Simões, Barreiro, Santos, Sousa-Silva & Tagnin (eds.) *Linguística, Informática e Tradução : Mundos que se Cruzam*, Oslo Studies in Language 7(1), 2015. 21–37. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

CORPÓGRAFO, TERMINOLOGIE, PHRASÉOLOGIE

FRANÇOISE BACQUELAINE

ABSTRACT

The Corpógrafo results from interdisciplinary collaboration between linguists and computer engineers under Belinda Maia's direction. This user-friendly tool for building and using tailor-made corpora allows not only for terminology extraction and management, but also for any research based on monolingual, comparable or parallel corpora. This paper presents the Corpógrafo's evolution from the first to the fourth version, and two experiences of its use in three languages (English, French and Portuguese). The first experience is in the field of Bluetooth technology terminology extraction and management. The second deals with four Portuguese structures containing the universal quantifier *cada* and expressing progression, «dropper», proportion between two sets of events or entities and proportion between a set and a subset of events or entities. These experiences show the strengths, weaknesses and limits of the Corpógrafo.

[1] INTRODUCTION

Le Corpógrafo, la Terminologie et la Phraséologie marquent les trois étapes d'un parcours sous la direction de Belinda Maia. Le Corpógrafo nous a été présenté en première année de master (Mestrado em Terminologia e Tradução) et a été utilisé dans le cadre d'un travail de groupe en Terminologie au deuxième semestre 2005–2006 pour produire une base de données terminologiques (BDT) dans le domaine des télécommunications sans fil. Cette initiation au Corpógrafo a déterminé la suite : mémoire de master en Terminologie (2006–2008, soutenance en janvier 2009) et découverte en 2009–2010, grâce au Corpógrafo, de l'objet linguistique de notre thèse de doctorat actuellement en cours. Le choix du Corpógrafo pour contribuer à cet hommage rendu à celle qui dirige nos recherches depuis bientôt dix ans s'est donc imposé.

Dans un premier temps, nous nous inspirons d'un rapport de la Linguateca (Santos 2005) et de publications des membres de l'équipe de Porto pour retracer l'évolution du Corpógrafo depuis le Gestor de Corpora en 2003 (Sarmento & Maia 2003 ; Maia & Sarmento 2003), jusqu'à la quatrième et dernière version présentée en 2008 (Maia & Matos 2008) [1]. Une fois dressé le portrait du Corpógrafo, deux expériences de recherche sur corpus en contexte de formation sont présen-

1. D'autres sources consultées mais non citées figurent dans la bibliographie d'autres articles de cet ouvrage, dont la plupart sont disponible sur le site de la Linguateca à http://www.linguateca.pt/.

tées. L'une concerne l'élaboration d'une BDT trilingue (anglais, français, portugais) dans le domaine de la technologie de télécommunication sans fil Bluetooth. L'autre a révélé la prédominance du quantificateur universel portugais cada sur le quantificateur universel pluriel *todos* (*os*) dans certains corpus de spécialité, alors que *each* et *chaque* sont moins fréquents que *all* ou *tous* (*les*) en anglais et en français, quels que soient les corpus. Ces deux exemples illustrent bien deux des principales applications pédagogiques et scientifiques du Corpógrafo dans la perspective de l'utilisateur.

[2] CORPÓGRAFO

La genèse du Corpógrafo remonte à l'aube du XXI^{e} siècle, sous l'impulsion de Belinda Maia, dans le cadre du jeune projet de la Linguateca, dont l'objectif principal est la mobilisation de ressources linguistiques et la conception d'outils de traitement automatique de la langue portugaise pour assurer sa pérennité parmi les langues informatisées et numérisées [2]. Le poloCLUP (pôle de Porto de la Linguateca) a entamé ses travaux en octobre 2002. L'équipe responsable de la conception et de l'adaptation du Corpógrafo aux besoins de ses usagers depuis 2002 est dirigée par Belinda Maia et Diana Santos, qui symbolisent la collaboration entre experts en Linguistique et en Génie informatique nécessaire à la création et au développement de ressources pour le traitement automatique des langues (TAL). L'évolution du Corpógrafo de 2003 à nos jours se divise en trois phases. La première, que l'on pourrait caractériser de préparatoire, correspond au Gestor de Corpora. Naît ensuite le Corpógrafo qui va évoluer de la version 1 à la version 3 de 2004 à 2007 essentiellement grâce aux efforts de Luís Sarmento, Luís Miguel Cabral, Ana Sofia Pinto et Débora Oliveira. La troisième phase correspond à la version 4 que l'on doit à Sérgio Matos sous la direction de Belinda Maia et Luís Costa. Le poloCLUP s'est éteint en 2008 et le bon fonctionnement du Corpógrafo dépend désormais de la bonne volonté de l'équipe composée de Belinda Maia, Diana Santos, Sérgio Matos et Luís Miguel Cabral.

Le Gestor de Corpora (GC) a été présenté à Lancaster en mars 2003 (Maia & Sarmento 2003) et à Braga en juin 2003 (Sarmento & Maia 2003), soit à peine quelques mois après le lancement du projet. La volonté de créer un outil répondant aux besoins d'enseignants et d'étudiants de trois domaines principaux (linguistique, traduction et TAL) et d'impliquer les utilisateurs dans l'évolution du Corpógrafo a déterminé le choix d'une architecture modulaire extensible et adaptable aux besoins formulés par les utilisateurs. Le GC permettait à l'utilisateur de créer un compte privé sur le Web où il pouvait stocker des fichiers PDF, PS, HTML, RTF ou MsWord convertis en texte grâce au module Perl EXTEX, révolutionnaire à l'époque. D'autres modules lui permettaient d'effectuer d'autres opérations : (1)

2. Le pari semble gagné : «De acordo com estimativas recentes, o português é a quinta língua mais usada na internet, sendo ultraprassada apenas pelo inglês, chinês, espanhol e japonês» (Branco et al. 2012, pg. 14).

éditer le texte pour le «nettoyer» et le diviser de façon semi-automatique en segments, conformément aux besoins de l'utilisateur ; (2) constituer un ou plusieurs corpus à partir de sélections de fichiers ; (3) réaliser des études de fréquence et des recherches de collocations (Sarmento & Maia 2003, pg. 27).

