla » (*mandibules*), « makadla » (*palpes*), « tykadla » (*antennes*), « chapadlo » (*tentacule*), « chodidlo » (plante du pied),...

renvoie désormais de manière univoque à une machine :

« letadlo » [let – a-dlo]: une machine permettant de voler ➞ *avion*,
« vozidlo » [voz – i-dlo]: une machine permettant de transporter ➞ *véhicule*,
« čerpadlo » [čerp – a-dlo]: une machine permettant de puiser ➞ *pompe*,...

L'analyse automatique de tels suffixes permet non seulement une analyse morphologique, mais elle permet de plus de donner une classification sémantique sommaire, ce qui joue un rôle déterminant dans des applications telles que l'indexation de textes techniques, scientifiques ou médicaux, les programmes de veille scientifique, ...

1.3. Evolution phonologique

Le troisième est une évolution phonologique dont les résultats offrent des régularités appropriées à des opérations de calcul. Remarquons au passage que l'évolution phonologique des langues slaves de l'Ouest (polonais, bas-sorabe, haut-sorabe, tchèque, slovaque) permet de comprendre par quels phénomènes ces différentes langues se sont écartées progressivement les unes des autres.

Il est habituel de présenter l'évolution phonologique du tchèque en trois périodes (Lamprecht, Šlosar, Bauer 1986) :

I.	**Du protoslave à la fin du 10ème siècle:**
1.	Métathèse des liquides (évolution de: *tort, tolt, tert, telt*)
	1.a. métathèse en o. et 1.b. métathèse en e
2.	Contraction
3.	Disparition et vocalisation des jers
4.	Evolution des nasales
II.	**De la fin du 10ème siècle à la fin du 14ème siècle**
5.	Passage g ⇨ h
6.	Evolution du r mouillé en ř : r' ⇨ ř
7.	Transformations 7.a : 'a ⇨ ě et 7.b: 'u ⇨ i
8.	Dépalatalisation
III.	**De la fin du 14ème siècle à la fin du 16ème siècle**
9.	Transformation du « u long » en diphtongue: ú ⇨ ou
10.	Contraction ie ⇨ í
11.	Contraction (ó ⇨) uo ⇨ ů
12.	Changement aj ⇨ ej

Fig. 3 : l'évolution phonologique du tchèque

Nous nous référerons à certains de ces phénomènes lors de la présentation d'exemples de calculs, notamment dans la reconnaissance des emprunts dans des textes tchèques.

2. Des exemples de calculabilité du tchèque

Nous présenterons deux exemples de calculabilité. Nous donnerons une démonstration de ce qui peut être fait pour reconnaître des emprunts, ce qui révèlera la raison d'un taux élevé de reconnaissance. Nous fournirons également un rapide aperçu de ce qui peut être fait dans le domaine de la morphologie tchèque.

2.1. Reconnaissance automatique d'emprunts

Certaines évolutions sont favorables à un traitement de la langue par reconnaissance de formes. Ainsi, l'évolution de « g » devenu « h » en tchèque, en slovaque et en haut-sorabe (II. 5. de l'évolution phonologique) permet d'utiliser le graphème « g » comme l'une des nombreuses marques de reconnaissance d'un mot d'origine étrangère.

Un autre exemple marquant est l'évolution du « o » long tchèque (ó) qui par l'intermédiaire d'une diphtongue (uo/uó) est passé au « u » long avec kroužek (ů) (III. 11. de l'évolution phonologique). Il en découle que tout mot contenant un « ó » ne peut pas être tchèque, car il y aurait « ů » à la place. On le voit aisément dans les mots suivants qu'il est inutile de traduire : *acetón, ambiciózní, chór, cirhóza, medailón, mykóza, ozón, penzión, prognóza, sezóna, skleróza, viróza*…

Il est également intéressant de constater que la reconnaissance automatique des emprunts est presque totale. Cela provient du fait que l'origine étrangère est généralement marquée par plusieurs critères comme nous allons le constater sur ces exemples :

