Base de données de la valence de verbes slovaques

P. Pognan et I. Šmilauer

Les travaux sur la valence des verbes slovaques ont été réalisés en parallèle avec un PHC franco-slovène Proteus [1] et en collaboration avec l'équipe du laboratoire ÚFAL (Ústav formální a aplikované lingvistiky) de l'Université de Prague. Cette équipe, l'équipe de Petr Sgall, est actuellement dirigée par Markéta Lopatková, l'un des auteurs du Vallex (*Valenční slovník českých sloves*, en français *Dictionnaire des valences des verbes tchèques* [2]).

Notre démarche est justifiée par :

- le souhait de lier au Vallex tchèque, suivant les mêmes principes, des « Vallex » d'autres langues slaves, ce qui donnerait un outil unique pour les langues slaves : outil soutenant la rédaction et la traduction pour chacune des langues contenues dans Vallex, mais aussi un outil de comparaison synchronique des langues slaves pouvant servir à la recherche et à la didactique ;
- le besoin de repenser et de refaire la composante « valence verbale » de bases de données généralistes préexistantes de langues slaves, dont, bien sûr, celles de slovaque et de slovène.

1 Définition d'un prototype (Patrice Pognan)

Les dépouillements effectués pour mettre en évidence les valences de certains verbes slovaques doivent être soigneusement conservés et structurés de manière telle qu'ils puissent être réutilisés pour l'enseignement du slovaque, pour la rédaction en langue slovaque que ce soit par des natifs ou par des étrangers ayant acquis un niveau suffisant de slovaque ou encore comme matériau permettant l'élaboration de stratégies et d'algorithmes d'analyse morphosyntaxique automatique du slovaque.

1. PHC Proteus 2012-2013, *Grammaire du slovène à l'usage des Francophones – Morphologie.*
2. M. Lopatková, Z. Žabokrtský, V. Kettnerová : *Valenční slovník českých sloves*, Univerzita Karlova v Praze, Karolinum, 2008.

Pour permettre cet ensemble de services, appuyés essentiellement sur un dictionnaire de valences, l'outil qui nous a semblé le plus adéquat est la base de données. En effet, un système de gestion de base de données (« SGBD ») a la souplesse nécessaire, puisqu'il permet de définir, changer, ajouter, retrancher des champs (contenant une information de base dont le type est défini), ajouter ou retrancher des tables (contenant des champs complémentaires constituant un ensemble de données ayant une finalité commune), changer les liens entre tables (ce qui détermine la manière dont les tables interagissent entre elles).

1.1 Présentation d'Access

Parce que le but premier était de fabriquer seulement un prototype de petite dimension et certainement aussi pour des raisons de facilité personnelle (nous connaissons bien le logiciel), mais aussi parce que ce produit est largement utilisé dans les milieux littéraires et qu'il permet de créer facilement des interfaces conviviales, notamment avec des systèmes d'onglets, nous avons choisi d'utiliser le SGBD Access de la suite Microsoft Office. Cela nous convenait d'autant mieux que nous avons dès le départ envisagé d'intégrer les acquis de ce prototype dans une base de données plus ancienne, elle aussi en Access, d'étude lexicale du slovaque et dont la destination est clairement les applications dictionnairiques.

Le logiciel Access offre la possibilité de créer seul ou à l'aide d'un assistant les tables où seront enregistrées les données, les formulaires qui sont les interfaces homme-machine, les états qui permettent de créer des sorties constituées d'éléments associés, même s'ils proviennent de tables différentes, en utilisant divers formats de fichiers, par exemple des fichiers Excel, des fichiers Word ainsi que des fichiers bruts, html ou xml. Un langage de programmation interne, VBA (Visual Basic pour Applications) complète cet ensemble et permet de programmer diverses actions non standard, par exemple positionnement sur certains enregistrements, navigation et recherche d'informations au sein de la base, ce qui peut également être fait, en partie, par des systèmes de requêtes.

Access a pour priorité de donner à des non-informaticiens la possibilité de construire des bases de données avec une structure pouvant aller jusqu'à une grande complexité et avec une interface homme-machine, relativement facile à élaborer (de manière presque intuitive) et elle aussi complexe dans la mesure où les données peuvent être présentées dans la conjonction de deux axes, un axe horizontal, représenté par des onglets thématiquement dédiés qui peuvent exercer des partitions sur les tables, et un axe vertical, représenté par la possibilité de longs ascenseurs. Si l'on rajoute à cela, la possibilité d'implanter dans l'un et l'autre axe autant de sous-formulaires, imbriqués (c'est-à-dire en « poupées gigognes » ou « poupées russes ») ou non, que l'on veut, on peut percevoir la complexité de telles interfaces. D'ailleurs, les programmes de transformations automatiques d'un type de SGBD à un autre (par exemple vers MySQL ou PostGreSQL), s'ils réussissent à faire correctement le transfert des tables et de leur intitulé, échouent à reconstruire une interface correspondant à celle d'Access.