En 2004, grâce à son architecture modulaire, des outils d'extraction et de gestion terminologique ont pu être ajoutés aux fonctions du GC pour faciliter le travail du terminographe : le Corpógrafo était né (Sarmento et al. 2004). La structure actuelle du Corpógrafo se mettait en place. Le menu principal (figure 1) offre aujourd'hui quatre options, dont les deux premières sont héritées du GC : (1) «Gestor» (Gestionnaire) pour la création et la gestion de corpus ; (2) «Pesquisa» (Recherche) pour l'analyse de corpus selon divers types de requêtes ; (3) «Centro de Conhecimento» (Centre de connaissance) pour la création et la gestion de bases de données et de relations sémantiques ; (4) «Centro de comunicação» (Centre de communication) où l'utilisateur peut trouver des informations sur le Corpógrafo. Outre ces quatre options du menu principal donnant accès à diverses fonctions, l'utilisateur dispose de quatre boutons (figure 2) lui permettant (1) d'accéder à la corbeille de fichiers supprimés ; (2) d'obtenir de l'aide sur la fonction qu'il est en train d'utiliser ; (3) d'envoyer des commentaires ; (4) d'éditer son profil.

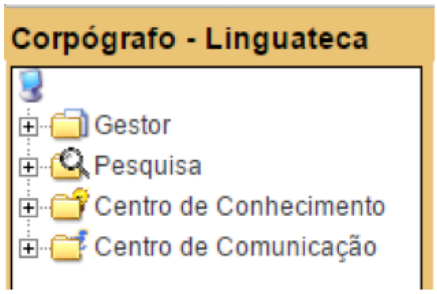


FIGURE 1 : Menu principal (Version 4).

FIGURE 2 : Boutons (Version 4).

Par rapport au GC, les fonctions de gestion de fichiers et de corpus de la première version du Corpógrafo ont été améliorées. L'utilisateur dispose d'un espace privé de 10 MB sur le serveur, non seulement pour stocker des fichiers et créer des corpus en recourant aux fonctions héritées du GC, mais aussi pour télécharger des sites à partir d'une adresse URL ou les explorer avant de sélectionner et de télécharger uniquement ce qui l'intéresse. Chaque fichier peut désormais être accompagné de métadonnées (titre, auteur, date, domaine de spécialité, type de texte, etc.). Certaines de ces métadonnées apparaissent automatiquement lors de certaines opérations d'analyse de corpus ou de constitution de base de données et elles permettent de classer les fichiers pour des recherches ultérieures. Chaque corpus est composé d'une sélection de fichiers qui peut être altérée à tout moment et le même fichier peut être intégré à plusieurs corpus. La figure 3 illustre les différentes options du gestionnaire.

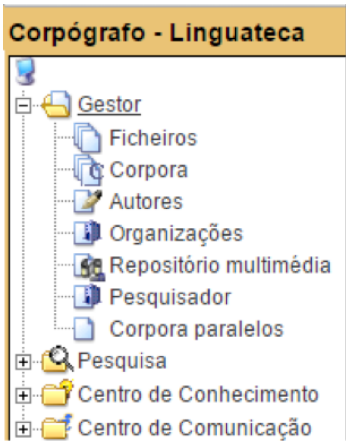
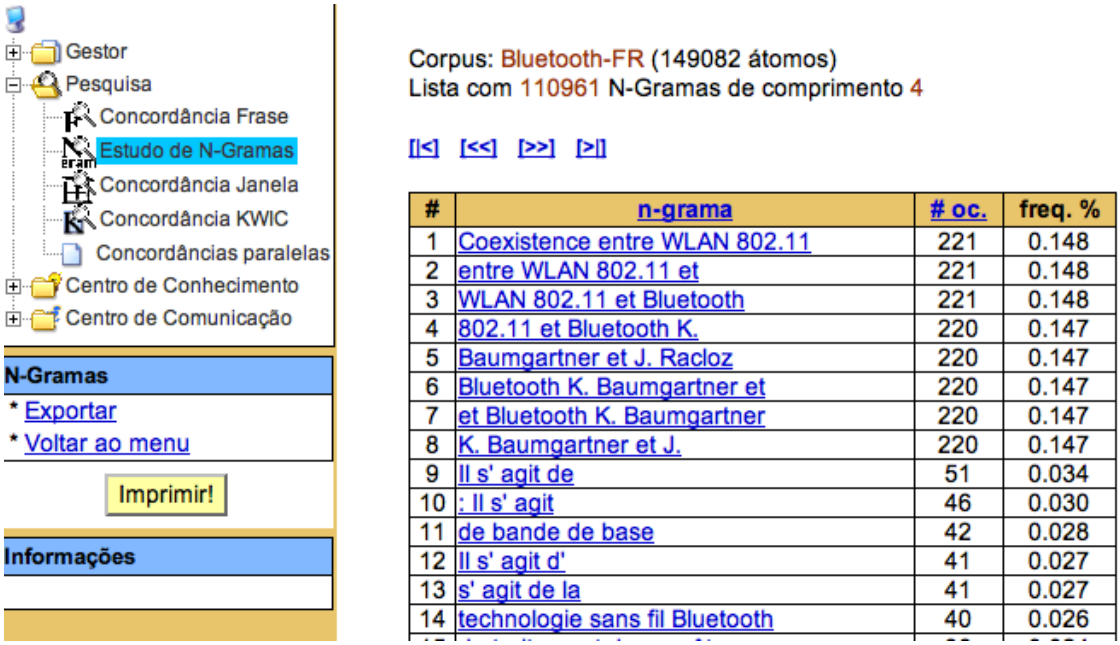


FIGURE 3 : Options du gestionnaire (Version 4)

L'option de recherche permet deux types de requête : l'une sur les n-grammes (co-occurrences d'un nombre paramétrable 'n' d'atomes), l'autre sur un mot, tel qu'un terme simple comme *piconet*, ou sur une séquence plus ou moins figée, telle qu'un terme complexe comme *access request address*. Dans le premier cas, l'utilisateur peut déterminer la longueur des n-grammes de 1 à 6 atomes (mots et signes de ponctuation). L'analyse de n-grammes permet d'obtenir des résultats absolus et relatifs sur la distribution des n-grammes dans le corpus analysé. Par exemple, une requête sur les 4-grammes du corpus Bluetooth en français (composé de 18 fichiers) nous apprend que le terme *technologie sans fil Bluetooth* y est attesté 40 fois, ce qui représente une fréquence de 0,026% [3] de l'ensemble des 4-grammes attestés dan ce corpus (figure 4). En cliquant sur ce 4-grammes, on apprend qu'il n'est utilisé que dans 4 des 18 fichiers, ce qui implique qu'il existe des variantes de ce terme central (à savoir *technologie Bluetooth* ou, tout simplement, *Bluetooth*). Les résultats peuvent être exportés vers un fichier CVS qui peut être lu par un tableur du type Excel.

