PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 89 JUNE 2008 97–106

De la théorie à l'application : VALLEX, une démarche exemplaire

Patrice Pognan

Abstract

VALLEX est le fruit du temps : le temps de réfléchir, le temps de tester, le temps de faire, le temps d'utiliser. VALLEX est le contre-exemple prototypique de tout ce que souhaitent les politiques actuelles de la recherche : c'est pour les chercheurs sérieux le réconfort d'apprécier la richesse qu'apportent la pérennité d'une équipe et de ses thèmes de recherche, l'effet cumulatif des connaissances d'une génération de chercheurs à l'autre. L'histoire de VALLEX prend ses racines dans les années soixante et ne peut pas être dissociée de l'histoire de Petr Sgall et de ses disciples qui pour vivre l'aventure de la recherche ont dû d'abord lutter pour la survie de leur équipe, de ses idées, de ses programmes.

1. La théorie

Nous avons jugé inutile une énième présentation de la théorie, la Description Générative Fonctionnelle (DGF) et préféré en commenter les aspects qui nous semblent primordiaux. Le lecteur trouvera des descriptions précises en particulier dans [Lopatková, 2003, PBML 79-80] et dans [Žabokrtský, Lopatková, 2007, PBML 87].

Nous donnons en annexe une bibliographie conçue de manière particulière : nous avons ordonné dans le temps quelques publications qui nous semblent importantes de l'ensemble de l'équipe. Le résultat est frappant sous plusieurs aspects.

La première remarque est claire : les années soixante sont la période de genèse de la théorie, la DGF, réalisée par Petr Sgall. Les années 70, 80 et 90 (trente années de travail !) sont globalement les années de développement de la théorie avec l'élaboration constante d'outils de test de cette théorie par P. Sgall et ses disciples – collaborateurs Eva Hajičová et Jarmila Panevová. (Nous en donnerons plus bas une interprétation plus fine). Enfin, les années 2000 voient apparaître sur le devant de la scène tout un ensemble de jeunes chercheurs « seconde génération » de la DGF tournés vers les applications informatiques, en particulier dans le cadre des travaux autour du Prague Dependency Treebank (PDT) et vers la réalisation concrète d'un dictionnaire

Please cite this article as: Patrice Pognan, De la théorie à l'application : VALLEX, une démarche exemplaire. The Prague Bulletin of Mathematical Linguistics No. 89, 2008, 97–106.

de valences, Vallex, ce qui marque la matérialisation de la théorie en une suite d'applications.

Il convient de souligner que la richesse d'applications bien fondées scientifiquement n'apparaît de manière évidente que dans la cinquième décennie après le début des recherches, que c'est une nouvelle génération de chercheurs qui, avec l'appui constant des chercheurs de la première génération, crée et affine les produits dont la validité est issue de la théorie. Ceci devrait être un guide de réflexion pour les « décideurs » … Le fait d'arriver vers les numéros 90 d'une revue bi-annuelle (le PBML) entièrement créée, nourrie et gérée par une équipe laisse également rêveur …

Dans un deuxième temps, nous allons considérer les développements « forts » de la DGF des années 70, 80 et 90.

– Les années 70 sont celles du renforcement de la DGF. Elles sont marquées principalement par l'analyse de la partition thème / rhème [Sgall, Benešová, 1973], [Sgall, Hajičová, 1977, 1978], [Sgall, 1979] et les études sur le cadre verbal [Panevová, 1974, 1975, 1977], [Panevová, Sgall, 1976].

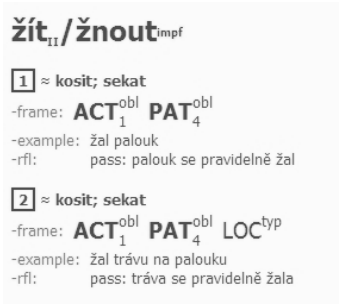
– Les années 80 sont celles de la maturité de la théorie et l'époque d'un faire-savoir important [Sgall, 1980, 1984], [Sgall, Hajičová, Panevová, 1986]. Ce sont également les années de recherche déterminante d'une part, vers la syntaxe profonde, le niveau tectogrammatical [Hajičová, Panevová, 1984] qui permet de formuler une interprétation sémantique de la phrase et du texte et d'autre part, pour la patiente mise en exergue de ce que nous considérons comme le maillon fondamental pour l'automatisation et les applications de type Vallex, l'ordre systémique sans lequel rien ne serait possible [Hajičová, Sgall, 1986].

