Analyse contrastive de la valence de différents groupes sémantiques de verbes slovaques avec leurs correspondants français

D. Lemay et P. Pognan

Nous l'avons vu, les dictionnaires de valence ne sont pas très nombreux. Ils sont souvent le résultat du développement d'une théorie de la valence vérifiée ensuite par les faits au sein de projets pouvant durer sur des décennies.

Même si l'idée remonte aux années 1970, par exemple chez Maurice Gross, le portage de dictionnaires de valence existants sur des supports informatiques consultables ou téléchargeables à distance est un phénomène récent. On constate aisément qu'ils sont encore peu nombreux.

Simultanément l'intérêt pour le multilinguisme, porté autrefois en grande partie par les espoirs de traductions automatiques, a repris de l'ampleur depuis quelques décennies, souvent comme un enjeu primordial de l'Union européenne. Dans ce cadre les études sur l'intercompréhension, notamment des langues romanes, ont participé au regain d'intérêt pour le multilinguisme. Le besoin de faire face à ce défi pour des besoins de traductions humaines ou automatisées, de traduction vers des langues étrangères ou même de rédaction dans sa langue maternelle ou une langue étrangère bien dominée incite à aborder la valence aussi sous l'angle du multilinguisme. Cet aspect devient une préoccupation majeure de certaines équipes de recherche.

Au niveau le plus bas se situe la simple traduction du sens des entrées dans une ou plusieurs langues étrangères. C'est le cas de *Dicovalence*. L'origine se situe vraisemblablement dans le statut multilingue de la Belgique. Le néerlandais apparaît très tôt en traduction de l'entrée française et est rejoint par l'anglais depuis la version *Dicovalence 2* dont voici un exemple :

VAL$	se coaliser : P0 (PP<contre>)
VTYPE$	predicator simple pseudo_se
VERB$	COALISER / coaliser

...
TR_DU$ een verbond aangaan, een bondgenootschap sluiten,
z verbinden, z verenigen, samenspannen
TR_EN$ enter into / conclude a coalition (against)
...

Dès que l'on dépasse ce stade, les difficultés commencent et vont crescendo au fur et à mesure que l'on traite des langues typologiquement de plus en plus distantes les unes des autres. Nous distinguerons les études de la valence pour des langues du même groupe (ce seront des langues slaves) des études pour des langues de groupes [1] différents (une langue slave et le français ou l'anglais).

Le laboratoire ÚFAL (Ústav Formální a Aplikované Lingvistiky – Institut de linguistique formelle et appliquée), héritier de Petr Sgall et continuant à travailler dans le cadre de la Description Générative Fonctionnelle, s'est de tout temps préoccupé des questions de multilinguisme. Ce souci est omniprésent dans les travaux de traduction automatique et se répercute par conséquent sur l'étude de la valence, incontournable pour les niveaux syntaxico-sémantiques. Plusieurs systèmes de traduction automatique entre langues proches ont été élaborés, par exemple *Česílko* (tchèque – slovaque) ou *Ruslan* (tchèque – russe).

Avant de nous intéresser à l'étude de la valence entre le slovaque et le français, et, également entre le slovaque, le tchèque et le français, nous aborderons, d'une part, l'étude de la valence dans le cadre de deux langues du même groupe linguistique (le tchèque et le russe) et l'étude de la valence dans le cadre de langues appartenant à deux groupes différents (le tchèque et l'anglais), d'autre part.

1 La valence comparée russe – tchèque

Natalia Klyueva conduit depuis plusieurs années des études sur la valence en tchèque et en russe, notamment dans le cadre de systèmes de traduction automatique entre tchèque et russe, langues typologiquement proches. Dans le cadre d'une traduction automatique entre ces deux langues, Natalia Klyueva a recours à une analyse de surface [2] en suivant les formalismes du système Ruslan (Oliva, 1992). L'expérience a porté sur un ensemble d'environ 1850 verbes. 70 % des verbes possèdent la même valence. Si les différences sont réduites dans le choix des cas non prépositionnels, il n'en va pas de même dès que les complémentations comprennent des prépositions (c'est le cas de 5,6 % des verbes). Cela concerne des différences « systémiques » telles que « do(g,v(a)) », puisqu'en tchèque la préposition « v » (dans, à) avec mouvement (donc accusatif) a été remplacée par « do » avec le génitif.