En fait, les problèmes avec Access commencent dès que l'informatique est requise. En premier lieu, parce que la programmation en VBA pour Access, contrairement à la

programmation en VBA pour Excel, est peu explicitée par des ouvrages spécialisés. De plus, si la base de données construite suit réellement la norme Unicode UTF-8, il n'en est pas de même de l'interface de programmation dénommé « VBE » (éditeur du Visual Basic) qui n'est pas à la norme Unicode et « casse » réellement le code lorsque l'on a besoin de programmer le traitement d'une langue non codée en latin-1 (la norme des langues « occidentales ») ou, encore pire, si l'on prend en compte des situations multilingues. Cette situation désavantageuse ne peut être compensée qu'au prix d'acrobaties informatiques d'une inutile complexité.

Les systèmes d'entrée-sortie, c'est-à-dire d'import-export des données, assurent l'essentiel, mais ils doivent être bien connus de ceux qui y ont recours. Encore faut-il relativiser les choses, puisqu'un mode essentiel d'import se fait, et vraisemblablement depuis bien avant qu'il ne soit ainsi dénommé, à l'aide de fichiers CSV utilisés abondamment par ceux qui travaillent sur des textes bruts (c'est-à-dire au format du bloc-notes que vous trouvez sur votre ordinateur). Bien entendu, grâce à VBA, il est possible de programmer aussi des entrées-sorties.

1.2 Présentation du prototype

Parmi les diverses formes du Vallex, nous avons retenu comme référence l'organisation graphique de l'interface HTML des pages Web. Une entrée de Vallex s'y présente sous la forme suivante :

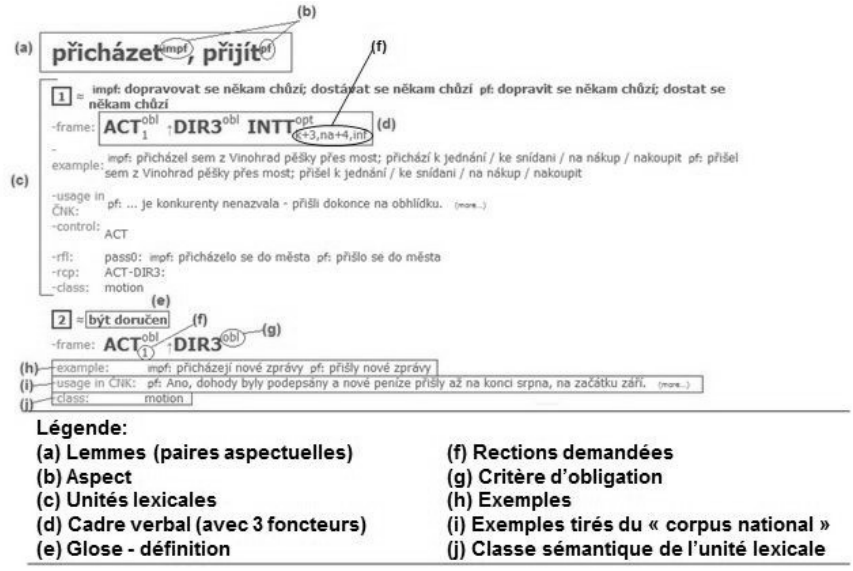


Figure 1 – Une entrée de Vallex sur le Web (forme en HTML) [3]

3. Extrait de : Grégoire Labbé (2013), *Traitement de la valence de verbes de mouvement slovènes sur la base de la valence de verbes tchèques*, mémoire de master 1, Inalco – Paris et Karlova Univerzita – Prague.

Comme l'indique le renvoi (b), les lemmes (a) constituant l'entrée sont toujours accompagnés de leur aspect (perfectif, imperfectif, itératif ou bi-aspectuel). Chacune des significations accompagnée de son cadre verbal et de tout ce qui la détermine est dénommée par nos collègues tchèques « unité lexicale » (c).