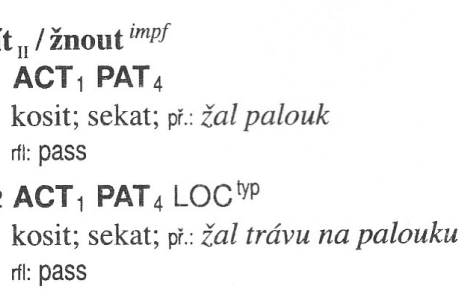
– Les années 90 voient l'apparition de concepts avancés tels que celui de contrôle [Panevová, 1996] et le renforcement des applications de grande envergure. Notons que l'équipe est connue sur toute son histoire pour ses applications dans les domaines de la traduction automatique, de l'indexation et de la recherche d'information.

Mais c'est certainement la prise en compte de l'ensemble de quarante ans de travaux (de 1960 à 2000) qui fait de la Description Générative Fonctionnelle la théorie (et la pratique !) capable de pleinement transformer les travaux de Tesnière en un système de calcul de la langue.

2. L'application Vallex

VALLEX existe en trois versions : une version HTML consultable en ligne, une version XML permettant l'utilisation du dictionnaire par programmation et une version papier qui vient d'être publiée [Lopatková, … 2008]. Cette version contient le dictionnaire de valences (environ 350 pages) dont l'organisation graphique s'inspire heureusement de l'interface HTML. Le dictionnaire est précédé d'une introduction détaillée de 20 pages et d'une bibliographie abondante.





Nous avons, à gauche, la forme issue de la consultation HTML et à droite, celle adoptée dans le dictionnaire. On y observe deux simplifications : les exemples ne sont pas donnés pour les formes réflexives et réciproques (un Tchèque reconstruit facilement ce type de construction, mais les exemples peuvent être utiles à un lecteur étranger) et les foncteurs représentant les participants internes sont considérés par défaut obligatoires. Ils ne portent un exposant que s'ils sont facultatifs (exposant « opt ») :

splácet *impf*, **splatit** *pf*
1 **ACT**$_1$ **ADDR**$_3^{opt}$ **PAT**$_4$ RCMP$_{za+4}^{typ}$ MANN^{typ}

Ce dictionnaire a été pensé comme outil pour le public tchèque. En témoigne l'introduction rédigée en tchèque. Il nous semble cependant regrettable que l'on n'ait pas pris en considération l'usage qu'un étranger, même sans connaissance du tchèque, peut en faire ne serait-ce qu'à titre d'exemple pour des travaux sur d'autres langues. Doubler l'introduction tchèque par une introduction dans une ou plusieurs langues internationales (au moins en anglais, mais aussi peut-être en français, espagnol, allemand) aurait été bienvenu. Cela paraît d'autant plus surprenant (et même contradictoire) que le rapport technique interne au laboratoire possède une très bonne introduction en anglais [Lopatková, … 2006] et qu'un article en anglais a été publié en juin 2007 dans le Prague Bulletin of Mathematical Linguistics [Žabokrtský, … 2007]. Le travail était quasiment fait ! A défaut, le lecteur étranger devra donc se munir du numéro 87 de cette revue.

Etant donnée l'existence de cet article, nous ne reprendrons pas, dans cette même revue, la description détaillée de Vallex. Nous nous contenterons d'insister sur quelques points qui nous semblent importants.