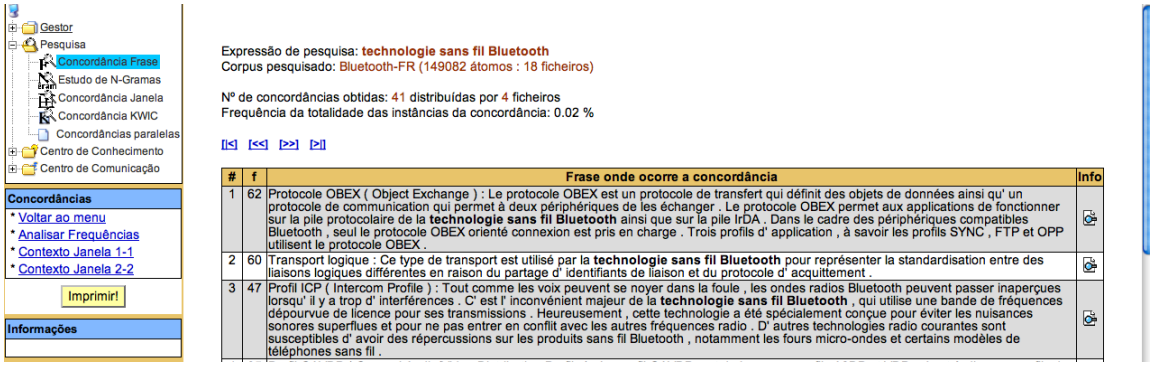
Les résultats des requêtes sur des expressions simples ou complexes peuvent être obtenus sous différentes formes selon qu'on veuille les analyser en contexte («Concordância Frase») ou découvrir des collocations ou toute autre forme de co-occurrence («Concordância Janela» (Concordance Fenêtre) et «Concordância KWIC») (Sarmento et al. 2004, pg. 451). Les options «Concordância Frase» et «Concordância KWIC [4]» (figures 5 et 6) donnent des résultats absolus et relatifs sur la distribution de l'expression de requête dans les divers fichiers du corpus et permettent d'accéder aux métadonnées du fichier de chaque occurrence (bouton «info»).

3. D'après nos calculs, cette fréquence est même plus élevée : $4000/110.961 = 0,36\%$.
4. Key Word In Context.



#	n-grama	# oc.	freq. %
1	Coexistence entre WLAN 802.11	221	0.148
2	entre WLAN 802.11 et	221	0.148
3	WLAN 802.11 et Bluetooth	221	0.148
4	802.11 et Bluetooth K.	220	0.147
5	Baumgartner et J. Racloz	220	0.147
6	Bluetooth K. Baumgartner et	220	0.147
7	et Bluetooth K. Baumgartner	220	0.147
8	K. Baumgartner et J.	220	0.147
9	Il s' agit de	51	0.034
10	: Il s' agit	46	0.030
11	de bande de base	42	0.028
12	Il s' agit d'	41	0.027
13	s' agit de la	41	0.027
14	technologie sans fil Bluetooth	40	0.026

FIGURE 4 : Requête sur n-grammes (Version 4)



#	f	Frase onde ocorre a concordância	Info
1	62	Protocole OBEX (Object Exchange) : Le protocole OBEX est un protocole de transfert qui définit des objets de données ainsi qu' un protocole de communication qui permet à deux périphériques de les échanger . Le protocole OBEX permet aux applications de fonctionner sur la pile protocolaire de la **technologie sans fil Bluetooth** ainsi que sur la pile IrDA . Dans le cadre des périphériques compatibles Bluetooth , seul le protocole OBEX orienté connexion est pris en charge . Trois profils d' application , à savoir les profils SYNC , FTP et OPP utilisent le protocole OBEX .	
2	60	Transport logique : Ce type de transport est utilisé par la **technologie sans fil Bluetooth** pour représenter la standardisation entre des liaisons logiques différentes en raison du partage d' identifiants de liaison et du protocole d' acquittement .	
3	47	Profil ICP (Intercom Profile) : Tout comme les voix peuvent se noyer dans la foule , les ondes radios Bluetooth peuvent passer inaperçues lorsqu' il y a trop d' interférences . C' est l' inconvénient majeur de la **technologie sans fil Bluetooth** , qui utilise une bande de fréquences dépourvue de licence pour ses transmissions . Heureusement , cette technologie a été spécialement conçue pour éviter les nuisances sonores superflues et pour ne pas entrer en conflit avec les autres fréquences radio . D' autres technologies radio courantes sont susceptibles d' avoir des répercussions sur les produits sans fil Bluetooth , notamment les fours micro-ondes et certains modèles de téléphones sans fil .	

FIGURE 5 : «Concordância Frase» (Version 4)

Quant à l'option «Concordância Janela» (figure 7), elle permet de classer les résultats par ordre alphabétique d'après les atomes qui précèdent ou qui suivent l'expression de requête. La version 4 comporte une cinquième fonction permettant d'explorer des corpus parallèles, mais nous ne l'avons jamais utilisée.

Mais ce qui a valu le changement de nom du GC, c'est bien la possibilité de créer et de gérer des BDT. L'extraction terminologique se fonde sur l'analyse de n-grammes et un ensemble de restrictions lexicales à partir d'un dictionnaire électronique. L'utilisateur peut créer des BDT dont le modèle se base sur la norme ISO 12620. Chaque «terme vedette» ou «entrée» de la BDT correspond à une fiche comportant divers champs pouvant être complétés ou non (langue, données morphologiques, source, définition, exemples en contexte, relations sémantiques avec

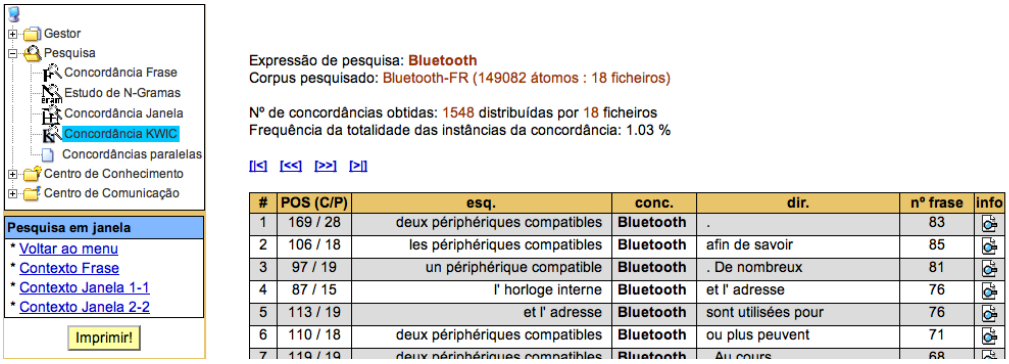


FIGURE 6 : «Concordância Frase» (Version 4)

