Trotter, David/Bozzi, Andrea/Fairon, Cédrick (éd.) (2016). *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16 : Projets en cours ; ressources et outils nouveaux*. Nancy, ATILF :
http://www.atilf.fr/cilpr2013/actes/section-16.html.

Le corpus multilingue InterCorp et les possibilités de son exploitation

L'objectif de cet article est de présenter un nouvel outil de recherche linguistique, le corpus multilingue InterCorp (www.korpus.cz / intercorp), et les possibilités de son utilisation. Après avoir montré la composition du corpus et le cadre institutionnel de tout le projet dans la section 1, nous allons préciser dans la section 2 les procédés techniques de la constitution du corpus et détailler les fonctions les plus importantes du moteur de recherche qui permet de l'exploiter. Pour terminer, nous donnerons dans la Conclusion quelques exemples de recherches déjà réalisées sur le corpus InterCorp, tout en soulevant les problèmes méthodologiques les plus pertinents qui pèsent sur toute exploitation de corpus parallèle.

1. La composition du corpus et son cadre institutionnel

Le corpus multilingue InterCorp implique actuellement 39 langues, parmi lesquelles des langues romanes – le catalan, le français, l'espagnol, l'italien, le portugais et le roumain. Le catalan n'a été intégré au corpus InterCorp qu'en 2013, mais grâce au dynamisme des coordinateurs et grâce à l'aide du Ministère de l'Éducation nationale de la Principauté d'Andorre, cette section se développe rapidement. Les langues slaves sont également bien représentées : nous y trouvons les sections biélorusse, bulgare, croate, macédonienne, polonaise, russe, slovaque, slovène, serbe, ukrainienne et tchèque. L'anglais ne manque naturellement pas dans ce corpus multilingue, mais d'autres langues germaniques y sont également présentes : l'allemand, le danois, le néerlandais, le norvégien et le suédois. En outre, deux langues difficiles à traiter par les outils du TALN, l'arabe et l'hindi, ont été récemment ajoutés à la liste des langues pour lesquelles le corpus InterCorp offre des données linguistiques ajoutées[1].

InterCorp est un ensemble de corpus parallèles, c'est-à-dire un ensemble de paires composées de textes originaux et de leurs traductions respectives, et les textes sont

[1] Voici la liste complète des langues représentées dans le corpus InterCorp (par ordre alphabétique) : l'albanais, l'allemand, l'anglais, l'arabe, le biélorusse, le bulgare, le catalan, le croate, le danois, l'espagnol, l'estonien, le finnois, le français, le grec, l'hebreu, l'hindi, le hongrois, l'italien, l'islandias, le japonais, le letton, le lithuanien, le macédonien, le malais, le maltais, le néerlandais, le norvégien, le polonais, le portugais, le roumain, le russe, le slovaque, le slovène, le serbe (en alphabets latin et cyrillique), le suédois, l'ukrainien, et le tchèque – langue pivot. L'ajout d'autres langues, telles que le chinois et le turque, le romani, le vietnamien est aussi envisagé.

alignés au niveau des phrases. Pour des raisons techniques ainsi que conceptuelles, la langue tchèque est la langue pivot de l'ensemble du corpus – chaque texte doit avoir sa contrepartie tchèque. En outre, tous les textes sont alignés d'après la segmentation du texte tchèque, ce qui donne au corpus l'homogénéité nécessaire. Néanmoins, dans les recherches sur corpus, il est possible de laisser le tchèque de côté et d'effectuer la recherche seulement sur les langues « étrangères », par exemple sur l'anglais, l'allemand, le catalan, l'espagnol, le finnois, le français et l'italien[2]:

[FR] *« Moi, se dit le petit prince, si j'avais cinquante-trois minutes à dépenser, je marcherais tout doucement vers une fontaine … »* (Antoine de Saint-Exupéry, *Le Petit prince*)

[ALL] *„Wenn ich dreiundfünfzig Minuten übrig hätte", sagte der kleine Prinz," würde ich ganz gemächlich zu einem Brunnen laufen…"* (trad. par Leitgeb, Grete; Leitgeb, Josef)

[ANG] *"As for me," said the little prince to himself, "if I had fifty-three minutes to spend as I liked, I should walk at my leisure toward a spring of fresh water."* (trad. par Katherine Woods)

[CS] *Kdybych já měl padesát tři minuty nazbyt, řekl si malý princ, šel bych docela pomaloučku ke studánce…* (trad. par Zdeňka Stavinohová)

[HIN] „अगर मेरे पास तिरपन मिनिट बिताने को होते „ छोटे राजकुमार ने सोचा, , तो मैं धीरे - धीरे एक जलाशय की ओर चल पड़ता ... ।» (trad. par कशिोर बलवीर, जगवं श)

[IT] *„Io», disse il piccolo principe, «se avessi cinquantatré minuti da spendere, camminerei adagio adagio verso una fontana…»* (trad. par Nini Bompiani Bregoli)

[RO] *«Eu, îşi spuse micul prinţ, dacă aş avea de irosit cincizeci si trei de minute, aş porni în linişte spre o fântână.»* (trad. par Benedict Corlaciu)

Comme nous l'avons signalé ci-dessus, la langue tchèque se trouve en tant que langue pivot au centre du corpus aussi pour des raisons conceptuelles : en effet, tout le projet est centré en particulier sur les recherches contrastives du tchèque par rapport à d'autres langues. Le projet, financé dans son intégralité par le Ministère de l'Éducation nationale tchèque, est réalisé par l'Institut du Corpus national tchèque (www.korpus.cz) qui fait partie de la Faculté des Lettres de l'Université de Prague. L'Institut du Corpus national tchèque a été fondé déjà en 1994 afin de créer le corpus de référence pour la langue tchèque, et cet objectif a été atteint en 2000, quand le corpus SYN2000 a été rendu librement accessible sur Internet. Dans sa partie principale, synchronique, le Corpus national tchèque contient actuellement plus d'un milliard de mots et il est constamment actualisé. Le corpus central est complété par plusieurs corpus spécifiques, par exemple le corpus diachronique, des corpus oraux, des corpus de correspondance ou des textes linguistiques, etc. Depuis 2005, InterCorp, le corpus parallèle, fait également partie de ce vaste projet.