Il est intéressant que les auteurs aient suivi la Description Générative Fonctionnelle de P. Sgall dans la constitution d'entrées possédant simultanément tous les lemmes aspectuels, ce qui a pour mérite de montrer que la bipolarité aspectuelle tant prônée peut s'étendre de manière très fréquente à une triade due à l'itératif sachant que l'on peut être en présence d'un nombre de lemmes beaucoup plus élevé : jusqu'à 6 !

Dans une approche linguistique centrée sur le verbe, il convient, en premier lieu, d'insister sur la nécessité de posséder des informations précises sur le cadre verbal. L'information sur les groupes compléments du verbe (ce que les auteurs appellent en anglais à juste titre « comple-

mentation » étant donné qu'il peut s'agir de réalisations sous forme de syntagmes nominaux, de locutions adverbiales ou même de propositions subordonnées) représente une description syntaxico-sémantique précise signification par signification du verbe (ses différents sens). Nous pensons que le terme d'unité lexicale utilisé par les auteurs tant en anglais qu'en tchèque pour nommer le cadre verbal de chacune des significations du verbe n'est pas réellement approprié. C'est cette information qui fait la validité d'un tel dictionnaire pour la rédaction en tchèque et la traduction du ou vers le tchèque. Pour chacun des sens, ces cadres verbaux sont divisés en participants internes (actants), en « quasi-actants » (la différence, l'intention et l'obstacle) et en participants externes, libres (complémentation libre - circonstants). A l'intérieur du cadre verbal chaque foncteur peut être obligatoire ou facultatif et accompagné en indice de ses rections syntaxiques.

Les rections syntaxiques des foncteurs peuvent être gérées par classes d'équivalence notées par un symbole du type « AIM » (but) regroupant un ensemble de valeurs telles que « aby » (afin, pour), « ať » (que), « do+2 » (dans, à + génitif), …, « v zájmu+2 » (dans l'intérêt de + génitif), …c'est-à-dire des connecteurs syntaxiques, des prépositions simples ou dérivées, … Ce type de démarche reflète l'implémentation qui peut être faite pour une analyse automatique du tchèque.

Dans le même genre d'idée, certains foncteurs de temps ou de lieu pouvant alterner sont représentés par un foncteur prototypique (au nombre de 5), ce qui offre la souplesse nécessaire à une bonne analyse automatique.

Enfin, l'affectation à environ 45(cadre pour chacun des sens d'un verbe) d'une catégorie sémantique générale (il en existe pour le moment 22), par exemple « transport », « mouvement », « phase d'une action », … rapproche de travaux de nature sémantico-cognitive. Cette direction, pour le moment exploratoire, devrait être sérieusement étudiée et affinée.

Nous nous permettrons de souligner que l'usage du dictionnaire ne dispense pas de la consultation de Vallex sous sa forme HTML qui reste nécessaire grâce à la souplesse et la multiplicité des accès que donne l'informatique. Nous pouvons, en effet, y trouver un accès par ordre alphabétique des entrées verbales comme dans le dictionnaire, mais aussi en plus un accès par ensembles de configurations aspectuelles, par nombre de sens de chacun des verbes, par foncteur, par rection syntaxique, par classe sémantique, par type de contrôle, un accès pour les homographes, pour les formes réfléchies, pour les formes réciproques, …

3. Développements ultérieurs potentiels

3.1. Développements souhaitables dans le cadre de ÚFAL

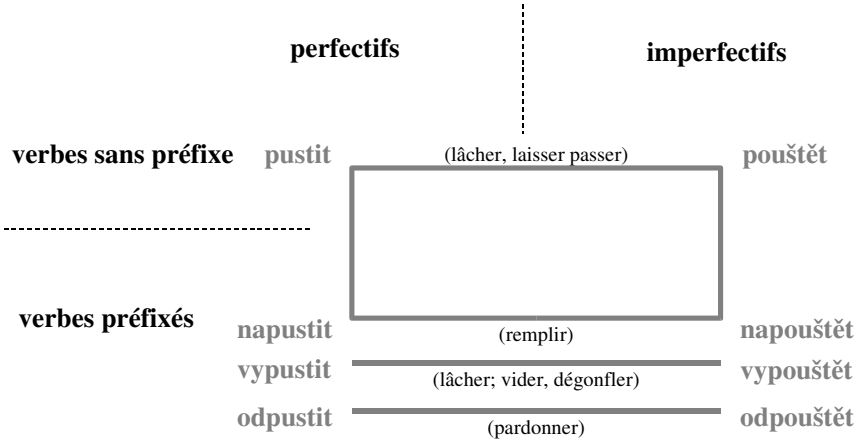
3.1.1. Outil tout à fait remarquable quelle que soit la forme considérée, Vallex requiert à notre avis encore au moins quelques développements.