Une telle unité lexicale contient une « glose » (e) qui est la signification pour laquelle on a créé un cadre verbal (d) avec un certain nombre de « foncteurs » (3 pour la première glose, 2 pour la seconde). Chacun de ces foncteurs est accompagné en exposant d'un critère d'obligation (g) qui a pour valeur « obligatoire », « optionnel » ou « typique » (pour les circonstants) et en indice de la rection requise (f) qui peut être un cas simple, une préposition suivie du cas régi, d'un infinitif, d'une proposition subordonnée.

L'unité lexicale contient également des exemples (h), des exemples extraits du corpus national (i), l'indication des valeurs réfléchies et réciproques et l'indication de la classe sémantique éventuelle d'appartenance.

La structure prévue dans nos bases de données devra répondre à celle d'une entrée de Vallex.

La table principale nommée « verbe » constitue l'entrée de la base de données et présente les verbes et leur aspect. Elle est liée à une table secondaire intitulée « cadre verbal » qui assure la description et la gestion des cadres verbaux. Cette table fait appel à une sous-table qui gère les exemples afférents. Deux autres tables, dont la définition du contenu est identique à celle du slovaque et possédant chacune une sous-table chargée des exemples, l'une pour le tchèque, l'autre pour le français, sont liées à la table « cadre verbal » et assurent la correspondance entre les trois langues.

On trouvera ci-dessous un extrait d'écran présentant la structure de la base de données [fig. 2].