Le corpus InterCorp est le résultat de la coopération de plusieurs établissements et organismes en République tchèque. La préparation des textes en langues étrangères

[2] Le moteur de recherche permet aussi des recherches unilingues sur le sous-corpus choisi ; il suffit de choisir seulement une langue dans le formulaire initial.

est assurée par les coordinateurs qui appartiennent en général aux différents départements linguistiques de la Faculté des Lettres de l'Université Charles à Prague. Quant aux outils informatiques nécessaires pour la préparation et l'exploitation du corpus, ils sont souvent issus de la coopération avec des spécialistes en TALN d'autres établissements, par exemple l'Institut de la linguistique théorique et computationnelle de la Faculté des Lettres de l'Université à Prague (http://utkl.ff.cuni.cz), l'Institut de la linguistique formelle et appliquée de la Faculté des Mathématiques et de la Physique de la même Université (http://ufal.mff.cuni.cz/) ou la Faculté de l'Informatique de l'Université Masaryk de Brno (http://www.fi.muni.cz). L'Institut du Corpus national tchèque est ensuite responsable de la coordination de toutes les activités, du stockage central des données, de leur traitement informatique et de leur mise en ligne via le moteur de recherche (‹http://kontext.korpus.cz›).

La première version du corpus InterCorp a été mise en ligne en 2008, mais depuis, le corpus ne cesse de croître : après la dernière actualisation en mai 2015, l'ensemble du corpus InterCorp contient 1 597 462 625 mots. Le noyau du corpus est constitué par les textes littéraires, ce qui représente la spécificité du projet en comparaison avec d'autres corpus parallèles, tels que OPUS ou Hansard par exemple. Le sous-corpus littéraire, contenant actuellement presque 280 millions de mots, est complété par quatre grands ensembles de textes. Premièrement, il s'agit de textes journalistiques tirés de serveurs multilingues *Project Syndicate* (www.project-syndicate.org) et *Presseurop / Voxeurop* (www.voxeurop.eu) en dix langues (51 177 809 mots au total)[3]. En deuxième lieu, il est possible d'effectuer les recherches sur les textes juridiques de l'Acquis communautaire de l'Union européenne, disponibles dans toutes les langues de communication des états membres (450 498 235). Le troisième ensemble de textes n'a été ajouté qu'en avril 2013 – il s'agit du corpus EuroParl, constitué par les débats au Parlement européen (277 951 947 mots). La dernière collection, intégrée au corpus en 2014, contient des sous-titres de films (Subtitles, 539 060 969 mots).

[3] Les collections *Presseurop* et *Project Syndicate* sont disponibles en allemand, anglais, espagnol, français et italien ; le russe est inclus seulement dans le *Project Syndicate*, et le néerlandais, le polonais, le portugais et le roumain font partie uniquement de *Presseurop*.

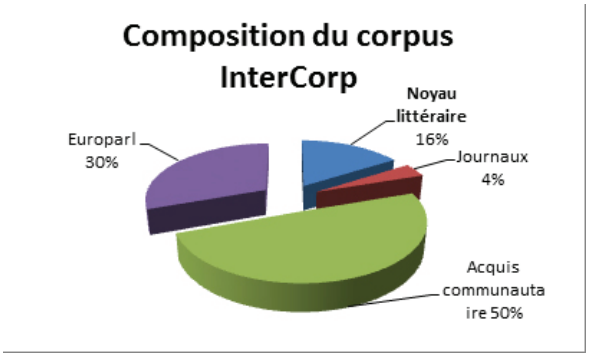

Figure 1 – Composition du corpus InterCorp (par types de textes)

Figure 1 montre les données pour l'ensemble du corpus InterCorp, mais il faut préciser que les différentes langues ne sont pas représentées dans le corpus de manière égale, comme le révèle *Figure 2* ci-bas. Parmi les langues les mieux représentées, nous trouvons par exemple l'espagnol : ce sous-corpus contient plus de 100 millions de mots dans sa totalité, dont 17 millions dans le noyau littéraire. Dans cette section du corpus, nous pouvons effectuer des recherches par exemple sur plusieurs textes de Gabriel Garcia Marquez ou de Mario Vargas Llosa. L'italien et le français ne sont pas en reste – leurs sous-corpus comptent presque 90 millions de mots (le français) et 65 millions de mots (l'italien), dont par exemple le français 9 millions de mots dans le noyau littéraire. Dans ces sections, l'utilisateur peut effectuer des recherches par exemple dans les textes de Patrick Chamoiseau, Albert Camus ou Amélie Nothomb pour le français, et dans des œuvres de Umberto Eco ou Alessandro Baricco pour l'italien.