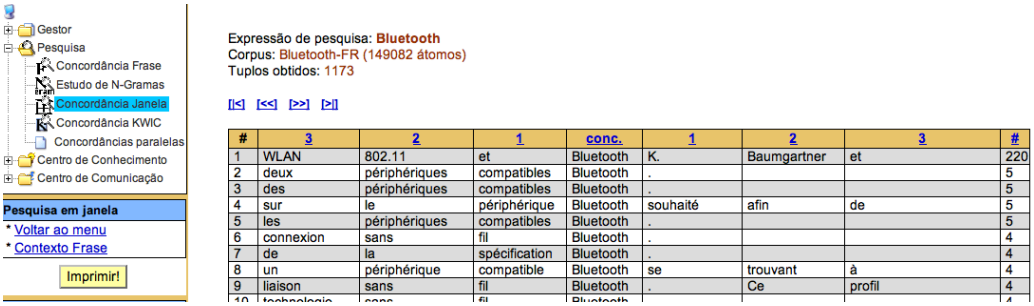


FIGURE 7 : «Concordância Frase» (Version 4)

d'autres termes, équivalent(s) dans une ou plusieurs autres langues, etc.). À partir de ces données, le Corpógrafo crée automatiquement des glossaires au format HTML. Il peut aussi produire des thesauri au format HTML, en se fondant sur les relations sémantiques entre termes, et exporter ces informations vers un fichier dont le format permet de visualiser le réseau sémantique ainsi produit (*ibidem*).

La version 2 est présentée l'année suivante à l'Université de Leeds (Maia & Sarmento 2005). De la mi-mai 2004 à la mi-mai 2005 (Santos 2005, pgs. 19–20), les efforts de Luís Sarmento se sont concentrés sur l'amélioration de la gestion terminologique (extraction semi-automatique de termes et de relations sémantiques, production de cartes conceptuelles à partir des relations sémantiques) et ceux de Luís Cabral sur l'interface SAGI, d'après une enquête auprès des utilisateurs, tandis que Débora Oliveira s'est chargée de l'adaptation de la documentation aux nouvelles fonctions du Corpógrafo.

La division du menu principal en quatre options est présentée pour la première fois en 2006, alors que le Corpógrafo en est déjà à sa troisième version (Maia & Sarmento 2006, pg. 55). Si les fonctions de gestion et de recherche ont légèrement évolué lors de cette deuxième phase, les efforts se sont toutefois concentrés sur la semi-automatisation de l'extraction terminologique, de l'extraction de définitions et de l'identification de relations sémantiques (Sarmento et al. 2006, pgs. 1503–1504).

La fonction d'extraction semi-automatique des termes existait déjà dans la première version et a été décrite ci-dessus. Son avantage, c'est que le corpus n'a pas besoin d'être annoté. Cette fonction est plus performante sur de petits corpus hautement terminogènes que sur de grands corpus en raison du bruit important que ceux-ci produisent malgré les filtres mis en place en anglais et en portugais.

Mais les progrès de la première à la troisième version sont plus sensibles au niveau de l'extraction semi-automatique de définitions et de l'identification semi-automatique de relations sémantiques qui étaient absentes de la version 1. Ces deux modules fonctionnent sur des stocks de structures [5] propres à introduire une définition ou une relation sémantique. Les stocks de structures «définitionnelles» sont assez importants pour l'anglais et le portugais (plus de 120 structures dans chaque langue). Cette fonction est plus efficace lorsque le corpus est composé de textes didactiques généralement riches en définitions. L'identification des relations sémantiques est plus complexe, non seulement parce qu'elles sont nombreuses et variées, mais aussi parce qu'elles sont souvent implicites (idem, pg. 1504). Les relations sémantiques peuvent également être établies manuellement et l'utilisateur peut créer d'autres relations que celles qui sont pré-définies selon ses besoins. La volonté de produire des ontologies et des thesauri de domaine de spécialité à partir de ces relations sémantiques était louable mais les résultats obtenus se sont révélés peu probants et la recherche sur ces fonctions ont été abandonnées.

La quatrième version du Corpógrafo se caractérise surtout par l'intégration du moteur Nooj à plusieurs fonctions de recherche du Corpógrafo pour repérer des unités phraséologiques en anglais, en français et en portugais, et par l'insertion de fonctions permettant d'aligner des corpus parallèles et des segments de corpus comparables, de lancer des recherches sur ces corpus alignés et de créer des bases de données phraséologiques ou lexicales (Maia & Matos 2008, pgs. 80–81). La longueur des n-grammes peut désormais aller jusqu'à 15-grammes, la possibilité d'exporter les bases de données a été optimisée et il est possible d'obtenir des statistiques sur la fréquence des termes ou des éléments lexicaux à partir des bases de données ou de visualiser les relations sémantiques à partir de la fiche d'un terme (figure 8).

Le Corpógrafo a ainsi évolué selon les objectifs de ses concepteurs et les besoins de ses utilisateurs. Aujourd'hui, il remplit la mission qui lui a été assignée : c'est un outil conçu prioritairement pour le traitement automatique du portugais accessible gratuitement à tout chercheur dont le travail comporte des recherches sur corpus, quel que soit son domaine et quelles que soient ses compétences informatiques. Les deux expériences décrites ci-dessous font apparaître certaines de ses failles et certains de ses atouts dans la perspective de l'utilisateur.

5. Des exemples de ces deux types de structures sont fournis par Sarmento et al. (2006, pg. 1503).

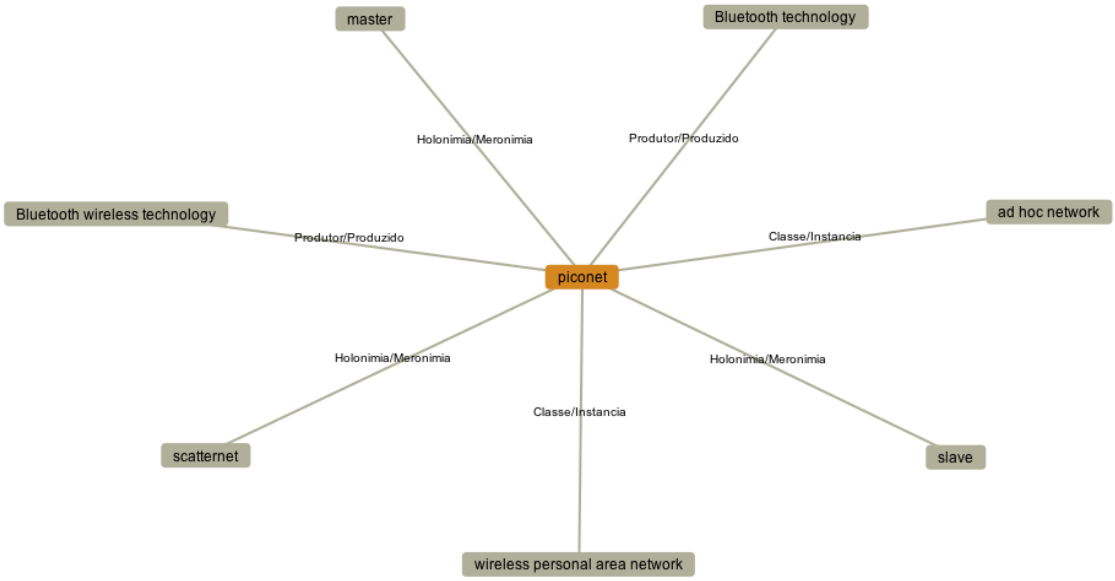


FIGURE 8 : Relations sémantiques (Version 4)