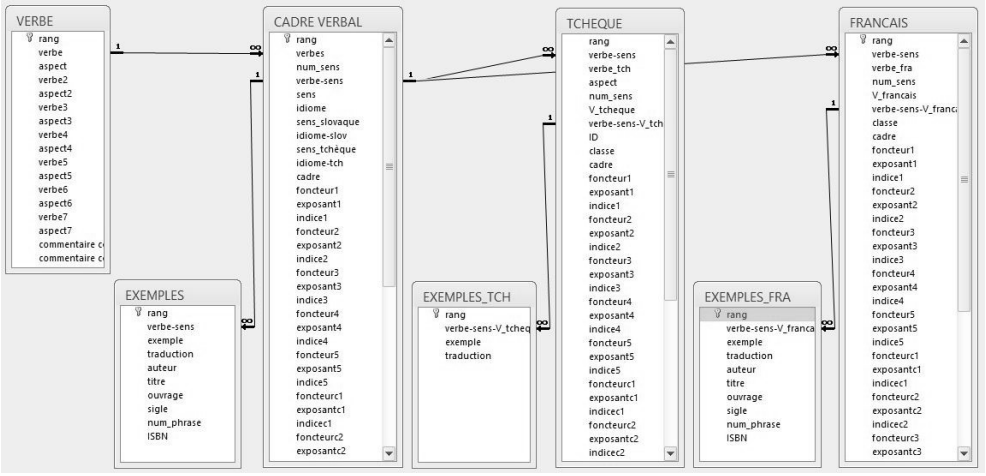


Figure 2 – Structure de la base de données de la valence verbale

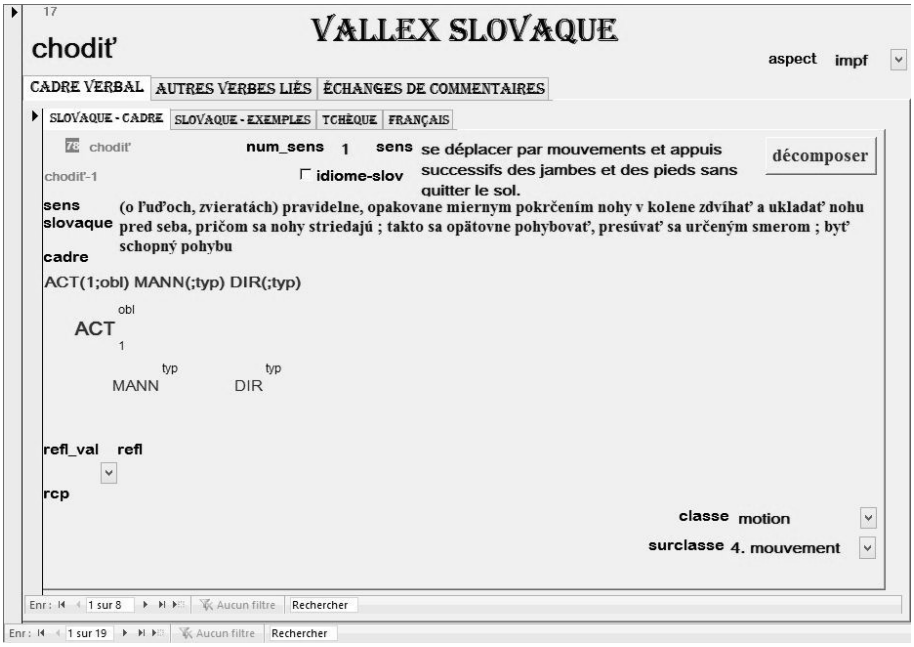


Figure 3 – Cadre verbal slovaque – entrée dans la base

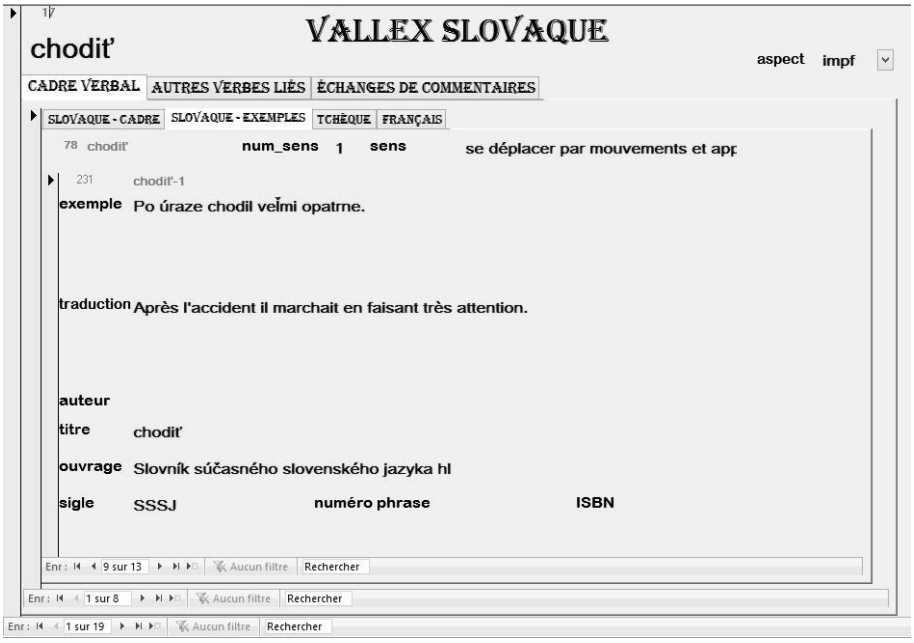


Figure 4 – L'onglet des exemples slovaques

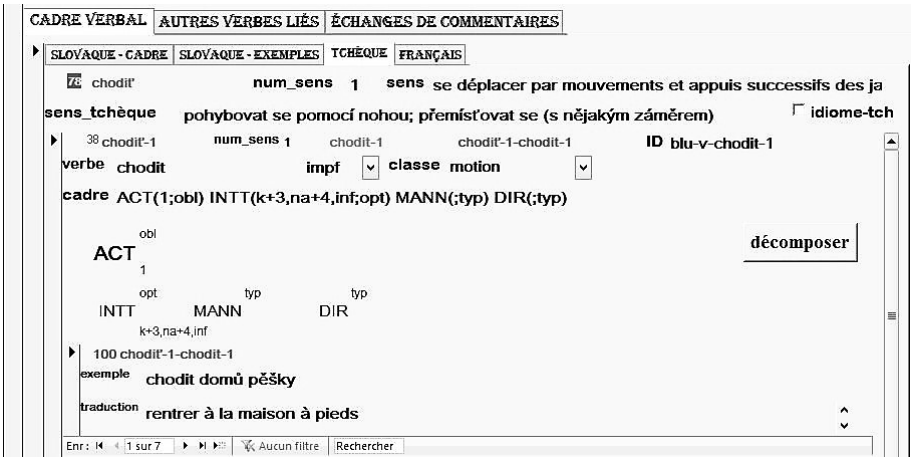


Figure 5 – L'onglet tchèque

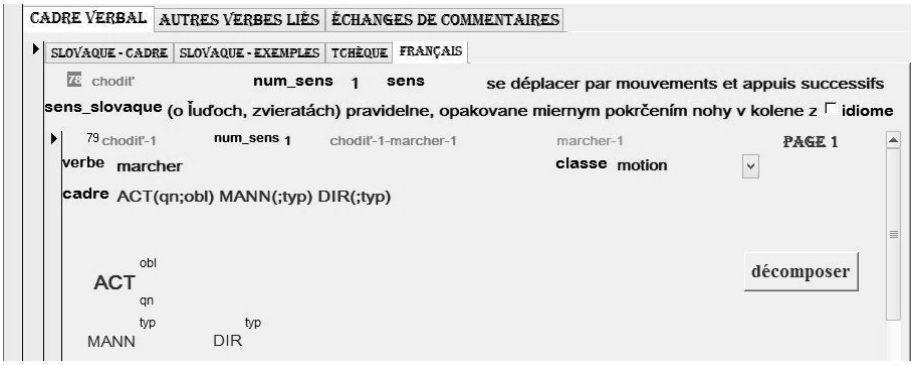


Figure 6 – L'onglet français

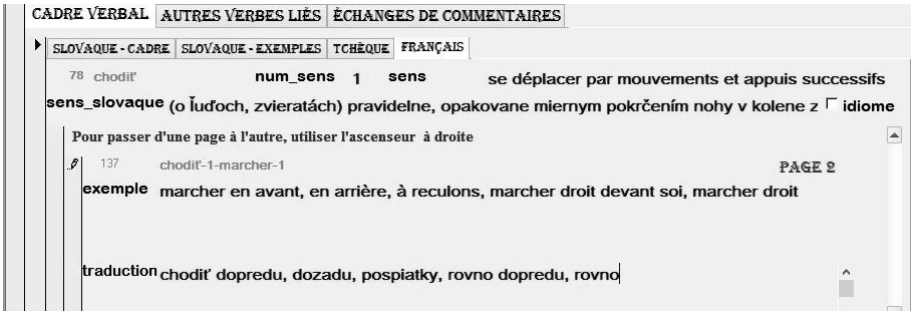


Figure 7 – Des exemples dans l'onglet français

L'entrée dans la base de données s'ouvre sur l'onglet « cadre verbal » qui lui-même donne accès au cadre verbal slovaque, puis aux exemples slovaques ainsi qu'aux parties tchèque et française. On le voit sur cet exemple, les onglets offrent une entrée structurée dans la base de données.

Les exemples slovaques sont regroupés avec les indications afférentes dans un onglet spécialisé [fig. 4].

Toujours pour le verbe slovaque « *chodiť* », il est possible de consulter l'onglet tchèque qui reproduit les informations extraites du Vallex tchèque pour le verbe tchèque « *chodit* ». On y trouve en premier lieu avec le verbe, l'indication de l'aspect et l'appartenance à une classe sémantique, une définition à peu près équivalente à celle du slovaque. Vient ensuite le cadre verbal présenté dans un mode analogue à la forme HTML du Vallex en ligne, suivi d'une éventuelle série d'exemples.

Enfin, nous pouvons également consulter l'onglet concernant le français qui offre comme le slovaque et le tchèque [fig. 5] un verbe avec uniquement un prime-actant [fig. 6].

Cet onglet est muni d'un ascenseur qui cache en sa partie basse l'ensemble des exemples [fig. 7].