Les différences de taille entre les sections linguistiques dépendent tant du taux d'activité des différents coordinateurs que de la disponibilité des textes – par exemple le nombre de traductions du portugais en tchèque (et vice versa) est assez limité, et ce sous-corpus doit ainsi être complété par des traductions d'une troisième langue, par exemple des livres de J.R.R. Tolkien ou de J.K. Rowling. Il faut également préciser que le noyau dit « littéraire » contient aussi des textes appartenant à d'autres genres, par exemple des essais (en français des textes d'Albert Camus ou de Saint-Exupéry) ou des textes scientifiques, comme ceux de Georges Duby[4]. Dans la section française, nous pouvons trouver également les versions parallèles de cinq volumes des aventures d'Astérix et Obélix[5].

[4] Avant la fin de l'année 2013, des textes de Michel Foucault, Henriette Walter et Ferdinand de Saussure ont été ajoutés au corpus.

[5] En définissant le corpus de travail, l'utilisateur peut préciser les genres qu'il désire inclure dans sa recherche.

Langue	Noyau	Journaux	Acquis	Europarl	Subtitles
ar	34 325	0	0	0	0
be	2 152 724	0	0	0	0
bg	5 240 831	0	13 816 405	9 083 403	0
ca	4 632 696	0	0	0	0
da	3 016 838	0	21 679 997	13 915 841	14 429 778
de	27 681 897	6 207 922	21 723 929	13 089 209	8 366 765
el	0	0	25 069 611	15 403 662	23 714 597
en	15 488 167	6 488 284	24 207 801	15 580 109	52 101 283
es	17 475 748	7 140 829	27 001 343	15 885 394	36 378 715
et	0	0	15 962 544	10 899 550	10 296 031
fi	3 426 226	0	16 455 144	10 175 256	15 097 653
fr	9 170 042	7 321 278	27 351 591	17 178 444	25 961 848
he	0	0	0	0	16 221 237
hi	408 616	0	0	0	0
hr	15 479 547	0	0	0	19 092 559
hu	5 387 533	0	19 176 514	12 306 692	21 239 634
is	0	0	0	0	1 584 758
it	7 247 545	3 359 150	24 849 477	15 489 468	14 653 613
ja	0	0	0	0	113 320
lt	358 253	0	18 392 644	11 212 864	557 961
lv	1 336 888	0	18 744 927	11 688 597	280 117
mk	3 741 900	0	0	0	1 877 210
ms	0	0	0	0	3 520 701
mt	0	0	14 133 133	0	0
nl	9 961 680	3 269 635	24 746 144	15 563 231	29 362 826
no	4 815 797	0	0	0	0
pl	17 516 332	2 378 025	20 627 627	12 811 143	26 572 483
pt	2 393 287	3 369 337	28 602 556	16 484 692	43 391 919
ro	3 432 615	2 737 807	8 199 565	9 446 369	34 128 511

ru	3 337 545	3 174 152	0	0	6 885 753
sk	7 401 998	0	19 222 784	12 734 444	5 134 150
sl	900 221	0	19 645 598	12 240 548	17 024 593
sq	0	0	0	0	2 003 579
sr	8 823 894	0	0	0	20 776 850
sv	8 138 161	0	20 585 800	13 840 373	14 693 861
tr	0	0	0	0	21 190 828
uk	5 054 034	0	0	0	246 059
vi	0	0	0	0	1 473 591
cs	84 718 325	5 731 390	20 303 101	12 922 658	50 688 186
TOTAL	278 773 665	51 177 809	450 498 235	277 951 947	539 060 969

Figure 2 – Composition du corpus InterCorp (par langues en nombre de mots)

Les textes pour le corpus sont choisis en fonction de plusieurs critères. Le premier critère limitant le choix de textes est la date de la création de l'original – étant donné le caractère synchronique du corpus, seuls les textes écrits après la seconde guerre mondiale sont admis. Néanmoins, pour certains textes appartenant au patrimoine littéraire universel, des exceptions ont été autorisées ; c'est ainsi que nous pouvons trouver dans le corpus *Le Voyage au bout de la nuit* de Ferdinand Céline ou des textes de l'auteur tchèque Karel Čapek, qui est mort en 1938. Deuxièmement, les coordinateurs sont encouragés à choisir pour le corpus les textes susceptibles d'être traduits en plusieurs langues, pour favoriser les intersections entre les différentes sections linguistiques. Ainsi, il n'est pas surprenant de trouver parmi les textes les mieux représentés dans le corpus des best-sellers anglais, tels que *Le Seigneur des anneaux* ou *Harry Potter*. Cependant, parmi les textes les mieux « cotés », nous trouvons aussi un texte français, *Le Petit Prince* d'Antoine de Saint-Exupéry, qui est représenté dans le corpus en 25 versions linguistiques. Plusieurs textes tchèques sont également présents dans un grand nombre de versions linguistiques, des romans de Milan Kundera (en particulier *La Plaisanterie* ou *L'Insoutenable légèreté de l'être*) et le *Brave soldat Chvéïk* de Jaroslav Hašek[6].

[6] Le corpus InterCorp est ouvert à la coopération avec d'autres projets : à titre d'exemple, en 2013, l'auteur du corpus ASPAC (*The Amsterdam Slavic Parallel Aligned Corpus*, cf. Waldenfels 2006) a fourni ses textes parallèles au projet InterCorp, ce qui a permis d'améliorer leur alignement et de les rendre accessibles via le moteur de recherche *NoSketch Engine*. De même, le Corpus national tchèque a accueilli et rendu disponible le corpus de sorabe (DOTKO et HOTKO).