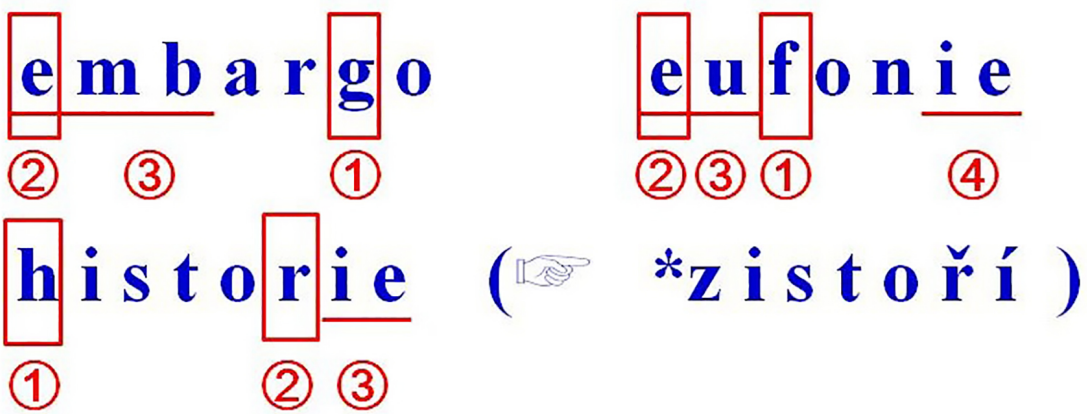

Fig. 4 : redondance des critères de reconnaissance des emprunts

Ces « locomotives à vapeur » montrent bien la redondance des critères de reconnaissance.

Dans le premier cas, « embargo » le « g » suffit à noter l'origine étrangère et dans l'algorithme de reconnaissance les autres critères ne seront pas testés.

Le deuxième critère de reconnaissance présent est le « e » en tant que premier caractère du mot. Déjà en 1792, Josef Dobrovský notait dans « Dějiny české řeči a literatury » : « Slovan nechává samohlásku 'a' na začátku slova zřídka kdy, 'e' pak nikdy bez joty. Říman říká 'est', Slovan 'jest' » (Un Slave ne laisse que très rarement la voyelle « a » en tête de mot, le « e » jamais sans yod. Un Latin dit « est », un Slave « jest »). De fait, en tchèque, à une douzaine d'exceptions près (en majorité des mots outils) pour « a » (ce qui vaut dans toutes les langues slaves de l'Ouest), à une exception pour « e » et deux exceptions près pour « i », tout mot ayant « a », » e » ou « i » en tant que caractère de tête est d'origine étrangère. En haut-sorabe, tout mot débutant par une voyelle est d'origine étrangère, sauf cas de métathèse non réalisée : « ert » (bouche) qui est « ret » en tchèque (lèvre).

Nous avons comme 3ème critère de reconnaissance la suite « emb » qui est la transcription de la nasale française avec la règle du « mpb » connue de tous les petits écoliers français : « n » devant « m », « p » ou « b » devient « m ». En tchèque, la conservation de la forme des nasales françaises se fait suivant l'un des deux patrons suivants:

- un patron général V – « N » – C_1 où C_1 représente toute consonne sauf « m », « p » ou « b ».
- un patron particulier V – « M » – C_2 où C_2 représente les consonnes « p », « b » et « f ». « m » disparaît (par rapport à la situation française), parce que dans les emprunts, sauf en ce qui concerne les noms propres, les géminées sont réduites à un seul caractère. D'autre part, « f » apparaît dans les emprunts d'origine grecque en tant que transcription de « ph ».

Dans le 2ème cas, « eufonie », « f » est un graphème étranger en dehors de sa présence dans les verbes tchèques « doufat » (espérer) et « zoufat » (désespérer) et dans la racine « fuk » / « fouk » (souffler). Nous avons ensuite, à nouveau, la présence de « e » en tête.

Les critères 3 et 4 sont intéressants. Le tchèque a connu deux phénomènes de contraction, le premier aux environs des 9ème et 10ème siècles (I. 2. de l'évolution phonologique), le second au 15ème siècle (III. 10. de l'évolution phonologique), qui ont eu pour effet de transformer un certain nombre de diphtongues en voyelles longues. Il en résulte que seule la diphtongue « ou » est autochtone. Même « au » et « eu » marquent une origine étrangère. Ici, les diphtongues « eu » et « ie » marquent, l'une et l'autre, une origine étrangère. Notons qu'il demeure un parallélisme important entre slovaque (« ie ») et tchèque (« í »), par exemple « viera » face à « víra » (*foi*).

Le 3ème exemple illustre un autre mode de reconnaissance des emprunts. Les deux exemples précédents ont été reconnus grâce à des « marqueurs d'origine étrangère ». Dans celui-ci, en revanche, il s'agit du non-respect de règles de fonctionnement tchèque, ici du non-respect de la palatalisation. Si le mot « historie » avait été tchèque, l'évolution phonologique l'aurait transformé en « zistoří », « h » serait devenu « z » et « r » aurait été palatalisé en « ř ». Nous aurions également assisté à la transformation de la diphtongue « ie » en voyelle longue « í ».