[3] TERMINOLOGIE

Pour réaliser l'étude de néonymie comparée de la terminologie Bluetooth en anglais, en français et en portugais (Bacquelaine 2009) [6], nous avons utilisé plusieurs fonctions des options «Gestor», «Pesquisa» et «Centro de Conhecimento» pour constituer une BDT à partir de corpus comparables sur mesure, pour obtenir des statistiques sur les termes et pour exporter la BDT vers un fichier consultable en dehors de l'outil [7].

Toute recherche commence évidemment par la création d'un ou plusieurs corpus à partir de fichiers [8] téléchargés au moyen des fonctions du «Gestor». La plupart des formats des fichiers récoltés étaient convertibles par EXTEX. Seules la présentation PPT de Schiller (2008) [9] et la leçon du professeur Nuno Almeida Almeida (2007) filmée le 29 mai 2007 ont été exploitées manuellement. La préparation des fichiers s'est limitée au strict minimum. Le Corpógrafo se fonde sur le point final pour segmenter le texte en phrases. Ainsi, les termes *IEEE 802.15* et *IEEE 802.15.1* ont été séparés automatiquement en deux ou trois segments qui ont dû être réunis. Nous avons également décidé de rassembler plusieurs phrases en un seul segment pour obtenir des définitions et des contextes comportant tous les

6. Une version mise à jour est actuellement en cours de publication (Bacquelaine 2015).

7. Ce fichier «BDT Bluetooth» peut être consulté à http://web.letras.up.pt/franba/BDT-Bluetooth/CGshirleyBluetooth.html

8. Les corpus comparables sont décrits dans notre mémoire (idem, pgs. 66–71).

9. Cette version consultée le 15/09/2008 n'est plus disponible. Dans les références, nous indiquons donc l'adresse url de la nouvelle version non datée (Schiller (s/d)).

éléments nécessaires à leur compréhension, notamment les antécédents des pronoms. Le nettoyage n'a pas besoin d'être parfait car on peut corriger les extraits dans la fonction BDT. Il faut toutefois veiller à ce que les termes soient orthographiés correctement pour être reconnus par les différentes fonctions. Quatre corpus comparables (anglais, français, portugais européen et portugais du Brésil) ont été constitués à partir des fichiers contenant les documents «nettoyés». Ces corpus ont ensuite été exploités grâce aux fonctions des options «Pesquisa» et «Centro de Conhecimento».

Les différentes fonctions de recherche permettent de limiter la recherche à un corpus sélectionné dans un menu déroulant. Elles fonctionnent très bien quelle que soit la langue. Nous avons utilisé trois des cinq fonctions de recherche : «Concordância Frase», «Concordância Janela» et «Concordância KWIC». La première a permis de trouver des définitions et des contextes pour les sigles de moins de quatre lettres qui ne sont pas reconnus par les fonctions de l'outil BDT. En effet, cela en ralentirait les performances, ce qui est contraire à l'intérêt de la majorité des utilisateurs. Ces trois fonctions ont été très utiles pour repérer les variantes à partir du co-texte et les termes composés à partir de noyaux terminologiques tels que *protocol*, *layer*, *link*, *channel* ou *logical transport*.

L'outil BDT permet d'extraire des candidats terminologiques à partir d'un corpus sélectionné. Au cours de la deuxième phase de développement du Corpógrafo, des filtres ont été mis en place en anglais et en portugais pour diminuer le bruit causé par la ponctuation, les pronoms, les prépositions, les auxiliaires, les déterminants, etc. Une option permet à l'usager d'accéder d'un simple clic au contexte et aux références du fichier d'origine avant de sélectionner le candidat. Cette sélection entraîne automatiquement la création de la fiche correspondante comportant plusieurs données insérées automatiquement : la langue et les références du fichier d'origine du terme. L'extraction terminologique semi-automatique fonctionne mieux en anglais et en portugais qu'en français, mais le bruit reste important étant donné le volume des corpus. Toutefois, la liste des termes anglais à inclure dans l'échantillon a été établie en concertation avec le professeur Almeida à partir de sa leçon filmée, du manuel de Schiller (2003) et de sa présentation PPT en ligne (2008). En tout, 35 termes EN sur 122, 47 termes PT sur 146 et 5 termes FR sur 205, soit un peu plus de 18% des termes, ont été insérés semi-automatiquement. Les autres fiches ont été créées au fur et à mesure des besoins.

La fiche terminologique du Corpógrafo propose dix champs principaux : «Dados Gerais» (Données générales), «Pesquisadores» (Chercheurs), «Autores», «Fontes» (Sources), «Morfologia», «Definições», «Contextos», «Relações Semânticas», «Termos Relacionados» et «Equivalentes de Tradução». Nous les avons complétés tous sauf le champ «Pesquisadores» puisqu'il n'est nécessaire que lorsque plusieurs chercheurs travaillent sur la même BDT.

Le champ «Dados Gerais» contient le terme vedette et ses principales caractéristiques : langue, type (sigle, abréviation, etc.), statut (normalisé, admis, etc.), registre (courant, technique, etc.), fréquence d'emploi et origine (emprunt, néologisme, etc.).

Les champs «Autores» et «Fontes» identifient, d'une part, les auteurs des fichiers d'où proviennent les termes vedettes, les définitions et/ou les contextes, et, d'autre part, les entités publiques ou privées dont ces auteurs relèvent. Ces informations apparaissent automatiquement si la fonction d'extraction terminologique semi-automatique a été appliquée, mais elle peuvent aussi être insérées manuellement à partir de menus déroulants des listes d'auteurs et d'entités enregistrés lors de la première étape grâce aux fonctions du «Gestor».