2. Procédés techniques de la constitution du corpus et son exploitation

Le traitement informatique des textes choisis pour le corpus est effectué en plusieurs étapes, mais la procédure est différente dans le cas des textes du noyau littéraire et ceux des collections de textes, telles que Presseurop ou Europarl.

Les textes destinés au noyau littéraire sont océrisés et entièrement relus par des collaborateurs du projet, dans la plupart des cas étudiants universitaires de la langue concernée. Les textes sont ensuite exportés à l'aide d'un macro de Visual Basic dans un format quasi-XML qui identifie les frontières entre les paragraphes et phrases et transforme les caractères spéciaux (&, <, >) en entités. Pour le tchèque, nous utilisons le segmentateur basé sur des règles, pour les autres langues, il s'agit d'un algorithme à l'apprentissage non-supervisé (*Punkt sentence tokenizer*, cf. Kiss / Strunk, 2006). Dans l'étape suivante, le coordinateur de la section linguistique concernée enregistre le texte dans la base de données rassemblant tous les textes du corpus et lui assigne les informations bibliographiques nécessaires.

Le coordinateur principal du projet effectue ensuite l'alignement automatique de la paire de textes donnée, à l'aide du logiciel *hunalign* (cf. Varga et al., 2005, http://mokk.bme.hu / rerources / hunglishcorpus /), et il enregistre le résultat dans l'éditeur de textes parallèles nommé *InterText*. Ce logiciel, créé directement pour le projet InterCorp (cf. Vondřička, 2010, http://wanthalf.saga.cz / intertext), permet aux collaborateurs du projet de corriger l'alignement automatique appliqué ultérieurement aux textes :

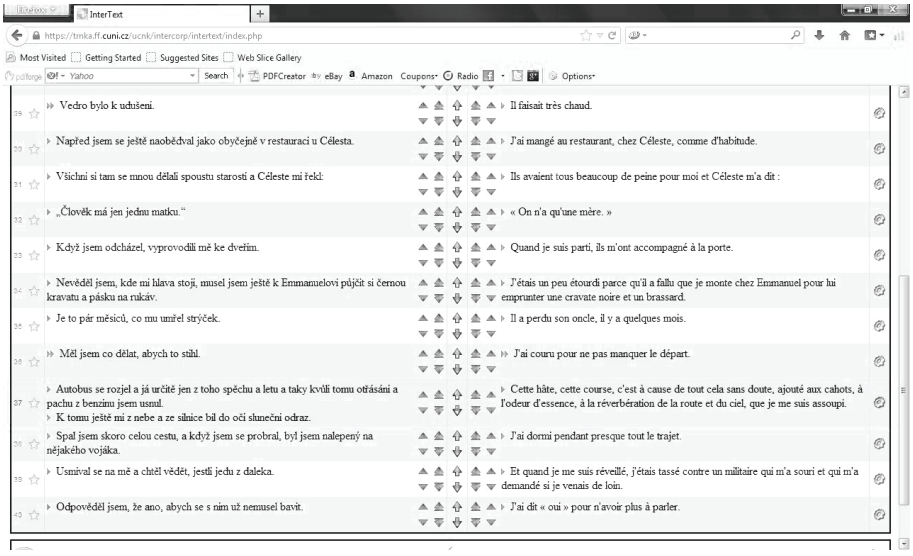

Figure 3 : L'éditeur de textes parallèles InterText

En ce qui concerne les textes des collections (*EuroParl*, *Acquis communautaire*, *Project Syndicate, Presseurop et Subtitles*), la relecture et la correction semi-manuelle de l'alignement automatique n'y sont pas appliquées. Par conséquent, les textes de ce type peuvent être ajoutés au corpus plus rapidement, en plus grande quantité et à un coût considérablement moins élevé que les textes du noyau littéraire, mais le résultat final est proportionnellement moins fiable. De plus, les collections de Presseurop / Syndicate et de l'Acquis communautaire présentent un inconvénient limitant leur utilité pour les recherches contrastives, parce qu'elles ne permettent pas d'identifier la direction de la traduction, c'est-à-dire la langue source. Pour les recherches contrastives et traductologiques, cette lacune représente un défaut majeur, étant donné les spécificités de la langue de traduction. Les techniciens participant au projet tentent actuellement de remédier à ce problème en identifiant la langue source à partir des données bibliographiques incluses directement dans les textes, et cette information devrait être disponible dans la prochaine version pour les textes journalistiques et pour Europarl (pour l'Acquis communautaire, la procédure est envisagée).

Les textes issus des alignements automatique ainsi que semi-automatique sont ensuite exportés et les liens d'alignement sont stockés dans un fichier séparé. Finalement, les textes alignés sont dotés de la lemmatisation et de l'étiquetage morphologique. Toutes les langues impliquées dans le projet InterCorp sont lemmatisées, à l'exception du bulgare, du hongrois et du néerlandais[7]. Pour l'annotation morphologique, le logiciel le plus utilisé dans le corpus est TreeTagger (www.ims.uni-stuttgart.de / projekte / corplex / TreeTagger), mais le tchèque et le slovaque sont annotés à l'aide de l'annotateur *morče* (http://ufal.mff.cuni.cz / morce), qui est capable de refléter la richesse morphologique de ces langues (il offre 15 positions de sous-ensembles de marques morphologiques, cf. Hajič, 2004). Les textes sont enfin reliés avec les données bibliographiques et indexés par le logiciel (gestionnaire de corpus) Manatee (cf. Rychlý, 2007). La nouvelle version du corpus est mise en ligne approximativement tous les six mois. L'actualisation qui a eu lieu en avril 2013, a apporté deux changements importants : le corpus a été considérablement élargi et le moteur de recherche utilisé pour l'exploitation du corpus est désormais *NoSketch Engine*, offrant un large éventail de possibilités de tri et de traitement statistique des concordances[8].