La partie qui concerne le cadre verbal du slovaque est intéressante pour décrire le fonctionnement du verbe au sein d'une base de données à finalité lexicographique ayant notamment des applications dans le domaine de la réalisation de dictionnaires et de lexiques bilingues. C'est pourquoi nous avons implanté cette partie du prototype dans une base de données générale du slovaque.

2 Insertion du cadre verbal dans une base de données générale conçue par Patrice Pognan

Nous avons, dans les années 2008-2009, développé un prototype de base de données, outil multifonctionnel et multilingue, sorte de matrice générale que l'on peut ensuite dupliquer pour l'étude lexicale et grammaticale de langues particulières au prix d'aménagements minimes. Cet outil rassemble des connaissances lexicales pour la réalisation de dictionnaires, des connaissances grammaticales synchroniques et diachroniques situées à différents niveaux (phonétique, phonologie, morphologie, syntaxe, sémantique) et, dans la mesure du possible, des connaissances spécialisées (scientifiques, médicales, techniques).

Ce modèle, utilisé pour plusieurs langues, est modifié (et espérons-le, amélioré) régulièrement. C'est dans cette perspective que nous avons intégré dans ce modèle les acquis de notre programme actuel sur la description et la consignation de la valence verbale.

Nous reprendrons ici, bien évidemment, la base de données du slovaque en nous concentrant sur les aspects liés à la valence.

La structure de la base de données lexicales montre l'emplacement de la table documentant la valence reliée à l'enregistrement des différentes significations d'un verbe [fig. 8].