La conception du champ «Morfologia» semble bien reposer sur l'assomption que la plupart des termes appartiennent à la classe des noms et seuls le genre et le nombre du terme peuvent être définis par le terminographe. Certains domaines techniques tels que le tricot ou le crochet [10] présentent pourtant beaucoup de verbes qui sont des termes et la terminologie Bluetooth comporte plusieurs adjectifs et de nombreux sigles, qui correspondent, certes, à des entités nominales, mais dont certains combinent les lettres aux chiffres et parfois même à la ponctuation (*L2CAP*, *IEEE 802.15*). Les termes complexes sont segmentés automatiquement (mais pas les sigles) et la classe grammaticale de chaque élément qui le compose peut être sélectionnée à partir d'un menu déroulant qui propose les options NC (nom commun), NP (nom propre), AJ (adjectif), VB (verbe), PP (préposition) et AD (adverbe). Ce système fermé limite les possibilités de classement et la segmentation du terme est imparfaite et ne peut être améliorée. Par exemple, la contraction de la préposition et de l'article défini, en français et en portugais, et l'article défini singulier ou la préposition *de* élidés devant un nom commençant par une voyelle en français sont considérés comme un seul mot (et donc un seul atome), ce qui pose des problèmes de classement. Le pronom latin *hoc* dans *réseau ad hoc*, les articles et les conjonctions — tels que les articles définis et la conjonction *et* dans *interface entre l'hôte et le contrôleur* — ne peuvent être classés. Il est vrai que les articles et les conjonctions sont plutôt rares dans les terminologies et que cet outil a été programmé pour l'anglais et le portugais. L'apostrophe a très peu de chance d'être employée dans les terminologies anglaises et elle est très rare en portugais. Le problème ne se pose que pour les termes complexes de plus en plus souvent représentés par des sigles [11]. Une solution pourrait être de les classer comme un tout, syntagme nominal, verbal, adjectival ou adverbial et des traits morphologiques d'autres classes de mots devraient pouvoir figurer dans ce champ.

10. Notre mémoire de Licence en Philologie germanique s'intitule «Deutsch-französische Terminologie des Strickens und Häkelns» (Bacquelaine 1980) et la plupart des termes sélectionnés pour l'analyse sont des verbes à particule séparable ou inséparable.

11. Sablayrolles parle même de «siglomanie» en néologie (2000, pg. 263).

Les deux champs suivants contiennent la ou les définition(s) et le ou les contexte(s) d'emploi permettant de repérer les collocations et autres phraséologismes propres au terme et au domaine. La plupart des définitions et des contextes ont été extraits automatiquement des corpus préparés à cet effet. Quelques définitions portugaises ont été transcrites du document audio-visuel et quelques anglaises du livre de Schiller (2003). D'autres ont été reformulées à partir de plusieurs sources.

Etant donné la complexité des relations sémantiques dans ce domaine, nous n'avons pas établi systématiquement les relations sémantiques entre termes dans la BDT. Nous avons préféré construire trois micro-structures à partir de la documentation et des entretiens avec l'expert : les réseaux ad hoc Bluetooth (Bacquelaine 2009, pg. 29), les modes, états et adresses des appareils compatibles Bluetooth (idem, pg. 31) et le système principal Bluetooth (idem, pg. 77). Ces trois micro-structures donnent une idée des relations entre la plupart des noyaux conceptuels désignés par les termes de la BDT. Par contre, les relations de synonymie («Termos relacionados») ont été établies systématiquement, car elles permettent de déterminer le nombre de concurrents pour le même concept. Quelques rares cas d'antonymie ont également été signalés.

Enfin, le dernier champ contient les équivalents de traduction que l'on sélectionne par langue dans un menu déroulant. Cette fonction a été aménagée depuis. En effet, le menu déroulant s'allongeait au fur et à mesure que la BDT s'enrichissait et il ne comporte plus désormais que les initiales — majuscules ou minuscules — des termes enregistrés dans la BDT. Cet aménagement parésente l'avantage de raccourcir le menu déroulant mais aussi l'inconvénient d'occulter les termes : le terminographe ne peut plus choisir le terme dans le menu, il doit savoir ce qu'il cherche pour pouvoir le trouver. Il faut donc que la fiche de l'équivalent ait été créée préalablement et notre organisation en trois étapes, l'anglais, puis le français, puis le portugais, s'est révélée très pratique. On peut aussi passer d'une fiche à l'autre grâce à des hyperliens entre synonymes, antonymes et équivalents de traduction. Cette fonction ajoutée en 2006 à la demande des utilisateurs se révèle très utile pour vérifier si aucun terme concurrent n'a été oublié.

Une autre fonction très utile de la BDT est celle qui permet d'obtenir des statistiques sur chaque terme. Elle distingue non seulement le nombre d'occurrences au singulier et au pluriel, sauf en français, mais aussi le nombre d'occurrences par fichier. Ces données permettent de comparer l'usage selon les auteurs, les types de texte ou les registres. Il est aussi possible d'obtenir des statistiques générales par langue et par corpus. Ces dernières ne tiennent compte que des termes de plus de trois lettres extraits et insérés semi-automatiquement, si bien que nous n'avons pu utiliser ces résultats efficacement en raison des nombreux sigles de trois lettres et des nombreux termes (81,6% du total) insérés manuellement.

On peut aussi associer à chaque fiche un ou plusieurs médias (images, films ou enregistrements sonores numérisés), mais quelques problèmes doivent encore

être résolus. D'une part, aucun champ n'est prévu pour indiquer la présence de ces fichiers, d'autre part, ils n'ont pas été exportés avec les autres données de la BDT.

Dans l'ensemble, cette expérience terminographique a été très positive. Si les performances des fonctions de gestion et de recherche sont remarquables, les fonctions d'extraction semi-automatique de termes et de définitions peuvent être améliorées, notamment en français, mais elles ont quand même facilité la tâche terminographique. Certaines contraintes, telles le nombre minimum de quatre lettres par terme pour obtenir des résultats statistiques ou les options réduites de classement morphologique, devraient pouvoir être levées ou aménagées par l'utilisateur, comme c'est le cas des relations sémantiques. Étant donné que tous les utilisateurs n'ont pas besoin de tous les champs prévus pour les fiches terminologiques, ceux-ci pourraient être activés selon les besoins de chacun. Les corpus FR et PE (portugais européen) créés pour cette première expérience ont été réutilisés dans la deuxième. Cette possibilité de «recyclage» représente un autre avantage du Corpógrafo.