Actuellement, les entrées verbales ne sont constituées que des ensembles aspectuels présentant le même radical, c'est-à-dire pour ce qui nous intéresse la même combinaison préfixe(s)-racine. Il convient de savoir qu'en terme de formation morphologique de l'aspect, on peut distinguer quatre groupes, de volumes très inégaux :

– les paires aspectuelles (2 ou 3) formées sur des verbes différents, par exemple brát, vzít (prendre) - traitées dans Vallex.

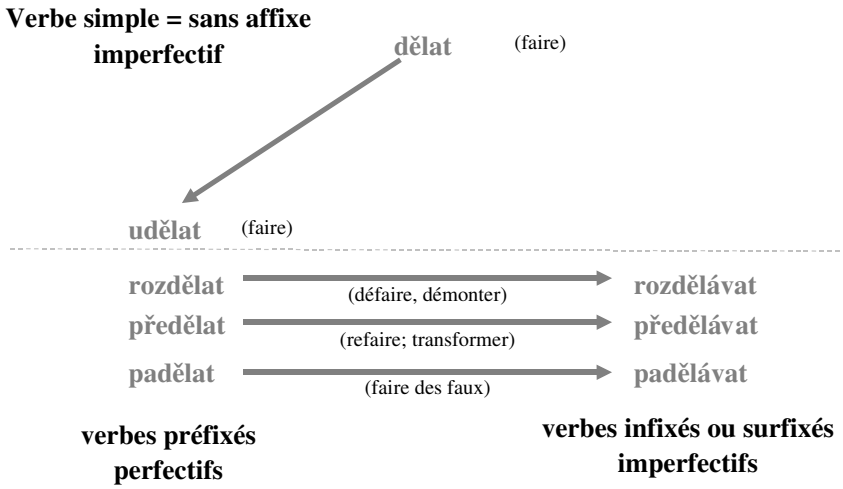
– les verbes bi-aspectuels, généralement des emprunts à des verbes étrangers qui sont regroupés dans une (sous-)classe en -ovat – catégorie traitée dans Vallex.

– la formation aspectuelle « en carré » pour les verbes perfectifs simples (c'est-à-dire non préfixés), par exemple :



Cette catégorie est traitée dans Vallex, au prix d'un renvoi d'une entrée perfective « koupit », « pustit » vers une entrée commune classée alphabétiquement suivant la forme imperfective : « kupovat, koupit » (acheter), « pouštět, pustit » (lâcher).

– la formation aspectuelle « en triangle », très majoritaire (vraisemblablement au moins 90 % des verbes sont concernés) :



Cette construction part de la forme simple imperfective, c'est-à-dire sans préfixe ni suffixe en dehors du morphème d'infinitif. Le verbe simple, imperfectif, forme son perfectif correspondant à l'aide d'un préfixe dit « zéro » parce que vide de sémantique. L'un des problèmes délicats (par exemple pour l'apprentissage de la langue) est que chaque verbe a un préfixe déterminé pour cet usage.

Les verbes préfixés sont perfectifs, c'est-à-dire que leur présent morphologique a une valeur de futur sémantique. Ils ont un sens différent de celui du verbe simple et ne peuvent donc pas lui correspondre en terme d'aspect. Ils ont besoin d'un imperfectif exprimant le même sens, mais dont le présent morphologique sera un présent sémantique. Cette valeur est obtenue ici par la présence d'un infixe d'itération qui sert aussi à la formation de l'imperfectif. Ces valeurs ayant même combinaison préfixe – racine, « horizontales », sont consignées dans Vallex.