2.2. Analyse morphologique automatique

Le tchèque, langue à flexion externe, marque la longueur des voyelles. La longueur participe à la calculabilité de manière importante comme on peut le voir avec cet exemple: le « -ý » : en tant que dernier caractère d'un mot, donne, à 3 exceptions près, une valeur univoque d'adjectif, forme longue, dur au masculin singulier nominatif ou accusatif inanimé. Cette seule opération permet de déterminer le radical d'un ensemble d'environ 16000 adjectifs.

La calculabilité élevée du tchèque est due aussi au fait qu'en tchèque, « la morphologie est partout », comme nous allons le voir ci-dessous :

Le tchèque possède une série de préfixes avec deux valeurs (brève et longue) : « do » / « dů », « na » / « ná », « při » / « pří », « pro » / « prů », « po » / « pů », « u » / « ú », « vy » / « vý », « za » / « zá ». La série longue est marquée : à une dizaine d'exceptions près, elle n'apparaît que dans des mots de nature nominale, c'est-à-dire non verbale. Cela apporte dans le processus de calcul pour l'analyse des valeurs morphologiques complémentaires qui permettent de lever des ambiguïtés non résolues par la seule désinence. L'opposition entre « výpočtu » (substantif au génitif, datif ou locatif singulier) et « vypočtu » (verbe à la 1ère personne du singulier du futur [aspect perfectif]) est une excellente illustration de la situation : seule la longueur différencie les deux mots. En termes d'analyse automatique, le substantif est reconnaissable et analysable en tant que tel grâce au préfixe long « vý ». Comme pratiquement toujours en linguistique, la situation n'est pas réversible et le préfixe « vy » ne désigne pas nécessairement un verbe. La série brève des préfixes est utilisée aussi bien dans le système nominal que verbal.

Le schéma qui suit montre comment le préfixe long « vý » lève l'ambiguïté dans le cas de la désinence « á », aussi bien désinence d'adjectif que de verbe. Dans ce cas, c'est une valeur morphologique située en partie préfixale qui permet de mener l'analyse à terme.

De manière encore plus surprenante, c'est une valeur située au centre, en tant que support vocalique de la racine, qui peut lever l'ambiguïté de la désinence. Nous obtenons ici un véritable schème comme dans les langues sémitiques : « í ____ á » qui ne connaît que quelques interférences avec des adjectifs (au féminin nominatif singulier ou au neutre nominatif et accusatif au pluriel). La valeur donnée par ce schème est « verbe, 5ème classe, imperfectif, 3ème personne du singulier, présent ».

Ces deux procédés permettent d'analyser correctement deux formes proches telles que « výborná » (adjectif) et « vybírá » (verbe).

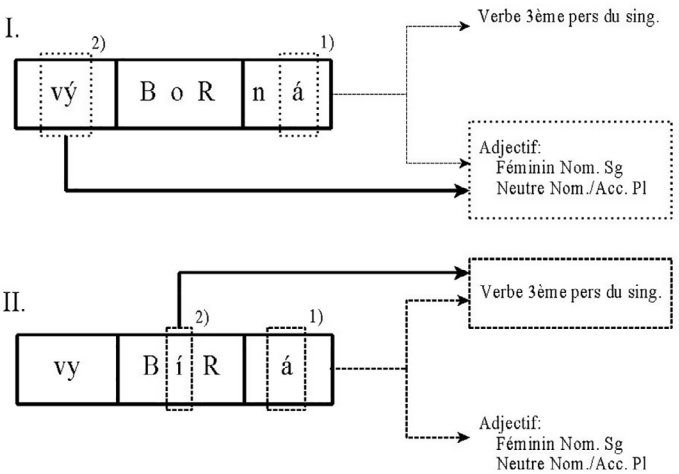

Fig. 5 : les valeurs morphologiques ne sont pas que dans les désinences

Je n'ai présenté ici qu'une petite minorité d'éléments qui participent à une analyse heuristique automatique de la morphologie tchèque. Mais étant donné qu'il s'agit d'une heuristique déterministe appuyée sur des faits saillants, les résultats de l'analyse morphologique peuvent ne pas être complets. L'étape d'analyse syntaxique qui repose sur une grammaire dirigée par un arbre de dépendances canonique y remédie en très grande partie en prenant en compte un contexte structuré.