[4] PHRASÉOLOGIE

Pour comparer l'usage des quantificateurs universels *chaque/cada*, *tout/todo o* et *tous les/todos os* en français et en portugais européen dans le cadre du séminaire de Sémantique du Doctorat en Linguistique de la FLUP (1er semestre 2009–2010), nous avons analysé trois corpus dans chaque langue au moyen des fonctions «Concordância Frase» et «Concordância Janela». Les premiers corpus comparables comportent 140 224 atomes en portugais et 110 928 en français. Ils se composent d'articles de presse sur la deuxième guerre en Irak extraits de divers journaux portugais, belges et français accessibles gratuitement en ligne. Ces extraits ont été récoltés à partir de mots clés (*guerre*, *Irak*, date, euphémismes courants liés à la guerre et à la mort) sur deux périodes, la première entre le 18 et le 24 mars 2003 (début de la deuxième guerre) et la seconde entre le 18 et le 24 mars 2006 (troisième anniversaire du début de la deuxième guerre). En effet, ces corpus avaient été constitués pour étudier les obstacles à la traduction des euphémismes (Bacquelaine 2006) et ont été réutilisés en tant que corpus représentatifs du registre courant en français et en portugais. Toutefois, deux corpus scientifiques ont été ajoutés dans chaque langue, car les quantificateurs *tout* et surtout *todo o* sont plutôt rares dans le registre courant. Nous avons ainsi réutilisé les corpus comparables sur les télécommunications sans fil (352 813 atomes en portugais et 149 082 en français) et créé un nouveau corpus à partir d'un texte de départ dans le domaine du droit constitutionnel comptant 67 463 atomes en portugais et de sa traduction remaniée qui compte 81 490 atomes en français. Ces deux derniers corpus sont ainsi comparables et parallèles à la fois, mais ils ont tous été traités de la même façon.

Les attestations des trois quantificateurs, au féminin et au masculin, en français et en portugais, ont été extraites grâce à la fonction «Concordância Frase». Ces résultats bruts ont été copiés-collés sur une feuille de calcul Excel où ont été réalisées les opérations de sélection des segments pertinents [12] et de classement de ceux-ci d'après les noms sur lesquels ils opèrent en vue de l'analyse qualitative et quantitative de ces données. Contrairement aux attentes des locuteurs natifs lusophones à qui ces résultats ont été présentés [13], il s'est avéré que, dans l'absolu, *cada* est plus fréquent que *chaque*, quel que soit le co-texte ou le registre (courant, juridique ou technico-scientifique).

La fonction «Concordância Janela» permet de classer les résultats par ordre alphabétique d'après le co-texte, c'est-à-dire les mots qui précèdent ou suivent l'expression de requête (la ponctuation n'a évidemment aucun intérêt). Cette fonction a ainsi révélé les affinités particulières de chaque quantificateur avec certains noms. Par exemple, l'affinité de *cada* avec le nom *vez* est très forte. Elle a aussi mis en évidence la particularité de *cada* qui peut opérer sur un nom quantifié par un numéral cardinal supérieur à l'unité, ce que ni *chaque* ni *each* ne peuvent faire. Ces découvertes ont ainsi déterminé le choix de l'objet d'étude de notre doctorat en cours. En effet, la fréquence de *cada* — particulièrement élevée dans le corpus Bluetooth — s'explique en partie par la fréquence de quatre séquences semi-figées dont les traductions [14] en anglais et en français ne comportent ni *each* ni *chaque* et qui s'assimilent à la phraséologie au sens large. Les exemples (1) à (4) illustrent ces quatre séquences :

(1) Estima-se que as organizações (...) procurarão usar métodos cada vez menos convencionais e mais inesperados. (corpus Irak)

(2) Um dispositivo pode fazer parte de diferentes piconets ; porém, como as unidades de rádio só podem sintonizar uma das portadoras em cada instante, ele só pode comunicar com uma piconet de cada vez. (corpus Bluetooth)

(3) ...Blair (...) resisitiu a uma prova de fogo, ao conseguir que a maioria dos deputados desse o seu apoio à partipação britânica na guerra apesar de um em cada três parlamentares trabalhistas terem votado contra. (corpus Irak)

12. Les résultats de cette recherche ont été présentés au Colloque international «Traduction, terminologie et rédaction technique : des ponts entre le français et le portugais» en janvier 2011 et l'article «Apports de la sémantique et de la syntaxe à la traduction des quantificateurs universels français et portugais» a été accepté en juillet 2011 pour publication dans les Actes, qui se font toujours attendre à ce jour.

13. Il s'agit du groupe de sémantique du CLUP dirigé par le professeur Fátima Oliveira, composé notamment de Fátima Silva, Luís Filipe Cunha, António Leal, Purificação Silvano, Idalina Ferreira et Joaquim Barbosa.

14. Les équivalents de traduction ont été confirmés par une étude postérieure de corpus parallèles disponibles en ligne.

(4) Como o intervalo de geração de células é menor, a atribuição de um canal LCH por cada 3 tramas deixa de ser suficiente para satisfazer o requisito da conexão... (corpus Bluetooth)

En (1), *cada vez* se combine à un comparatif (*mais*, *menos*, *maior*, *menor*, *melhor* ou *pior* [15]) pour exprimer la progression dans un sens ou dans l'autre. La progression s'exprime par d'autres moyens en anglais (p. ex. *more and more*, *ever more*, ou la lexicalisation du concept en recourant à *to increase*, *increasing* ou *increasingly* pour exprimer l'augmentation quantitative ou l'intensification qualitative) et en français (p. ex. *de plus en plus*, *de moins en moins*, *de mieux en mieux*, *de mal en pis* ou diverses lexicalisations du concept telles que se *multiplier* ou *croissant*). Nous avons baptisé «compte-gouttes» la relation exprimée en (2), où *uma* correspond à un numéral cardinal restreignant à l'unité la quantité de *piconets* où l'appareil Bluetooth peut communiquer *de cada vez* (*at a time* et *à la fois* sont les équivalents les plus fréquents). Les exemples (3) et (4) illustrent les deux derniers types de relations quantifiées par *cada* en portugais. En (3), *um em cada três parlamentares trabalhistas* (*one in (every) three Labour MEPs* et *un député travailliste sur trois*) exprime une proportion entre un ensemble et un sous-ensemble tandis qu'en (4), *um canal LCH por cada 3 tramas* (*one LHC channel for every three frames* et *un canal LHC pour trois trames*) exprime une proportion entre deux ensembles distincts.

Les relations de proportion sont les plus complexes et les prépositions *em* ou *por* peuvent entraîner l'un ou l'autre type de proportion, mais il ne s'agit pas ici d'entrer dans les détails de l'expression de ces quatre relations quantifiées par *cada*, mais bien de démontrer la performance et l'utilité des fonctions de recherche du Corpógrafo qui dévoilent des aspects insoupçonnés des langues naturelles.