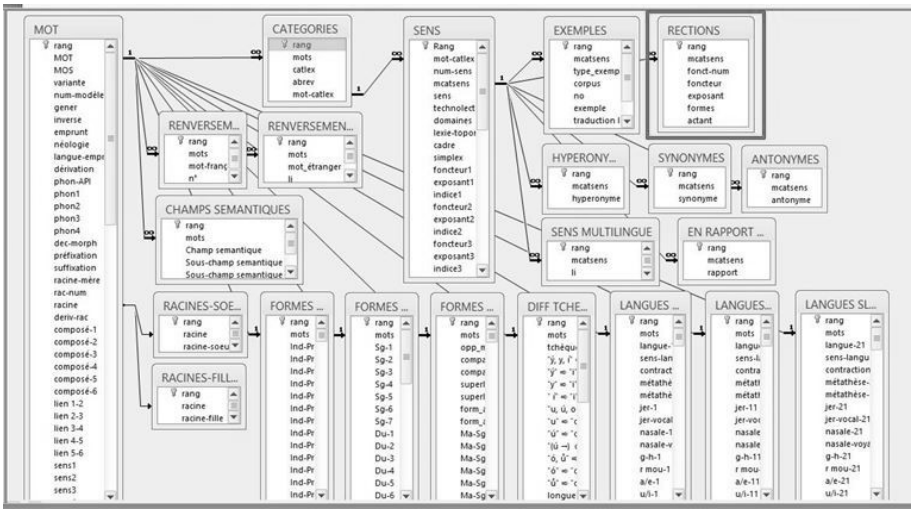


Figure 8 – La structure de la base de données lexicale

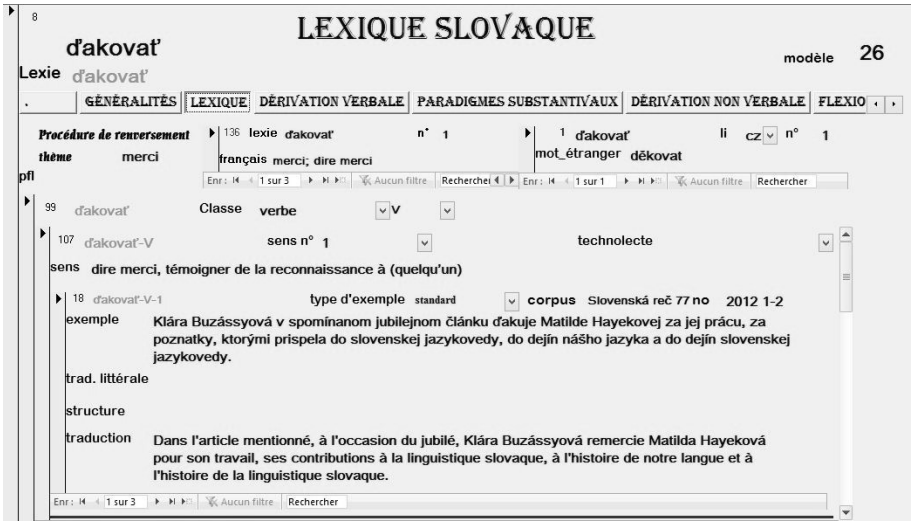


Figure 9 – Des exemples illustrant le sens n° 1 du verbe « *ďakovať* » (remercier)

La table concernant la valence est une table-fille de la table consacrée au sens et elle apparaît dans l'interface en tant que sous-formulaire du formulaire « sens » dans un ascenseur en compagnie de la sous-table « exemples », elle-aussi liée à la table « sens ».

Voici des exemples pour le verbe « *ďakovať* » (remercier) [fig. 9].