En tant qu'enseignant de la grammaire tchèque à des adultes non-tchécophones, l'univocité de certaines désinences m'a frappé. C'était le cas de « -ý » (cf. supra), mais c'est également le cas de « -ů », désinence du génitif pluriel de tous les masculins. Il s'agit de l'exploitation d'un morphème marqué, désinence qui exprime une valeur grammaticale univoque, donc totalement fiable. A partir de cette constatation, il a été possible d'avoir recours au texte pour déterminer des valeurs morphologiques, ce qui ouvre le ***traitement de grandes masses de données textuelles***.

3. Recours complémentaire aux paradigmes de flexion nominale

La flexion substantivale est abordée par 14 tableaux de désinences, y compris les variantes: 6 pour les masculins et 4 pour les féminins et les neutres. Pour l'ensemble des substantifs masculins, la désinence marquée est le « ů » dont la valeur grammaticale (grammatème) est : « Substantif Masculin Pluriel Génitif ».

Cette désinence permet de reconnaître tous les modèles de masculins sans savoir toutefois à quel paradigme ils appartiennent. En particulier, on ne sait pas si le substantif est un masculin consonantique ou vocalique.

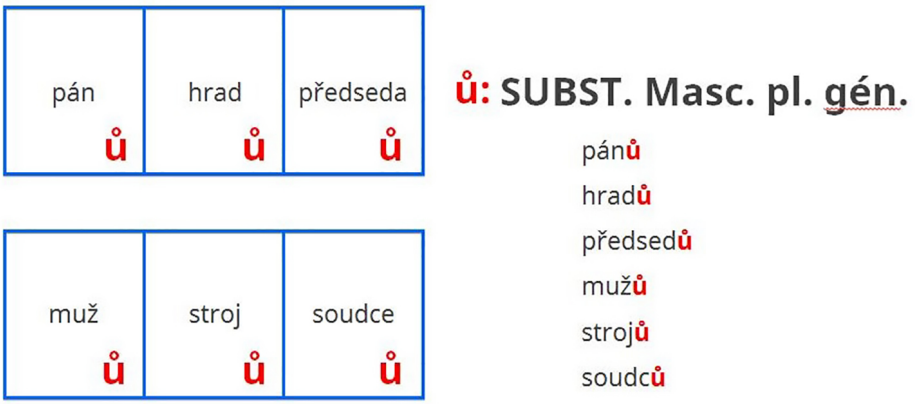

Fig. 6 : les paradigmes de flexion des substantifs masculins

Par des séries de raisonnements sur les modèles de flexion, le texte va pouvoir apporter les éléments qui, étape par étape, vont déterminer le modèle et les valeurs afférentes. Le système réalisé en Python fonctionne en 5 cycles successifs et utilise des raisonnements qui s'appuient sur la connaissance des modèles et sur un cheminement à l'intérieur de ceux-ci. Je ne donnerai que quelques exemples évidents.

Les substantifs candidats à un traitement sont repérés par la désinence non ambiguë « ů » et reçoivent le grammatème afférent : substantif masculin au génitif pluriel. A ce stade, tous les mots

relevés appartiennent à l'ensemble des 6 modèles. Le radical (sans le « ů ») est conservé pour des explorations ultérieures.

Si le radical obtenu se retrouve dans le texte comme mot autonome, alors il ne peut pas être un masculin vocalique et n'appartient plus qu'à l'ensemble des 4 modèles consonantiques.

Si ce radical est rencontré dans le texte avec la désinence « ou », alors il est un masculin vocalique dur de modèle « předseda » (président) et reçoit le grammatème afférent. Le radical est ensuite testé dans le cycle suivant avec toutes les désinences du modèle y compris des sous-modèles « husita » et « cyklista » (nominatif pluriel en « é »).

De manière beaucoup plus inattendue, un radical de masculin avec la désinence « ě » renvoie uniquement au modèle « hrad » (château fort) (« na hradě » : au château).

A chaque étape de chaque cycle, le système calcule la forme de radical palatalisée et/ou la forme avec « e » intercalaire à partir desquelles il pourra rechercher un certain nombre de formes casuelles :