L'accès au corpus est disponible à partir du site ‹http://kontext.korpus.cz› et il est gratuit après inscription. Pour s'inscrire et obtenir le mot de passe, il suffit de remplir le formulaire en ligne où l'on s'engage à ne pas utiliser le corpus à des fins commerciales et à signaler toute publication que l'on réalise grâce aux données issues d'InterCorp (www.korpus.cz / english / dohody.php). Le même mot de passe est également valide

[7] La lemmatisation du néerlandais est prévue pour la prochaine actualisation du corpus.

[8] L'ancien concordancier, nommé Park, est toujours disponible sur ‹http://korpus.cz / Park / login›. Dans l'évaluation des données statistiques issues des deux logiciels, il faut prendre en considération le fait que les calculs dans Park sont opérés sur le nombre de mots, tandis que ceux de NoSketch Engine sont basés sur le nombre de tokens (positions), incluant non seulement les mots, mais aussi la ponctuation. En linguistique de corpus, la deuxième approche est la plus courante.

pour le Corpus national tchèque et pour quatre larges corpus unilingues contenant chacun plus d'un milliard de mots – les WaCky corpora pour l'allemand, le français, l'italien et l'anglais britannique (cf. Baroni et al., 2009)[9]. Avant la fin de l'année 2013, nous avons ajouter aussi l'accès au corpus *Est républicain* dont les textes ont été rendus disponibles par le Centre national de Ressources Textuelles et Lexicales (‹http://www.cnrtl.fr / corpus / estrepublicain›). Tous les corpus unilingues sont lemmatisés et dotés de l'annotation morphologique.

La recherche sur InterCorp s'effectue en trois étapes, comme dans la plupart des corpus :

(1) La définition du sous-corpus de travail

L'utilisateur peut choisir deux ou plusieurs langues qu'il veut analyser, par exemple le français, l'anglais et le finnois, et définir le sous-corpus de travail en fonction de plusieurs critères :

(a) le genre du texte (roman, drame, texte juridique, scientifique, etc.) ;
(b) la langue source et la distinction original / traduction ;
(c) les textes concrets.

Le sous-corpus de travail peut être sauvegardé pour que l'utilisateur puisse y revenir à chaque session.

(2) La formulation de la requête

La requête dans le corpus InterCorp peut porter sur un mot simple (*chat*) ou sur un lemme (*chanter*), mais aussi combiner les expressions régulières et les marques morphologiques dans une expression CQL (*Corpus Query Language*). Par exemple, pour chercher les gérondifs en français, il est possible de profiter des marques morphologiques assignées par TreeTagger ([word="en"] []{0,2} [tag="VER:ppre"]), ou utiliser les expressions régulières ([word="en"] []{0,2} [word=".*ant"])[10]. Les requêtes dans les autres langues choisies peuvent rester libres, mais il est également possible de préciser que les phrases équivalentes en anglais doivent contenir la forme en –ing ([word=".*ing"]) ou l'équivalent systémique présumé du gérondif en tchèque, le transgressif présent ([tag="Ve.*"])[11]. Les informations affichées dans la concordance obtenue indiquent les données bibliographiques concernant les textes sources, mais

[9] La lemmatisation du néerlandais est prévue pour la prochaine actualisation du corpus.

[10] Avant de formuler la requête impliquant l'annotation morphologique, il est nécessaire de consulter les listes de marques (*tagsets*) pour les langues choisies (cf. la liste ‹http://www.korpus.cz / intercorp / ?req=page:info›). La définition et l'utilisation d'une même marque ainsi que la segmentation peuvent varier d'une langue à l'autre (cf. Rosen, 2010 et 2012).

[11] Pour créer la requête en tchèque sans avoir à connaître les listes de marques pour toutes les 15 positions de l'annotation morphologique, l'utilisateur peut profiter de l'application nommée *klikátko*, qui lui permet de sélectionner les marques directement d'après leur description (‹http://utkl.ff.cuni.cz / ~skoumal / morfo / ?lang=cs›).

aussi des données statistiques, concrètement la fréquence absolue et relative (i.p.m.) de l'élément analysé.

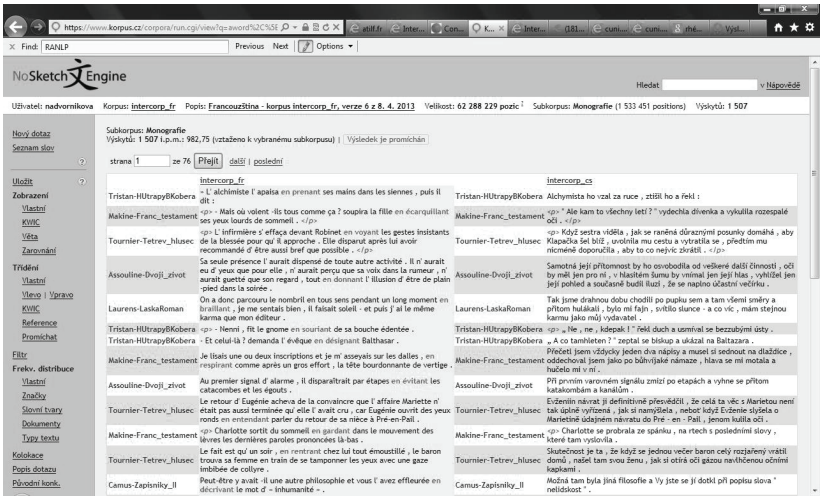

Figure 4 : Exemple de concordance dans le corpus InterCorp

(3) Le tri des résultats et leur traitement statistique

NoSketch Engine permet de trier et filtrer les résultats visionnés dans la concordance en fonction de plusieurs critères, par exemple par leur contexte de droite ou de gauche ou par les propriété de KWIC (forme, lemme, marque morphologique). Le concordancier offre également un grand nombre de possibilités de traitement statistique des données et de calcul de collocations (t-score, MI-score, logDice, etc.). La distribution de fréquence visualise de manière claire les fréquences absolue et relative dans les différents documents du corpus :