Par contre, la paire « verticale » n'existe pas dans Vallex. Ainsi, à l'entrée « dělat » ne trouve-t-on que l'itératif « dělávat ». La forme « udělat » est isolée, ce qui n'est pas logique. De même, « děkovat » et « poděkovat » sont séparés et non reliés par un renvoi.

L'une des vertus de Vallex est son usage possible pour l'apprentissage du tchèque. A cette fin, le traitement de la paire « verticale » est absolument nécessaire pour savoir quels sont les verbes qui se correspondent. Quelle que soit la langue slave, cette correspondance n'est pas évidente, même pour les autochtones. C'est pourquoi, à plus forte raison, sa notation dans Vallex nous semble indispensable.

3.1.2. Une autre caractéristique de Vallex est liée au traitement automatique du tchèque. En effet, Vallex fournit des données nécessaires à l'analyse (ou la génération) automatique du cadre

verbal et donc à une analyse / génération syntaxique et sémantique de la proposition. Dans cette visée, il nous semble nécessaire de rechercher dès ce niveau de présentation des données le maximum d'automatismes. Il est vraisemblable qu'un certain nombre de transformations puissent être exprimées par des systèmes de règles là où il y a pour le moment duplication du cadre verbal pour des sens qui ne sont pas différents, mais liés l'un à l'autre par transformation de structure. Nous avons à l'esprit des exemples tels que celui de « žít 2 / žnout » (faucher) dont nous avons donné des extraits Vallex plus haut :

« žal trávu na palouku » : il a fauché l'herbe du pré (m.à.m sur le pré)

« žal palouk » : il a fauché le pré

Personnellement, nous donnerions les cadres verbaux dans cet ordre (2 – 1 de Vallex) car faucher le pré, c'est toujours faucher le « x » qui se trouve dessus même si ce « x » n'est pas exprimé (herbe, trèfle, luzerne, …). Lorsqu'il y a omission de « x » PAT, le LOC (ici le pré) subit une translation vers la valeur PAT. Cette transformation est-elle calculable ou plus exactement est-elle transposable dans le formalisme de Vallex ? La Description Générative Fonctionnelle a depuis longtemps adopté la translation des actants situés au-delà du Patient dans un point de vue mêlant les aspects syntaxiques et les aspects sémantiques.

3.1.3. Dans le même ordre d'idée et pour éviter de construire également les cadres verbaux de lexèmes dérivés de verbes, le calcul systématique d'une catégorie lexicale à une autre serait-il envisageable, possible ? Nous pensons particulièrement aux cadres verbaux des substantifs verbaux ou des adjectifs issus de participes verbaux. Seraient-ils déductibles des cadres (des « unités lexicales ») du verbe correspondant ?

3.2. Développements possibles à l'extérieur de ÚFAL

Deux situations nous semblent possibles : la réalisation d'autres Vallex pour des langues autres que le tchèque et l'intégration de Vallex (tchèque ou autre langue pour laquelle pourrait être réalisé un dictionnaire de valences) dans des projets de dictionnaires ou de didactique du tchèque.

3.2.1. Pour le premier point, notre équipe envisage des études sur le cadre verbal en slovaque (Diana Lemay et nous-même) et en albanais (Klara Lagji).