Sinon les différences sont les conséquences d'une différence de cadre valenciel :

záležet (cz) | (n(n),na(l,ot(g))) | *зависеть* (ru) (dépendre de).

1. Nous parlons de « groupe », groupe slave, groupe germanique ou groupe roman, puisque nous nous situons toujours dans la famille des langues indo-européennes.

2. Klyueva N., Kuboň V. : « Verbal Valency in the MT Between Related Languages ». In : *Proceedings of Verb 2010, Interdisciplinary Workshop on Verbs*, Pisa, Italy, p. 160-164, 2010.

De manière plus récente [3], Natalia Klyueva a étudié les différences de valences sur la base de 1473 couples de verbes appartenant à 5 classes sémantiques du Vallex. Seulement 7,5 % (111 verbes) présentent une différence de cadre valenciel, suivant le tableau [fig1] :

Verb class	same frame	different frame	# of verbs
Change	309(95%)	14(5%)	323
Exchange	166(92%)	13(8%)	179
Motion	305(99%)	3(1%)	308
Communication	312(88%)	42(12%)	354
Mental Action	270(87%)	39(13%)	309
Total	1362(92%)	111(8%)	1473

FIGURE 1 – Différences dans le cadre valenciel entre tchèque et russe

L'observation d'un nombre plus grand de différences dans les cadres verbaux ayant des complémentations prépositionnelles demeure valable. Il faut noter que les différences systématiques sont nettement plus nombreuses que les différences accidentelles.

2 La valence comparée anglais – tchèque

Ce projet de dictionnaire de valences bilingue anglais – tchèque, dénommé CzEngVallex, a été présenté notamment dans le cadre de l'ACL [4] (Association for Computational Linguistics) et du PBML [5] (*Prague Bulletin of Mathematical Linguistics*). Ce qui est particulièrement intéressant, c'est la double utilisation prévue : recherche linguistique comparative autant que les travaux en traitement automatique.

Le projet se situe dans le cadre de la *Description générative fonctionnelle* et s'appuie sur un corpus arboré (« *treebank* ») annoté contenant valences et rôles sémantiques, le PCEDT (*Prague Czech-English Dependency Treebank* – Hajič *et al.*, 2011) et sur deux dictionnaires de valences, le PDT-Vallex pour le tchèque (environ 12 000 cadres valenciels pour environ 7 000 verbes) et le EngVallex pour l'anglais (environ 7 000 cadres valenciels pour un peu plus de 4 000 verbes) (Cinková, 2006). Ces trois ressources, bénéficiant d'une unicité théorique, la DGF, sont nécessaires à la construction du dictionnaire bilingue de valences, CzEngVallex, qui leur reste lié et permet ainsi d'assurer l'alignement des cadres verbaux tchèques et anglais.

Ce dictionnaire est en fait un fichier XML qui contient les cadres verbaux anglais où chaque sens anglais se voit attribuer les correspondants tchèques.

Un environnement particulier, basé sur l'éditeur Treex (dérivé de TrEd) qui permet d'afficher des structures multi-arborescentes, a été utilisé pour les travaux d'indexation de CzEngVallex à partir de données alignées.

3. Klyueva N. : *Comparing Czech and Russian Valency on the Material of Vallex*, *Proceedings of KONVENS 2012 (LexSem 2012 workshop)*, Vienna, 2012.

4. Urešová Z., Dušek O., Fučíková E., Hajič J., Jana Šindlerová J. : *Bilingual English-Czech Valency Lexicon Linked to a Parallel Corpus*, Proceedings of LAW IX – The 9th Linguistic Annotation Workshop, pages 124–128, Denver, Colorado, June 5, 2015. Association for Computational Linguistics.

5. Urešová Z., Fučíková E., Šindlerová J. : *CzEngVallex : a bilingual Czech-English valency lexicon*, *The Prague Bulletin of Mathematical Linguistics*.

Les difficultés rencontrées peuvent se situer à différents niveaux :

- Ce qui est un actant dans la structure anglaise est un circonstant dans la structure tchèque ou inversement, par exemple ce qui est le destinataire peut devenir une direction. Ces distorsions peuvent être typiques de certaines sous-classes sémantiques de verbes.