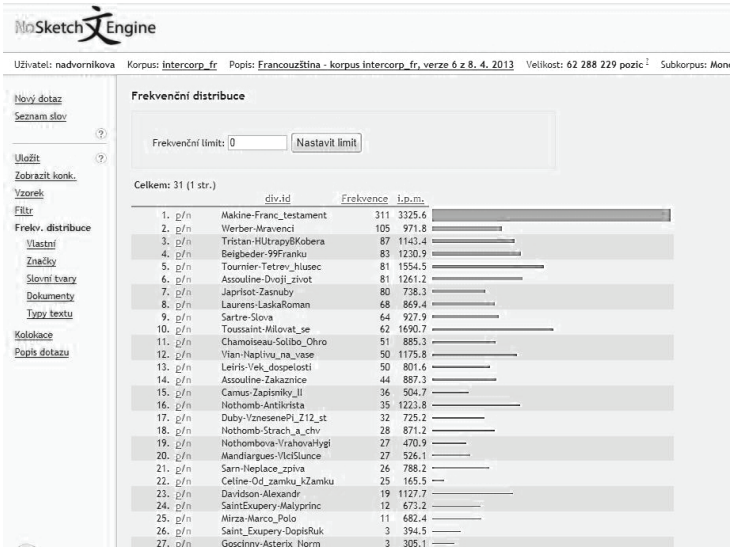

Figure 5 : Visualisation des fréquences relative et absolue de la distribution du gérondif dans le sous-corpus français

A la fin de la session, l'utilisateur peut télécharger la concordance finale dans un fichier Excel et continuer ses analyses.

3. Conclusion

Le corpus parallèle InterCorp était initialement destiné en particulier aux linguistes, mais la communauté de ses utilisateurs est actuellement bien plus variée : les traducteurs l'utilisent comme source d'inspiration en cas de problème, les enseignants de langues profitent de cet outil informatique pour faire varier leurs méthodes en classe et les étudiants en tchèque langue étrangère dispersés aux quatre coins du monde y cherchent les solutions qu'il serait impossible ou trop long de trouver dans les grammaires et les dictionnaires. Par ailleurs, l'Institut du Corpus national tchèque s'ouvre de plus en plus au large public et organise des stages d'initiation au travail sur corpus non seulement à Prague, mais aussi dans d'autres villes en République tchèque.

Néanmoins, les linguistes restent le noyau du groupe des utilisateurs du corpus et réalisent des recherches contrastives sur la plupart des langues représentées dans le corpus[12]: P. Čermák et P. Štichauer ont par exemple comparé les constructions factitives en espagnol, en italien et en tchèque (Čermák / Štichauer, 2010) ou le préfixe *re-* et le suffixe *-ble / -bile* à la lumière de leurs équivalents tchèques (voir Čermák, 2013 et

[12] L'Institut du Corpus national tchèque a déjà organisé deux colloques internationaux présentant des recherches effectuées sur InterCorp (voir Čermák / Corness / Klégr, 2010 et Čermák, 2011) et un autre colloque s'est tenu en septembre 2014 (http://www.korpus.cz / kl2014 /).

Bratánková / Štichauer, 2011). En allemand, T. Káňa et H. Pelouškova ont étudié les diminutifs ou la construction modale *ohne* + *zu* en comparaison avec leurs équivalents en tchèque (Káňa / Pelouškova, 2009 et 2011 et Čermák / Nádvorníková *et al.* 2015) ; et la section anglaise s'est concentrée sur l'analyse des prépositions (Malá / Šaldová / Klégr, 2010) ou l'ordre de mots et la portée de rhématiseurs (Martinková, 2012). L'atout principal du corpus consiste en particulier dans sa capacité d'offrir au chercheur de larges données authentiques qui lui permettent de vérifier les informations présentées comme incontestables dans les grammaires et les dictionnaires[13]. Le corpus InterCorp a ainsi permis de mettre en doute la tradition prétendant l'équivalence systémique entre le gérondif français et le transgressif présent tchèque : les données tirées du corpus montrent que l'équivalent français le plus fréquent de cette forme verbale tchèque n'est pas le gérondif, mais le participe présent (Nádvorníková, 2010b). Même les étudiants se lancent dans des recherches lexicales et grammaticales ponctuelles, explorant par exemple l'emploi de connecteurs (*en effet*) ou du conditionnel passé.