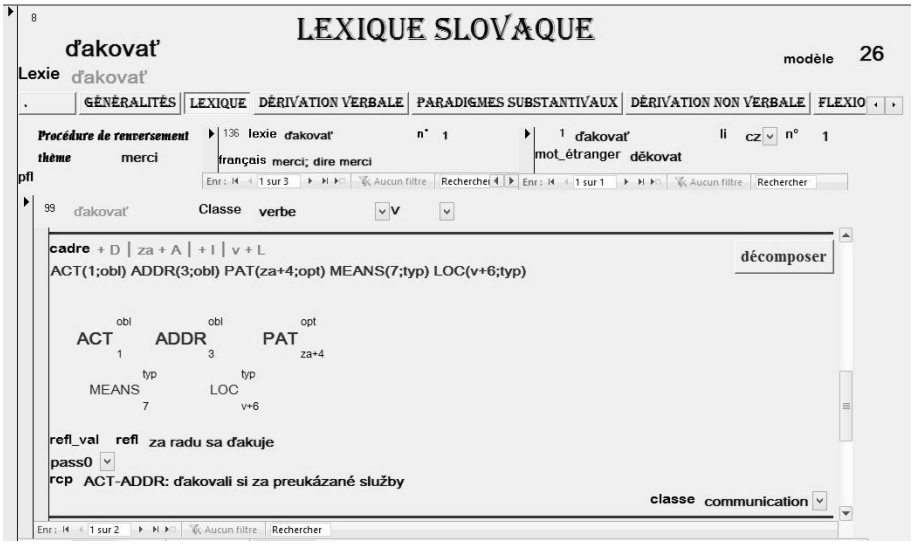


Figure 10 – Le sous-formulaire « cadre verbal » lié au sens n° 1 du verbe « ďakovať » (remercier)

Puis, plus bas dans l'ascenseur, le sous-formulaire du cadre verbal afférent au sens n° 1 du verbe [fig. 10].

3 Un « Vallex » slovaque sur la plateforme Cetlef conçue par Ivan Šmilauer

L'analyse contrastive de la valence de différents groupes sémantiques de verbes slovaques avec leurs correspondants français a abouti à la conception d'une base de données relationnelle, réalisée à l'aide du logiciel MS Access. Ce prototype contient 19 verbes slovaques et leurs équivalents en français et en tchèque.

Une base de données en Access est un fichier créé sur un ordinateur personnel et dont la caractéristique essentielle est d'être « mono-utilisateur ». Or, nous avons eu besoin d'un accès partagé par tous les membres du projet, ce qui nous a amené à la monter sur un serveur, en l'occurrence sur notre plateforme Cetlef.

Pour cette raison, les données structurées dans cette base ont été exportées en XML pour rendre possible un affichage sur une page Web. Celle-ci permet une présentation publique des données dans un but de recherche ou dans une visée pédagogique.

L'affichage de ces données sur la plate-forme CETLEF (voir le chapitre 8) est assuré par une génération dynamique des pages HTML à l'aide du langage PHP à partir du fichier source (le contenu structuré en XML de la base Access). Reprendre le schéma XML du Vallex pragois et concevoir un outil spécifique de saisie dans une base de données de type MySQL aurait été une procédure plus directe et plus évidente.

Pour faciliter la lecture des entrées du dictionnaire, la représentation graphique est inspirée de l'interface Web développée pour le Vallex tchèque [4].

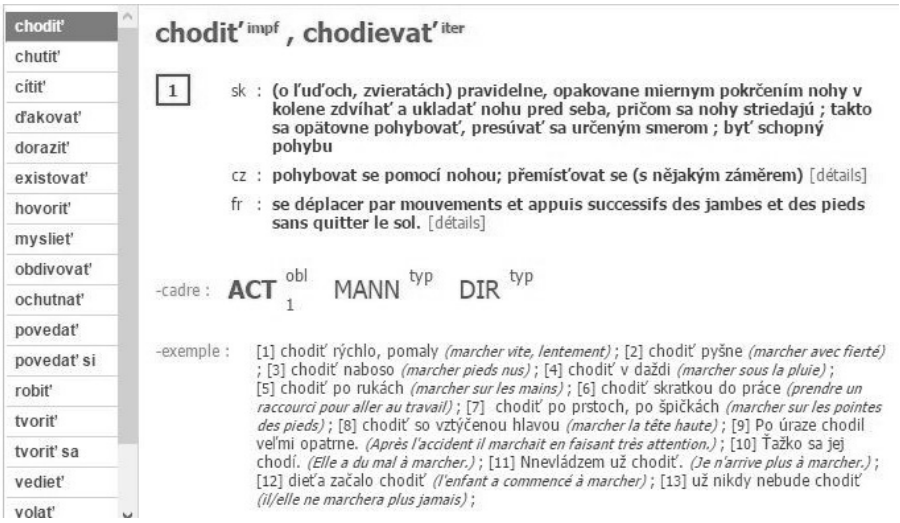


Figure 11 – Affichage d'un Vallex slovaque sur la plate-forme CETLEF

L'interface se présente avec une colonne à gauche qui contient la liste des verbes disponibles et une fenêtre centrale qui présente à son tour la description du verbe choisi. Chaque lexie d'un verbe polysémique est annotée par un nombre entier en commençant par '1'.