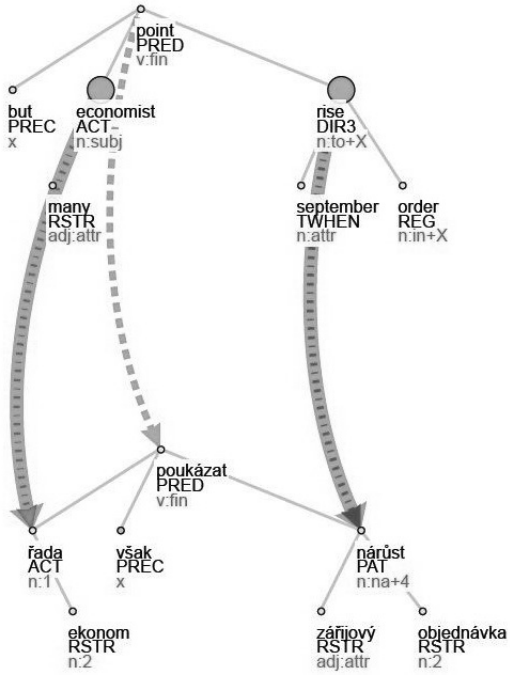


Figure 2 – Éditeur Treex – mise en correspondance des structures anglaise et tchèque

Dans l'exemple de la figure [fig2], les phrases concernées sont :

En : *But many economists pointed to a... September rise in orders...*

Cz : *Řada ekonomů však poukázala na... zářijový nárůst objednávek...*

On remarquera que dans cet exemple ce qui est noté « direction » dans la structure anglaise à cause du verbe « *to point to* » reçoit une valeur de « patient » dans la structure tchèque pourtant à partir du verbe correspondant, mot-à-mot « attirer l'attention sur ».

Des foncteurs différents peuvent être choisis pour une même structure sémantique en présence de structures morphosyntaxiques différentes. Mais d'une manière générale, cela illustre surtout la difficulté de traduire une expression linguistique en un schéma valenciel.

- L'absence de correspondant au foncteur d'une des structures dans l'autre structure, ce qui est nommé « alignement zéro ».

- L'usage de verbes réfléchis, notamment dans les traductions tchèques, déséquilibrent les structures.
- L'inconsistance de la notation valencielle au sein des données, ce qui peut être dû à une différence de granularité au sein des dictionnaires de valence. Les tournures idiomatiques peuvent également en être l'origine.
- La présence d'expressions numériques indéfinies comme dans l'exemple de la figure [fig2] (« *many economists* » [de nombreux économistes] face à « *řada ekonomů* » [une série d'économistes]) introduit des structures inverses : en anglais, « *economists* » est la tête alors qu'en tchèque, c'est « *řada* » (la série).

Les recherches entreprises dans le cadre de la construction d'un dictionnaire bilingue de valences permettent d'apprécier de manière scientifique les identités et les différences de structures valencielles en anglais et en tchèque, deux langues typologiquement différentes. La totale identité de structure existe, semble-t-il, dans environ 45 % des cadres valenciels. Souvent les cadres diffèrent par seulement un ou deux foncteurs.

La présence d'un tel dictionnaire est extrêmement utile pour la traduction, qu'elle soit humaine ou automatique. L'entreprise a été rendue possible par la présence simultanée et liée d'un corpus bilingue et d'un dictionnaire de valences pour chacune des langues en présence. Mais, en même temps, ces conditions favorables, où nous avons dans les faits deux langues sources, donnent vraisemblablement un contexte plus complexe à manipuler que la construction de structures valencielles d'une langue cible à partir d'une langue source.

3 La valence comparée slovaque – français

L'étude de la valence comparée slovaque – français se situe justement dans un contexte où le slovaque est la langue source et le français la langue cible à consigner en parallèle. L'ouvrage *Valenčné potencie slovies v kontraste*[6], publication liée à un contrat VEGA, pose de sérieuses bases à l'élaboration complète d'un dictionnaire de valences slovaque – français. Ces recherches récentes s'appuient naturellement sur une base de données d'où sont extraits, comme illustration des questions théoriques exposées en première partie, 36 verbes slovaques (avec leurs équivalents français) regroupés en 14 ensembles d'après l'aspect. On aura ainsi des regroupements où l'aspect porte sur une racine unique (le cas standard dans les langues slaves), par exemple : « *čakať* » / « *počkať* » / « *dočkať* » (attendre) ou sur des racines différentes (cas très rare) : « *povedať* » / « *hovoriť* » (avec le sens global de parler).