Tout en approfondissant leurs recherches sur le corpus multilingue InterCorp, les linguistes se rendent compte des contraintes méthodologiques multiples qui pèsent sur son exploitation. Premièrement, ils doivent prendre en considération les différences de taille entre les sections linguistiques du corpus ainsi que les spécificités des genres qui y sont inclus (romans, textes juridiques de l'Union européenne, etc.). De plus, les linguistes doivent accepter le fait qu'un corpus parallèle ne peut jamais être balancé et représentatif, parce que certains types de textes ne sont presque jamais traduits ; et l'oral est quasiment exclu de ce domaine[14]. Deuxièmement, les recherches sur corpus parallèles se heurtent au problème donné par la spécificité de la langue de traduction (*translationese*) qui tend à la normalisation et à la simplification (cf. par exemple Olohan, 2004, 91-104 ou Baker, 1996, 176-177). Les linguistes réagissent à ce problème d'une part en vérifiant les résultats sur les larges corpus unilingues, d'autre part en respectant strictement la direction de la traduction (la langue source)[15]. L'analyse bi-directionnelle, profitant des deux directions de la traduction, s'avère ainsi très prometteuse (cf. Malá, 2013). L'intégration de ce respect pour les spécificités de

---

[13] L'Institut d'études romanes de la Faculté des Lettres de l'Université à Prague a entamé récemment un vaste projet impliquant quatre langues romanes (l'espagnol, le français, l'italien et le portugais) dans la recherche sur cinq sujets, à savoir le préfix *re-*, le suffixe *-ble/ -bile*, les constructions inchoatives et factitives, et le gérondif. Les résultats de cette recherche sur corpus ont été publiés en 2015.

[14] Les textes du corpus Europarl ne reflètent pas la production orale spontanée et les textes sont très probablement post-édités. Le corpus parallèle de sous-titres de films pourrait refléter davantage la langue parlée authentique. (voir Čermák / Nádvorníková 2015).

[15] Pour le français, il est possible de vérifier les résultats sur FRANTEXT (www.frantext.fr) ; dans le cas du gérondif, cette procédure a révélé que la fréquence relative de cette forme verbale était plus élevée dans le noyau littéraire du corpus InterCorp (2 066 i.p.m.) que dans le corpus de 192 romans publiés après 1950 et disponibles dans FRANTEXT (1 629 i.p.m.). L'analyse plus détaillée des textes a ensuite démontré que cette différence était liée surtout à l'idiolecte d'un seul auteur, A. Makine, où la fréquence relative du gérondif s'élève à 5 135 i.p.m. (cf. Nádvorníková, 2012, 165-166).

la langue de traduction est le fruit de la coopération avec les traductologues, qui offre à la linguistique de corpus contrastive les méthodes d'analyse plus fines.

Les projets d'avenir pour le corpus multilingue InterCorp ne devraient donc pas être concentrés seulement sur l'élargissement du corpus et l'amélioration de son moteur de recherche (par exemple par l'ajout de l'alignement par mots), mais les linguistes utilisant les données parallèles devraient se lancer dans un débat méthodologique sérieux qui leur permettrait de profiter pleinement des vastes champs d'exploration que les corpus multilingues leur ouvrent.

Université Charles à Prague Olga NÁDVORNÍKOVÁ

Références bibliographiques

Baker, Mona, 1996. « Corpus-based translation studies: The challenges that lie ahead », in : Somers, H. (ed.), *Terminology, LSP and Translation: Studies in languagae engineering, in honour of Juan C. Sager*, Amsterdam, John Benjamins, 175-186.

Baroni, Marco / Bernardini, Silvia / Ferraresi, Adriano / Zanchetta, Eros, 2009. « The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora », *Language Resources and Evaluation*, 43 (3), 209-226.

Bratánková, Leontýna / Štichauer, Pavel, 2011. « Italský iterativní prefix *ri-* a jeho české protějšky », in : Čermák, F. (ed.), *Korpusová lingvistika Praha 2011 – 1 InterCorp*, Praha, Nakladatelství Lidové Noviny, 136-143.

Čermák, František / Corness, Patrick / Klégr, Aleš (ed.), 2010. *InterCorp: Exploring a Multilingual Corpus*, Praha, Nakladatelství Lidové noviny.

Čermák, František (ed.), 2011. *Korpusová lingvistika Praha 2011 – 1 InterCorp*, Praha, Nakladatelství Lidové noviny, 2011.

Čermák, František / Rosen, Alexandr, 2012. « The case of InterCorp, a multilingual parallel corpus », *International Journal of Corpus Linguistics*, 13, 3, 411-427. ‹http://utkl.ff.cuni.cz / ~rosen / public / 2012_intercorp_ijcl.pdf›.

Čermák, Petr, 2007. « Acerca de los corpora paralelos: el proyecto Intercorp », *Verba, Anuario Galego de Filoloxia* (Santiago de Compostela), 34, 377-382.

Čermák, Petr, 2012. « *Al + infinitivo*: estudio contrastivo checo-español », *Acta Universitatis Carolinae – Philologica*, 2 / 2009 – *Romanistica Pragensia* XVIII, 2010, 49-64.

Čermák, Petr, 2013. « Las posibilidades de estudio ofrecidas por los corpus paralelos: el caso del prefijo español *re-* », *Acta Universitatis Carolinae – Philologica*, 2 / 2013 – *Romanistica Pragensia XIX* (*Les langues romanes à la lumière des corpus linguistiques*), 2013, 123-136.

Čermák, Petr / Štichauer, Pavel, 2010. « Španělské a italské kauzativní konstrukce *hacer / fare + sloveso* a jejich české ekvivalenty », in : Čermák, F. / Kocek, J. (et al.), *Mnohojazyčný korpus Intercorp: Možnosti studia*, Praha, Nakladatelství Lidové Noviny, 70-90.