3.2.2. En relation avec ÚFAL, l'exploitation de Vallex comme composante syntaxico-sémantique de lexiques ou de dictionnaires tchèque - français nous semble nécessaire pour de tels projets. L'existence d'un dictionnaire français - tchèque ayant une composante Vallex pour les verbes nous semble encore plus nécessaire pour les besoins de Francophones souhaitant :

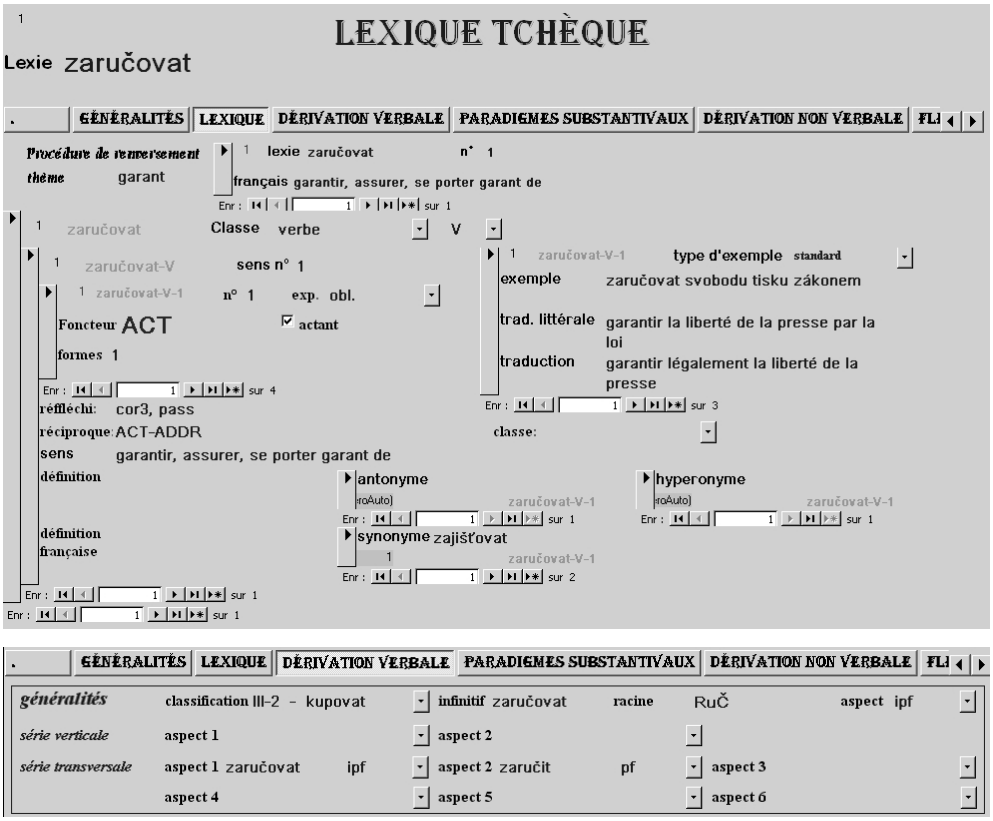
– apprendre le tchèque

– traduire en tchèque

– rédiger en tchèque.

C'est pourquoi nous définissons un projet de base de données tchèque – français englobant des informations Vallex qui sera par la suite renversée pour préparer un dictionnaire français

– tchèque avec la même masse lexicale.



Vallex peut également donner la matière à la constitution d'exercices sur serveur pour les apprenants du tchèque. Dans le cadre de la réalisation d'une méthode d'apprentissage du tchèque, nous ne négligerons pas cette possibilité. Cette méthode est envisagée à la suite de la méthode de slovaque réalisée dans le cadre du projet ALPCU (Lingua II) dont les auteurs sont Elena Baranová, Vlasta Křečková, Diana Lemay et nous-même.

En conclusion, nous soulignerons le fait que Vallex, heureux résultat d'une longue recherche, pourra à son tour donner lieu à d'autres développements en direction de la traduction, de la réalisation de lexiques et de dictionnaires et surtout de la didactique du tchèque.

Bibliographie chronologique de l'équipe ÚFAL

1961 – Sgall, P. : "*Functional Sentence Perspective in a Generative Description*". Prague Studies in Mathematical Linguistics, no. 2.

1967 – Sgall, P. : « *Generativní popis jazyka a česká deklinace* ». Academia, Prague.