Les exemples sont extraits de grands corpora : le slovaque « *sme2011* » provient du quotidien SME pour l'ensemble de l'année 2011 transmis par l'intermédiaire du « *SNK* » (*Trésor national de la langue slovaque*) et le corpus français « *Le Monde 0.3* » comprend lui aussi des articles publiés sur l'ensemble de l'année 2011. Le premier contient plus de 400 000 phrases et le second plus de 800 000.

6. Zázrivcová M., Chovancová K., Veselá D., Ráčková L., Křečková V., Klimová K., Očenáš I. : *Valenčné potencie slovies v kontraste*, Belianum, Université Matej Bel, Banská Bystrica, 2016, p. 292.

Ces recherches sont réalisées dans le respect de la tradition slovaque et donc les principes et l'architecture de ce travail suivent la présentation afférente du chapitre 6 du présent volume. Les entrées slovaques sont conformes à celles du dictionnaire *Valenčný slovník slovenských slovies na korpusovom základe* (VSSSKZ) présenté dans le chapitre 6.

Nous examinerons, en l'explicitant d'après les références adéquates, le dictionnaire bilingue pour les verbes « *povedať* » / « *hovoriť* » (parler).

povedať (Pf), hovoriť (Ipf)
I. « *rečou oznamovať* » (annoncer par la parole)
VŠ$_{slo}$: S$_n$ – VF – (S$_d$) – o Sl/VV$_{že/aby/či/kor}$ – ADV$_{mod}$

À cette struture valencielle slovaque correspondent deux structures valencielles françaises :

parler **VŠ$_{fra}$: S$_S$ – VF – (à S$_{COI}$) – de S$_{COI}$/PS$_{de\ ce\ qui\ /\ de\ ce\ que}$** – ADV$_{loc}$ – ADV$_{mod}$
annoncer **VŠ$_{fra}$: S$_S$ – VF – (à S$_{COI}$) –S$_{COD}$/PS$_{que}$** – ADV$_{loc}$ – ADV$_{mod}$ – ADV$_{temp}$.

Rappelons ici les quatre types de transformations que l'on trouve dans le VSSSKZ :

1. passif réfléchi ;
2. passif participial ;
3. construction résultative ;
4. participe.

Les formes réflexives impersonnelles du slovaque (VF$_{refl,\ imp}$) et passives impersonnelles (VF$_{pas,\ imp}$) ne sont pas des formes réfléchies en français à cause de la présence du pronom « on » (PRONpers, ind) dans la position valencielle à gauche du verbe.

À la fin de l'exposé de chacune des significations du verbe sont énoncées les transformations possibles. Nous avons ici la possibilité d'abord de transformations de type 1, à savoir 1b (passif réfléchi impersonnel) et 1d (structure impersonnelle) :

le slovaque **VŠ$_{slo}$1b) : VF$_{refl,\ imp}$ – (S$_d$) – o Sl/VV$_{že/aby/či/kor}$**
p. ex. : Keď sa povie, že... (quand on dit, que...) donne en français :
VŠ$_{fra}$1b) : PRON$_{pers,\ ind}$ – VF – (à/avec S$_{COI}$) – de S$_{COI}$
VŠ$_{fra}$1b) : PRON$_{pers,\ ind}$ – VF – (à S$_{COI}$) – S$_{COD}$/PS$_{que}$
p. ex. : Quand on parle de ... ou p. ex. : On annonce que ...

Les structures des transformations (2b) (participe passif impersonnel) sont identiques en slovaque et en français, celles de (3a) (construction résultative non possessive) également. Le verbe de la construction française (3a) est « dire » :

VŠ$_{slo}$3a) : [VF$_{byť}$ – VF$_{part\ pas\ rez}$] – (S$_d$) – o Sl/VV$_{že/aby/či/kor}$
donne en français :

$VŠ_{fra}$3a) : [$VF_{être}$ – $VF_{part\ pas\ rez}$] – (à S_{COI}) – de/sur S_{COI}/PS_{que}
p. ex. : Beaucoup a été dit sur le concept…

On trouve ensuite un deuxième sens :

II. « *rečou vyjadriť* » (exprimer par la parole)
avec les verbes français correspondants : « dire », « énoncer », « exprimer », « formuler ».