Les informations caractérisant chaque verbe sont les suivantes :

- l'entête de l'article avec l'infinitif du verbe et son opposition aspectuelle, voire la forme itérative ;
- la définition de chaque lexie du verbe, formulée en langage naturel et fournie en slovaque, en tchèque et en français ;
- le cadre verbal en slovaque pour la lexie en cours contenant les actants et les modifications libres obligatoires, optionnelles ou typiques, accompagnées des étiquettes portant sur la nature de l'élément et des morphèmes (cas, prépositions) à l'aide desquels cet élément est exprimé sur le niveau analytique (syntaxe de « surface » en PDT) ;
- enfin, les exemples en slovaque avec leur traduction en français sont donnés pour illustrer l'usage de la lexie courante de chaque verbe.

Les définitions en français et en tchèque (voir le 2e alinéa « • » ci-dessus) renvoient vers les informations équivalentes du slovaque ce qui permet une comparaison rapide des cadres verbaux des trois langues.

4. Voir http ://ufal.mff.cuni.cz/legacy/vallex/2.6.3/data/html/generated/alphabet/index.html.

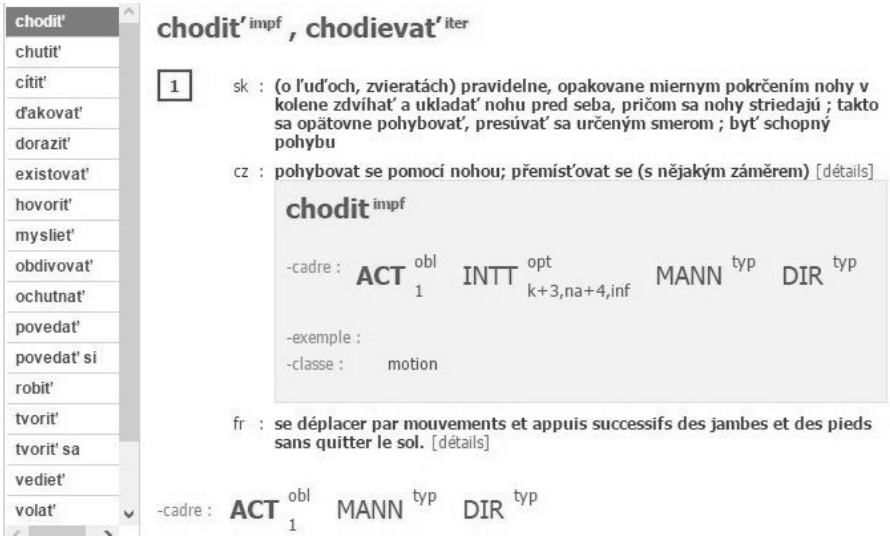


Figure 12 – Affichage du cadre verbal pour l'équivalent tchèque

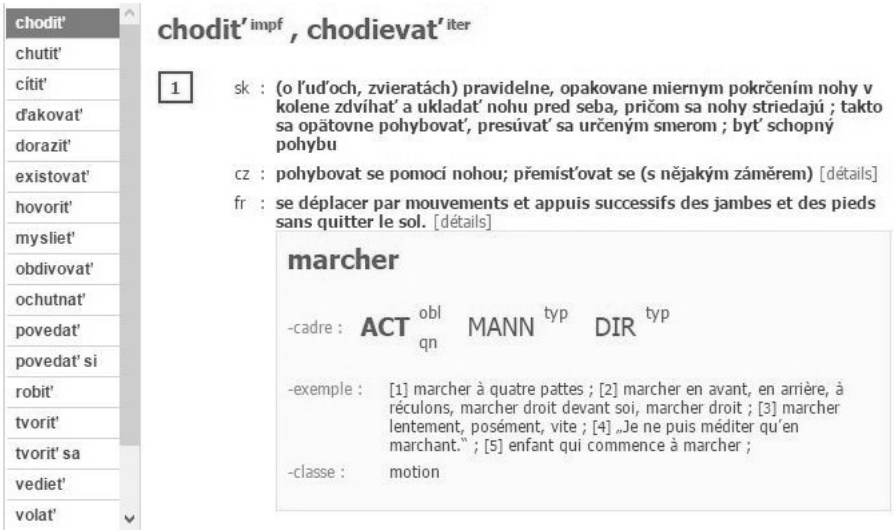


Figure 13 – Affichage du cadre verbal pour l'équivalent français

La solution actuelle est adaptée à un nombre limité de verbes et sert en tant que démonstration des possibilités. Si le nombre de verbes était amené à s'agrandir, il serait nécessaire de retravailler l'interface pour permettre une sélection plus confortable.