1973 – Benešová, J., Sgall, P. : "*Remarks on the Topic/Comment Articulation*". Prague Bulletin of Mathematical Linguistics no. 19.
1974–1975 – Panevová, J. : "*On Verbal Frames in Functional Generative Description*". Prague Bulletin of Mathematical Linguistics no. 22 & 23.
1976 – Panevová, J., Sgall, P. : "*Verbal Frames and Free Adverbials*". International Revue of Slavic Linguistics no. 1.
1977 – Panevová, J. : "*Verbal Frames Revisited*". Prague Bulletin of Mathematical Linguistics no. 28.
1977–1978 – Sgall, P., Hajičová, E. "*Focus on Focus*". The Prague Bulletin of Mathematical Linguistics no. 28 & 29.
1979 – Sgall, P. : "*Towards a Definition of Focus and Topic*". The Prague Bulletin of Mathematical Linguistics no. 31 & 32.
1980 – Sgall, P. : "*Case and Meaning*". The Prague Bulletin of Mathematical Linguistics no. 33.
1980 – Sgall, P. : "*A Dependency-Based Specification of Topic and Focus. Formal Account*". SMIL 1-2. Prague.
1984 – Sgall, P. (ed.) : *Contributions to Functional Syntax, Semantics and Language Comprehension*. Academia, Prague.
1984 – Hajičová, E., Panevová, J. : "*Elementary and Complex Units of the Tectogrammatical Level*". Prague Bulletin of Mathematical Linguistics no. 42, Prague.
1986 – Hajičová, E., Sgall, P. : "*The Ordering Principle*". Prague Bulletin of Mathematical Linguistics no. 45, Prague.
1986 – Sgall, P., Hajičová, E., Panevová, J. : *The Meaning of the Sentence in its Semantic and Pragmatics Aspects*. Academia & Reidel.
1990 – Panevová, J., Sgall, P. : "*Dependency Syntax, its Problems and Advantages*". Prague Series of Mathematical Linguistics no. 10.
1996 – Panevová, J. : "*More Remarks on Control*". Prague Linguistic Circle Papers, John Benjamin.
2003 – Bojar, O. : "*Towards Automatic Extraction of Verb Frames*". Prague Bulletin of Mathematical Linguistics no. 79-80.
2003 – Lopatková, M. : "*Valency in the Prague Dependency Treebank : Building the Valency Lexicon*". Prague Bulletin of Mathematical Linguistics no. 79-80.
2006 – Kolářová, V. : "*Valency of deverbal nouns in Czech*". Prague Bulletin of Mathematical Linguistics no. 86.
2006 – Lopatková, M., Žabokrtský, Z., Benešová, V. : "*Valency lexicon of czech verbs VALLEX 2.0* ". Technical Report 34, UFAL MFF UK, Prague.
2007 – Žabokrtský, Z., Lopatková, M. : « *Valency Information in VALLEX 2.0. Logical Structure of the Lexicon* ». The Prague Bulletin of Mathematical Linguistics, N° 87.
2008 – Lopatková, M., Žabokrtský, Z., Kettnerová, V. : "*Valenční slovník českých sloves*". Karolinum, Prague.
2008 – Panevová, J. : "*VALLEX 2.0 – Valency Lexicon of Czech Verbes and Its Theoretical Background* ". Conférence donnée au LaLIC (Université de Paris-Sorbonne et INALCO), Paris.

Bibliographie extérieure à l'équipe ÚFAL

Daneš, F. (1971) : "*On Linguistic Strata*". Travaux de Linguistique de Prague no. 4.
Daneš, F. (ed.) (1974) : *Papers on Functional Sentence Perspective*. Academia, Prague.
Firbas, J. (1971) : "*On the Concept of Communicative Dynamism in the Theory of Functional Sentence Perspective*". Sborník prací filosofické fakulty brněnské university. Brno.
Mathesius, V. (1936) : "*On some Problems of the Systematic Analysis of Grammar*". Travaux du Cercle Linguistique de Prague, no. 6.
Tesnière, L. (1959) : "*Eléments de syntaxe structurale*", Paris.

Annexe : l'interface HTML de VALLEX