Les significations qui ne sont pas communes aux deux éléments de la paire aspectuelle sont étudiées séparément. On trouve ainsi l'étude de *povedať* (Pf) seul avec les sens de (a) « dire / énoncer un jugement / accuser » ; (b) « fixer / préciser / établir / indiquer » ; (c) « dire / prononcer / déclamer ».

« *Hovoriť* » qui est l'imperfectif est vraisemblablement d'une fréquence élevée à en juger d'après le nombre de significations (6). La première a pour sens « parler / dialoguer / communiquer », la deuxième est exprimée en français par « parler et dire ». P. ex. : « Il parle avec modestie et conviction de ce projet. » Le troisième sens est traduit en français par les verbes « nommer / appeler », le quatrième est « parler (une langue) », le cinquième, rendu par « parler », a pour signification « prononcer un discours / prendre la parole en public ». Le sixième sens, particulier, réfère à l'aptitude d'utiliser la parole comme dans l'exemple : « Une seule chose est certaine, à Tripoli, les morts ne peuvent parler… »

En ce qui concerne les cinq premières significations, la seule transformation possible est la transformation 1b (passif réfléchi impersonnel). Comme précédemment, on observe un non parallélisme des structures slovaques et françaises du fait de la présence, en français, en position à gauche du verbe, du pronom personnel « on ». Ainsi, face à « hovorí sa » (avec la particule réfléchie « sa »), nous trouverons en français des exemples du type :

- « le média où l'on communique avec les autres »
- « on parle d'autant plus haut et fort que… »
- « ce que l'on nomme pudiquement le destin… »
- « on parle le catalan et le castillan… »
- « on a beaucoup parlé en 2010 de… »

4 La valence comparée slovaque – tchèque – français

C'est dans un tel contexte que nous avons conduit des tests en choisissant quelques verbes (19) appartenant à sept classes sémantiques différentes. Ce qui différencie ce travail-ci du précédent, c'est le choix d'une description suivant la tradition tchèque plutôt que slovaque dans l'optique de s'aligner sur Vallex. Ce serait le moyen de rendre Vallex multilingue, en premier lieu « multi-slave » ou à terme « panslave ».

Les classes et les verbes choisis sont présentés ci-dessous :

a. activité
robiť (faire, fabriquer), *tvoriť* (créer), *tvoriť sa* (se former)

b. parole
ďakovať (remercier), *hovoriť* (parler), *povedať* (dire), *povedať si* (se dire), *volať* (appeler)

c. émotion
cítiť ((res)sentir), *obdivovať* (admirer)

d. mouvement
chodiť (marcher), *doraziť* (arriver, atteindre)

e. sens
chutiť (avoir le goût de), *ochutnať* (goûter), *voňať* (émettre un parfum, sentir)

f. cognition
myslieť (penser), *vedieť* (savoir)

g. existence
existovať (exister), *žiť* (*vivre*)

Nous avons créé une base de données (Access) pour nous aider à définir un prototype. Cette base de données et les développements ultérieurs (plateforme en ligne Cetlef) sont présentés dans les deux chapitres suivants.

Notre étude s'appuie sur un dictionnaire général, le *Slovník súčasného slovenského jazyka* (*Dictionnaire du slovaque contemporain*) qui nous sert d'entrée notamment en ce qui concerne les différentes significations. La description s'appliquant au slovaque comme langue source, ce sont les éléments de cette description qui servent d'entrée pour le travail en français et le travail en tchèque. Nous sommes, en fait, dans une position double : d'une part, dans une situation où nous devons apparier des éléments existants pour les couples slovaque – tchèque et créer librement les éléments français pour le couple slovaque – français.

Signalons ici que nous avons délibérément adopté une description non pas tout à fait par pronominalisation comme le faisait Claire Blanche-Benvéniste, mais à l'aide des formes abrégées « qn » et « qch » qui nous semblent avoir des avantages sur des représentations de type « COD » et « COI ». En effet, on peut utiliser « qn » (COD) face à « à qn » (COI). L'opposition « qn » / « qch » permet d'introduire la notion d'animé et assimilé face à celle d'inanimé pour un foncteur donné.

Nous prendrons le verbe « *chodiť* » qui appartient à la classe « mouvement » comme illustration de notre démarche. Le *Slovník súčasného slovenského jazyka* (SSSJ ou S^3J) indique huit significations différentes pour ce verbe. Nous en trouvons neuf dans le Vallex 2.7 qui ne correspondront pas nécessairement à celles du slovaque, soit à cause d'une classification différente, soit à cause d'usages différents (en particulier, idiomatiques) du verbe concerné.