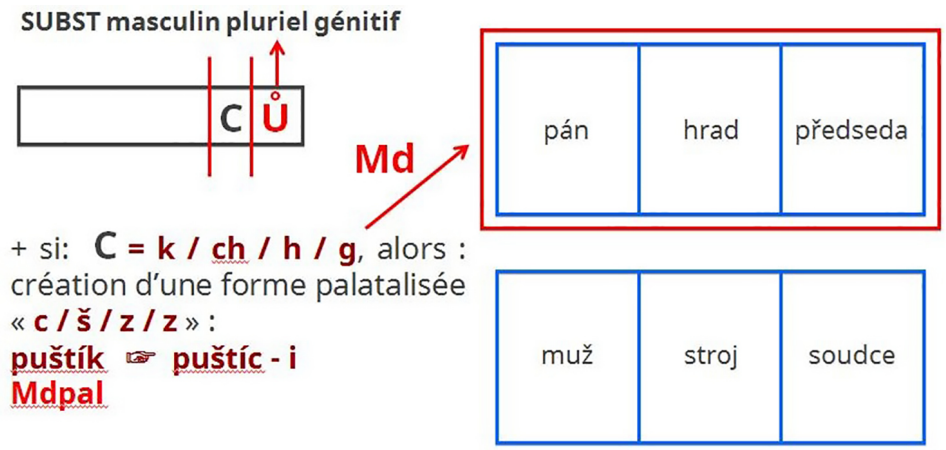

Fig. 7 : détermination des masculins durs et génération de la forme palatalisée

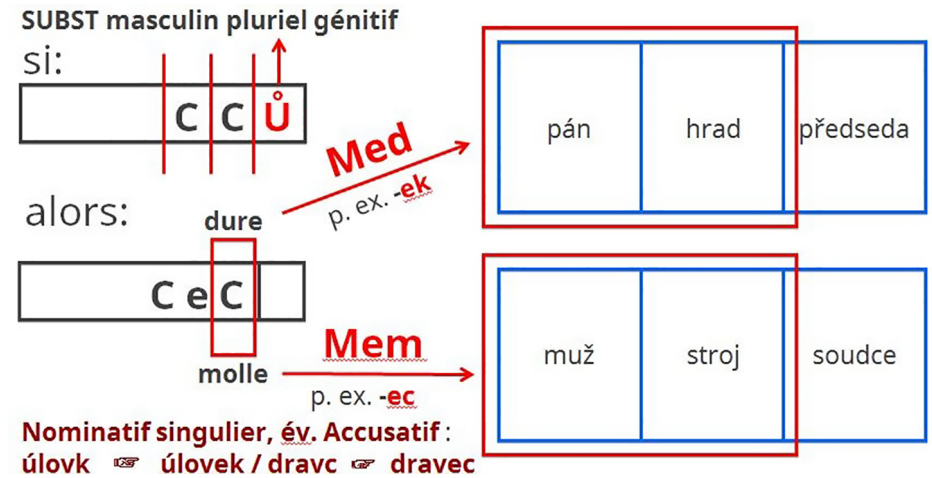

Fig. 8 : génération des formes avec « e » intercalaire

L'utilisation de désinences marquées est beaucoup plus simple pour le féminin et le neutre, car on ne peut utiliser ici que des désinences afférentes à un seul paradigme.

Pour les féminins, les désinences sûres sont celles de l'instrumental pluriel « -ami », « -emi » et « -ěmi » respectivement pour les modèles « žena » (féminin dur) (femme), « duše » (féminin mou) (âme) et « píseň » (féminin consonantique mou) (chanson). Pour le féminin dur, la désinence du locatif pluriel « ách » est valable tant que l'on ne traite pas des textes de la langue parlée, où cette désinence apparait également au masculin (« ve vlakách » au lieu de « ve vlacích » et au neutre (« v střediskách » au lieu de « ve střediscích ».

Pour le neutre, le modèle « stavení » (construction) est traité dans le cadre de l'analyse morphologique. La désinence du datif pluriel « -ům » combinée à l'augment du pluriel « -at- » (terminaison du mot en « -atům ») permet de traiter le paradigme « kuře » (poulet), un modèle à augment concernant au plan de la sémantique essentiellement des petits d'animaux.

Le traitement des féminins et des neutres ne comportent que deux cycles. Les autres occurrences de cas sont recherchées au sein de ce second cycle. Y sont générées les formes palatalisées (le datif et le vocatif singulier du modèle « žena ») et les formes avec un « e » intercalaire (génitif pluriel du modèle « žena » dont le calcul est complexe, nominatif et accusatif singulier du modèle « píseň », mais aussi d'un sous-modèle de « duše » « ploutev » (nageoire) reconnu par la forme « ploutvemi »). Au neutre, il est nécessaire de gérer les changements en fin de radical : à partir de « štěňatům » il est nécessaire de générer la forme « stěně » (chiot).