Čermáková, Anna / Fárová, Lenka, 2010. « Keywords in Harry Potter and their Czech and Finnish translation equivalents », in : Čermák, František / Corness, Patrick / Klégr, Aleš (ed.), *InterCorp: Exploring a Multilingual Corpus*, Praha, Nakladatelství Lidové noviny, 177-188.

Čermák, Petr / Nádvorníková, Olga *et al*., 2015. *Románské jazyky a čeština ve světle paralelních korpusů* [*Les Langues romanes et le tchèque à la lumière des corpus parallèles*], Praha. Karolinum.

Hajič, Jan, 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, Praha, Karolinum.

Káňa, Tomáš / Pelouśková, Hana, 2011. *Deutsch und Tschechisch im Vergleich: Korpusbasierte linguistische Studien* II, Brno, Masarykova univerzita.

Káňa, Tomáš / Pelouśková, Hana, 2009. *Deutsch und Tschechisch im Vergleich. Kopusbasierte linguistische Studien* I, Brno, Masarykova univerzita.

Kiss, Tibor / Strunk, Jan, 2006. « Unsupervised multilingual sentence boundary detection », *Computational Linguistics*, 32 (4), 485-525.

Malá, Markéta, 2013. « Translation counterparts as markers of meaning. The case of copular verbs in a parallel English–Czech corpus », *Languages in Contrast*, 13(2), 170-192.

Malá, Markéta / Šaldová, Pavlína / Klégr, Aleš, 2010. « English equivalents of the Czech preposition *v / ve* from the point of view of the 'open-choice principle' and the 'idiom principle' », in : Čermák, František / Corness, Patrick / Klégr, Aleš (ed.), *InterCorp: Exploring a Multilingual Corpus*, Praha, Nakladatelství Lidové noviny, 2010, 118-137.

Martinková, Michaela, 2012. Subject-operator inversion after sentence-initial *only* seen through its Czech equivalents, *Linguistica Pragensia*, 22 (2), 79-97.

Nádvorníková, Olga, 2010a. « Les corpus parallèles : L'Espace pour l'analyse contrastive », *Études Romanes de Brno*, 31 (1), 7-27.

Nádvorníková, Olga, 2010b. *The French* gérondif *and its Czech equivalents*, in : Čermák, František / Corness, Patrick / Klégr, Aleš (ed.), *InterCorp: Exploring a Multilingual Corpus*, Praha, Nakladatelství Lidové noviny, 2010, 83-96.

Nádvorníková, Olga, 2012. *Korpusová analýza faktorů sémantické interpretace francouzského gérondivu* [Thèse de doctorat], Filozofická fakulta Univerzity Karlovy v Praze, dir. H. Loucká.

Olohan, Maeve, 2004. *Introducing Corpora in Translation Studies*, London / New York, Routledge.

Rosen, Alexandr, 2010. « Morphological tags in parallel corpora », in : Čermák, František / Corness, Patrick / Klégr, Aleš (ed.), *InterCorp: Exploring a Multilingual Corpus*, Praha, Nakladatelství Lidové noviny, 2010, 205-234. ‹http://utkl.ff.cuni.cz / ~rosen / public / unitags_pap.pdf›.

Rosen, Alexandr, 2012. « Mediating between incompatible tagsets », in : Ahrenberg, Lars et al. (ed), *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC), Volume 10 of NEALT Proceedings Series*, Tartu, Northern European Association for Language Technology, 53-62.

Rosen, Alexandr / Vavřín, Martin, 2012. « Building a multilingual parallel corpus for human users », in : Calzolari, N. et al. (ed), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC´12)*, Instanbul, European Language Resources Association (ELRA), 2447-2452.

Rychlý, Pavel, 2007. « Manatee / Bonito – a modular corpus manager », in : *Ist Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno, Masarykova univerzita, 65-70.

Svášek, Martin, 2007. *Définition, élaboration et exploitation d'un corpus parallèle bidirectionnel français – tchèque tchèque – français* [Thèse de doctorat], Filozofická fakulta Univerzity Karlovy v Praze (dir. V. Petkevič) / INALCO (LALIC – CERTAL), Paris (dir. P. Pognan).

Varga, Dániel et al., 2005. « Parallel corpora for medium density languages », in : *Proceedings of RANLP 2005*, Borovets, Bulgaria, 590-596. ‹http://www.ldc.upenn.edu / Catalog / docs / LDC2008T01 / ranlp05.pdf›.

Vavřín, Martin / Rosen, Alexandr, 2008. « InterCorp: A Multilingual Parallel Corpus Project », in : Proceedings of the International Conference Corpus Linguistics - 2008, St. Petersburg State University, 97-104. ‹http://utkl.ff.cuni.cz / ~rosen / public / 2008_intercorp_peterburg.pdf›.

Vondřička, Pavel, 2010. « TCA2 - nástroj pro zpracovávání překladových korpusů » [TCA2 - a tool for processing translation corpora], in : Čermák, F. / Kocek, J. (et al.), *Mnohojazyčný korpus Intercorp: Možnosti studia*, Praha, Nakladatelství Lidové Noviny, 225-231.

Waldenfels von, Ruprecht, 2006. « Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alingment », in : Brehmer, B. / Zdanova, V. / Zimny, R. (ed.), *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)*, *Volume 9*, München, Verlag Otto Sagner, 123-138.

Cet article a été réalisé grâce au soutien financier du projet Český národní korpus (LM2011023), *soutenu par le Ministeère de l'Éducation nationale tchèque dans le cadre de l'activité* « Projekty velkých infrastruktur pro VaVal ».