Nous présenterons ici en exemple deux significations :

a. SK : **chodiť**

> *(o ľuďoch, zvieratách) pravidelne, opakovane miernym pokrčením nohy v kolene zdvíhať a ukladať nohu pred seba, pričom sa nohy striedajú;*

takto sa opätovne pohybovať, presúvať sa určeným smerom ; byť schopný pohybu.
([à propos des gens, des animaux] avec un léger ploiement de la jambe au niveau du genou, régulier et répété, lever la jambe et la déplacer devant soi, chaque jambe à son tour ; se mouvoir ainsi, se déplacer dans une certaine direction ; avoir l'aptitude de la marche).
ACT$_1$$^{\text{obl}}$ MANN$^{\text{typ}}$ DIR$^{\text{typ}}$
Ex. : *chodiť naboso* (marcher pieds nus)

FR : **marcher** Classe : **motion**.

se déplacer par mouvements et appuis successifs des jambes et des pieds sans quitter le sol.
ACT$_{\text{qn}}$$^{\text{obl}}$ MANN$^{\text{typ}}$ DIR$^{\text{typ}}$
Ex. : marcher à quatre pattes (sk : *chodiť po štyroch*)

CZ : **chodit** : **chodiť-1-chodit-1**. Classe : **motion**.

pohybovat se pomocí nohou ; přemísťovat se (s nějakým záměrem).
(se mouvoir avec les jambes, se déplacer [avec une certaine intention]).
ACT$_1$$^{\text{obl}}$ **INTT**$^{\text{opt}}{}_{\text{k+3,na+4,inf}}$ MANN$^{\text{typ}}$ DIR$^{\text{typ}}$
Ex. : *chodit na borůvky* (aller à la cueillette des myrtilles).

b. SK : **chodiť**

byť v ľúbostnom vzťahu s niekým ; mať známosť (idiomatique).
(avoir une relation amoureuse avec qn ; avoir une connaissance).
ACT$_1$$^{\text{obl}}$ **PAT**$^{\text{obl}}{}_{\text{s+7}}$
Ex. : *chodiť s dievčaťom* (fréquenter une jeune fille).

FR : **fréquenter** Classe : **social interaction**.

fréquenter quelqu'un.
ACT$_{\text{qn}}$$^{\text{obl}}$ **PAT**$_{\text{qn}}$$^{\text{obl}}$
Ex. : Il fréquente les filles (*Chodí s dievčatami*).

CZ : **chodit** : **chodiť-3-chodit-7**. Classe : **social interaction**.

mít partnera (idiomatique).
(avoir un partenaire).
ACT$_1$$^{\text{obl}}$ **PAT**$^{\text{obl}}{}_{\text{s+7}}$
chodit s někým (fréquenter quelqu'un).

À la lumière de ce travail expérimental, il apparaît que les classes sémantiques choisies au départ doivent être plutôt considérées comme des hyper-classes, puisque nous trouverons dans une comparaison multilingue d'autres classes sémantiques telles que ci-dessus « social interaction » qui correspond dans le cas présent à un emploi idiomatique.

Nous sommes ici dans le cas de langues slaves très proches l'une de l'autre, mais lorsque la distance est plus grande, par exemple entre le tchèque et le slovène les différences de structures s'accroissent. La distance est perceptible au niveau d'une expression de nature idiomatique en ce qui concerne le français qui doit recourir à un verbe qui n'appartient pas à une classe sémantique de mouvement (fréquenter) alors que le verbe est le même en slovaque et en tchèque.

Il apparaît très vraisemblable d'avoir plusieurs cadres valenciels dans une langue cible pour répondre à un cadre valenciel source. Ainsi, si « *chodiť* » pouvait aussi signifier « flirter », nous aurions un deuxième cadre valenciel tel que :

flirter avec qn.
ACT$_{\text{qn}}^{\text{obl}}$ **PAT**$_{\text{avec qn}}^{\text{obl}}$

A posteriori, pour réaliser la composante française d'un dictionnaire (ou base de données) slovaque – tchèque – français, il nous apparaît nécessaire, comme cela a été fait ailleurs, de « compiler » les dictionnaires de valences existants tels que LVF+1, *Dicovalence 2* et Lexvalf et consulter Verbənet.