4. Présentation de résultats

Voici un extrait de résultats obtenus sur l'un des textes d'essai consacré à une variété de chouette (Strix uralensis), en tchèque « puštík bělavý », dite simplement « chouette de l'Oural » en français :

```
puštík                ^V      puštík
                puštík        puštíc          puštíc
                SUBST         MCda    S       N       N       G
obecný                ^V++    obec-n-ý obec
                obecn obecný
                ADJd  MSNAV   M       S       NAV     NAV
a                     ^V
                a
                CONJc
při                   ^V
                při
                PREP                                          L
pozorování            P1      V+      po-ZoR-ov-án-í  ZoR
                pozorován     pozorování
                SUBST NSNGDAL.NPNGA                                   G
je
zřetelně
patrný                ^V++    PaTR-n-ý PaTR
                patrn patrný
                ADJd  MSNAV   M       S       NAV     NAV
```

Fig. 9 : quelques résultats de l'analyse morphologique

Quelques commentaires sont nécessaires pour permettre la lecture des résultats. Les catégories lexicales sont données en majuscules : VERB pour verbe, SUBST pour substantif, ADJ pour adjectif, divisé en ADJd pour adjectif dur et ADJm pour adjectif mou, ADV pour adverbe, CONJ pour conjonction, PREP pour préposition. L'analyse donne toujours une dérivation : V = verbale ou ^V = non verbale (= nominale). Cela permet d'avoir une dérivation nominale pour SUBST, ADJ, ADV, …., une dérivation verbale pour les verbes (VERB), mais aussi pour V SUBST = substantif verbal, V ADJ = adjectif dérivé d'un verbe. Le genre est donné par « M », « F » ou « N », le nombre par « S » ou « P » (le duel « D » est réduit en tchèque à seulement quelques formes).

Les Tchèques ne dénomment pas les cas par leur nom, mais par le numéro d'ordre qu'ils leur ont donné : 1 = Nominatif, 2 = Génitif, 3 = Datif, 4 = Accusatif, 5 = Vocatif, 6 = Locatif, 7 = Instrumental. Pour des raisons de compréhension des résultats, nous avons remplacé ces chiffres par le caractère initial de chaque cas en majuscule.

L'analyse produit un paradigme, p. ex. MCda = Masculin Consonantique dur animé et les valeurs grammaticales afférentes S = singulier et N = nominatif. Cela peut être aussi ADJd suivi d'un grammatème MSNAV qui est automatiquement décomposé en genre M (masculin) au S (singulier) à un cas qui peut être NAV, c'est-à-dire nominatif ou accusatif ou vocatif.

Le premier mot « puštík » (la chouette en question) est reconnu par le programme de reconnaissance des substantifs masculins par l'intermédiaire de la désinence « -ů ». Cette forme est reconnue comme étant un substantif de dérivation non verbale de type consonantique, modèle dur animé (paradigme « pán »), valeur désignée par le symbole MCda. Il est au singulier (« S ») et au nominatif (« N »). Il requiert le génitif (« G ») pour toute complémentation. Les formes de palatalisations « puštíc » sont correctement générées.

L'autre substantif de l'extrait « pozorování » (*observation*) est un substantif verbal issu du verbe « pozorovat » (*observer*). Il est noté simultanément comme étant un processus de premier rang (P1) par rapport à des substantifs suffixés en « -ba » qui sont des « P2 », p. ex. « orba » (*labourage*) du verbe « orat » (*labourer*). Ces indications sont importantes pour pouvoir jouer avec les catégories lexicales dans une paraphrase ou dans une traduction. Elles sont primordiales dans des applications d'indexation des textes scientifiques, techniques ou médicaux. Le découpage morphématique en « po-zor-ov-án-í » et l'extraction de la racine « ZoR » sont corrects. Contrairement à d'autres langues parentes, le système linguistique verbal impersonnel du tchèque est complet[2] : tout gérondif présent ou passé peut engendrer un adjectif mou, tout participe peut engendrer un adjectif dur. Le participe passé passif a la particularité d'être à la base de son adjectif dérivé et du substantif verbal. Ainsi, à partir de l'infinitif « pozorovat », le participe passé passif est « pozorován » (fléchi en nombre et en genre, p. ex. contrairement au polonais où seul le neutre singulier en « -o » est attesté) qui donne directement le substantif verbal par ajout de la désinence neutre « -í » = « pozorování » (fléchi en tant que substantif neutre) et l'adjectif par suffixation et alternance de longueur : « pozorovaný » (*observé*) (fléchi en tant qu'adjectif dur). Si l'une des formes est reconnue, les autres formes peuvent être générées automatiquement. C'est une autre facette de la calculabilité linguistique. Ici, la forme du participe passé passif « po-zor-ov-án » a été correctement obtenue et permettra de reconnaître les formes de l'adjectif dérivé et même, lorsqu'il existe, l'adverbe, lui-même dérivé de cet adjectif verbal.

2 Ensemble, le bas-sorabe et le tchèque représentent l'intégralité du système linguistique slave de l'Ouest.

Les deux adjectifs durs « obecný » (*commun*) et « patrný » (*que l'on remarque*) sont correctement reconnus. Les découpages morphématiques et la racine « PaTR » sont corrects. Je n'ai pas encore pris de décision sur la forme de racine de « obec », en « obec », « oBC » ou même « oB », puisque le mot « obec » (*commune*) vient vraisemblablement de la préposition « ob » (*autour de*).

L'adverbe « zretelně » (*visiblement*) issu de l'adjectif prédicatif « zretelný » n'est pas traité, l'implémentation correspondante devant intervenir prochainement.

Une telle analyse, heuristique déterministe, se doit d'exploiter au maximum toutes les informations que l'on peut extraire à partir de la forme. C'est ainsi que sont mis en place des « limites », des « drapeaux » (« flags » en anglais) qui sont liées à la morphologie (cas voulu(s) par une préposition, génitif requis par un substantif à l'exception de quelques substantifs verbaux requérant le datif ou l'instrumental). Ces indications font partie des résultats de l'analyse phono-morphologique dont les sorties sont les entrées de l'étape syntaxique. Ces diverses limites permettent d'y calculer la portée (en anglais « scope ») de telle ou telle valeur casuelle et ainsi de définir des entités syntaxiques (syntagmes nominaux, propositions, …).

Conclusion : les applications privilégiées de la calculabilité

Il ressort de ces différentes analyses que le traitement du tchèque par l'exploitation de la forme est non seulement possible, mais qu'il présente de nombreux avantages, notamment pour l'enseignement du tchèque. L'expérience a montré que ce qui est calculable par la machine est utilisable pour la formation des apprenants et bien assimilé par ceux-ci.

Cette approche par la prise en compte de la grammaire synchronique et diachronique dans le cadre du système linguistique du groupe de langues offre en plus une ouverture vers le multilinguisme, au minimum vers une compréhension passive des autres langues du groupe.

Cependant, l'enjeu essentiel pourrait être désormais, à condition de pouvoir accéder à de très grands corpus, la réalisation des entrées[3] d'un dictionnaire à partir des mots extraits. Le résultat est un fichier brut (.txt) structuré « à plat » en Unicode UTF-8. Ce type de fichier peut être transformé en base de données de quelque type que ce soit ou exploité directement, notamment transformé en fichier Word dont toute l'organisation et la mise en page, même complexe, peut être réalisé à l'aide du langage de programmation orienté objet VBA (Virtual Basic for Applications). Le savoir-faire a été acquis grâce à la réalisation du « Dictionnaire raisonné berbère - français » (Taïfi, Pognan 2017). Ainsi, les en-têtes de page sont faites automatiquement avec la numérotation centrale et à gauche le premier élément de classification sur la page (la racine pour le dictionnaire principal, un mot pour les index) et à droite le dernier sur la page. Il en va de même avec le traitement des entrées : racines colorées en bleu pour une meilleure lisibilité, numéro de la racine monté en exposant, etc.

3 Pour toutes les entrées qui ont été obtenues dans l'analyse par l'intermédiaire d'un paradigme de flexion (cf. l'analyse des substantifs), il est simultanément possible de générer l'ensemble des formes à l'image de ce qui a été réalisé dans le Polimorf polonais. Mais l'avantage premier de l'analyse présentée est de ne pas avoir besoin de cumuler des formes comme données d'analyse. Ces formes sont reconnues par une algorithmique linguistique (***calculabilité de la langue***) à une seule condition : que le texte soit écrit en tchèque. L'avantage supplémentaire est que, n'étant pas lié à une base de données ou à un dictionnaire, ce système sait analyser la néologie et les emprunts. Etant donné le type d'analyse, plus le texte est long, plus le résultat est précis. C'est une analyse particulièrement adéquate pour les « big data », en l'occurrence une analyse de grands corpus textuels.

On voit ci-dessous à la fois l'en-tête de la page 1080, mais aussi la première racine WZR reportée à gauche dans l'en-tête sous la forme WZR[1] :

Fig. 10 : génération de l'en-tête et de la typographie du dictionnaire berbère - français

Bibliographie

Bareš, Rudolf (1970) *Die Nomina auf -dlo. Ein Beitrag zur tschechischen Wortbildung*. Meisenheim am Glan: Verlag Anton Hain.

Brodda, Benny (1983) "An Experiment with Heuristic Parsing of Swedish." [In:] Pise: *First Conference of the European Chapter of the Association of Computational Linguistics*; 66–73.

Dobrovský, Jozef ([1792] 1951) *Dějiny české řeči a literatury*. Praha: Československý spisovatel.

Havránek, Bohuslav, Jedlička Alois (1960) *Česká mluvnice*. Praha: SPN.

Källgren, Gunnar (1991) "Parsing without Lexicon: the MorP System." [in:] Berlin: *5th Conference of the European Chapter of The Association for Computational Linguistics*; 143–148.

Lamprecht, Arnošt, Šlosar Dušan & Bauer Jaroslav (1986) *Historická mluvnice češtiny*. Praha: SPN.

Mareš, František V. (2000) *Cyrilometodějská tradice a slavistika*. Praha: Torst.

Marvan, Jiří (2000) *Jazykové milénium. Slovanská kontrakce a její český zdroj*. Praha: Academia.

Mazon, André (1952) *Grammaire de la langue tchèque*. Paris: Institut d'Etudes Slaves.

Meillet, Antoine, Vaillant André (1965) *Le slave commun*. Paris: Honoré Champion.

Mistrík, Jozef (1983) *Moderná slovenčina*. Bratislava: SPN

Pleskalová, Jana (2001) *Stará čeština pro nefilology*. Brno: Filozofická fakulta Masarykovy Univerzity.

Pognan, Patrice (1999) "Histoire de l'écriture et de l'orthographe tchèques." [In:] *Histoire, Epistémologie, Langage*. Vol 21(1); 27–62.

Pognan, Patrice (2001) "Introduction aux systèmes d'écriture des langues slaves de l'Ouest (polonais, bas-sorabe, haut-sorabe, tchèque, slovaque)." [In:] Jean Breuillard, Roger Comtet (eds.) *Slavica occitania*. Vol. 12: *Alphabets slaves et interculturalité*; 283–310.

Pognan, Patrice (2007) "Forme et fonction en analyse automatique du tchèque. Calculabilité des langues slaves de l'Ouest." [In :] *BULAG*. Vol. 32: *Les langues slaves et le français: approches formelles dans les études contrastives*; 13–33.

Pognan, Patrice, Taïfi Miloud (2015) "Traitements automatiques en lexicographie de langues non dotées." [In:] Katarína Gajdošová, Adriána Žáková (eds.) *Natural Language Processing, Corpus Linguistics, Lexicography, Eighth International Conference.* Bratislava: RAM-Verlag.

Pognan, Patrice, Sadi Nabila, Fitas Rachida, Salhi Mohand-Akli & Achour Ramdane (2018) "Analyse morphologique automatique du kabyle. Traitement de l'affixation. Présentations linguistique et algorithmique." [In:] *10ème Bayreuth-Frankfurt-Leidener Kolloquium zur Berberologie*. Bayreuth: Rüdiger Köppe Verlag.

Schlamberger-Brezar, Mojca, Perko Gregor & Pognan Patrice (2015) *Les bases de la morphologie du slovène pour locuteurs francophones*. Ljubljana: Filosofska Fakulteta, Univerza v Ljubljani.

Taïfi, Miloud (2017) *Dictionnaire raisonné berbère - français. Parlers du Maroc* (1223 p.) et Pognan Patrice *Index Formes de mots berbères - racines* (166 p.) et *Index Significations françaises - racines berbères* (319 p.). Rabat : IRCAM.

Woliński, Marcin, Miłkowski Marcin, Ogrodniczuk Maciej, Przepiórkowski Adam, Szałkiewicz Łukasz & Szejko Jan (2011) *PoliMorf — otwarty słownik morfologiczny*. Warszawa: IPI PAN.

Woliński, Marcin, Miłkowski Marcin, Ogrodniczuk Maciej, Przepiórkowski Adam, Szałkiewicz Łukasz (2012) "PoliMorf: a (not so) new open morphological dictionary for Polish." [In:] *Proceedings of the Eighth International Conference on Language Resources and Evaluation (*LREC'12). Istanbul: European Language Resources Association (ELRA); 860–864.

Zemb, Jean-Marie (1996–1997) *Comment les mots éclairent les mots. A. approche graphématique. B. approche sémantique. C. approche métaphorique*. Paris: Annuaire du Collège de France, Chaire de grammaire et pensée allemandes.