[5] CONCLUSION

Ces deux expériences démontrent incontestablement les atouts du Corpógrafo et les performances exceptionnelles des fonctions de gestion et de recherche, malgré ses défauts inévitables. Les besoins des utilisateurs ne sont pas uniformes et certains aménagements réalisés pour satisfaire les besoins des uns compliquent la tâche des autres. Certaines contraintes peuvent être gênantes. Il a été question de la limitation du menu de classement du champ «Morfologia» de la BDT, de la longueur minimale des termes fixée à quatre lettres pour améliorer les performances de l'outil alors que les sigles de deux ou trois lettres se multiplient dans la plupart des terminologies de pointe, des statistiques sur chaque terme qui ignorent également les termes de moins de quatre lettres et enfin des statistiques sur la BDT qui ne tiennent compte que des termes insérés de façon semi-automatique. La fonction BDT a été conçue en conformité avec la norme ISO 12620, mais, comme

15. *Inferior* et *superior* sont aussi possibles, mais ne sont pas attestés dans les corpus utilisés.

le fait remarquer Gouadec (2003), l'utilisateur préfère des répertoires terminologiques aussi simples que possible et donc «en contravention totale avec toutes les règles de la terminographie» (16'18" – 16'20" [16]). Comme nous l'avons dit, la possibilité de sélectionner les champs des fiches terminologiques selon les besoins de chaque utilisateur permettrait d'améliorer la présentation de la BDT exportée. La gestion des médias associés devrait pouvoir être améliorée par le signalement de leur présence et leur inclusion lors de l'exportation.

Certes, des corpus «prêt-à-porter», comparables ou parallèles, sont disponibles et exploitables gratuitement en ligne, mais ils ne contiennent pas toujours ce dont on a besoin, notamment lorsqu'il s'agit de terminologie ou de phraséologie spécialisée. Certes, d'autres outils permettent d'explorer l'immense corpus du Web (par exemple, WebCorp [17] voire Google) ou de constituer des corpus à partir de mots-clés associés au domaine (par exemple, BootCat [18]) [19], mais le Corpógrafo est sans doute le seul outil gratuit accessible en ligne [20] ou téléchargeable spécialement conçu pour le traitement automatique du portugais et fonctionnant aussi pour d'autres langues en caractères romains. Il permet en outre d'inclure des fichiers non disponibles sur Internet, tels que ceux fournis par les experts de la FEUP. Son architecture modulaire lui confère une grande flexibilité et la possibilité de s'adapter aux besoins formulés par les utilisateurs. Cette flexibilité l'apparente à un laboratoire où des pistes sont suivies jusqu'au bout ou abandonnées en chemin si les résultats s'avèrent négatifs. Les fonctions de gestion de fichiers et de corpus ainsi que les fonctions de recherche sont très efficaces et ne requièrent que peu d'efforts de préparation des matières premières. La possibilité de rectifier le texte des définitions et des contextes dans les bases de données ainsi que celle de créer des relations sémantiques constituent d'autres atouts du Corpógrafo.

Initialement conçu comme un outil d'aide à la recherche et à la formation en linguistique portugaise et en terminographie au service de la traduction, il ne peut se mesurer à des fonctions intégrées aux outils d'aide à la traduction telles que Termbase ou Multiterm, qui sont beaucoup plus pratiques pour les traducteurs professionnels. Il n'en reste pas moins que c'est un outil didactique performant pour initier les étudiants aux canons de la terminographie et de la lexicographie. En outre, il se révèle un allié fidèle et utile pour toute recherche sur un ou plusieurs corpus. Les corpus « sur mesure » constituent ainsi un investissement à long terme permettant d'utiliser les mêmes corpus ou d'en créer d'autres à partir des mêmes fichiers pour réaliser toute sorte de recherches fondées sur le langage et toute sorte de comparaisons entre langues ou registres au sens large (oral ou écrit,

16. Il s'agit d'un document sonore.

17. À http://www.webcorp.org.uk/live/.

18. À http://bootcat.sslmit.unibo.it/.

19. Ces deux outils concurrents, Webcorp et Bootcat, nous ont été signalés par Sílvia Araújo que nous tenons à remercier pour sa révision minutieuse et ses conseils judicieux.

20. http://labclup.letras.up.pt/corpografo/

langue courante ou langue de spécialité, types de textes, etc.), quel que soit le domaine de recherche (Terminologie, Traduction, Linguistique, TAL, Sociologie, etc.). Depuis que le projet n'est plus financé, ce qui est regrettable, le Corpógrafo n'a plus fait l'objet que d'aménagements ponctuels et son bon fonctionnement dépend désormais de la bonne volonté de quelques-uns. Qu'ils en soient remerciés.

RÉFÉRENCES

Almeida, Nuno. 2007. A tecnologia Bluetooth.

Bacquelaine, Françoise. 1980. Deutsch-französische Terminologie des Strickens und Häkelns. Université de Liège, non publié.

Bacquelaine, Françoise. 2006. L'euphémisme, un obstacle à la traduction. *Revista da Faculdade de Letras : Línguas e Literaturas* XXIII. 463–487.

Bacquelaine, Françoise. 2009. *La terminologie Bluetooth en anglais, en français et en portugais. Étude de néonymie comparée*. Porto : Faculdade de Letras da Universidade do Porto. MA thesis. Version de septembre 2008 revue après soutenance.

Bacquelaine, Françoise. 2015. *La terminologie Bluetooth en anglais, en français et en portugais -- Étude de néonymie comparée*. Presses académiques francophones.

Branco, António, Amália Mendes, Sílvia Pereira, Paulo Henriques, Thomas Pellegrini, Hugo Meinedo, Isabel Trancoso, Paulo Quaresma, Vera Lúcia Strube de Lima & Fernanda Bacelar. 2012. *A língua portuguesa na era digital - The Portuguese Language in the Digital Age*. Springer.

Gouadec, Daniel. 2003. Terminologie et traduction. Document audio, communication et discussion : 1'05''- 32'30''. http://archives.diffusion.ens.fr/diffusion/audio/2003_10_17_terminologie_02.mp3.

Maia, Belinda & Sérgio Matos. 2008. Corpógrafo V4 - Tools for Researchers and Teachers using Comparable Corpora. In Pierre Zweigenbaum, Éric Gaussier & Pascale Fung (eds.), *LREC 2008 Workshop on Comparable Corpora (LREC 2008)*, 79–82. ELRA.

Maia, Belinda & Luís Sarmento. 2003. GC - Integrated Web Environment for Corpus Linguistics. Présentation à la *Corpus Linguistics 2003 (CL2003)*. http://www.linguateca.pt/documentos/cl2003.pdf.

Maia, Belinda & Luís Sarmento. 2005. The Corpógrafo - an Experiment in Designing a Research and Study Environment for Comparable Corpora Compilation and Terminology Extraction. In *Proceedings of eCoLoRe / MeLLANGE Workshop, Resources and Tools for e-Learning in Translation and Localisation*, 45–48.

Maia, Belinda & Luís Sarmento. 2006. Corpógrafo - Applications. In *Third International Workshop on Language Resources for Translation Work Research & Training, Satellite event of LREC 2006 (LR4Trans-III)*, 55–58.

Sablayrolles, Jean-François. 2000. *La néologie en français contemporain. examen du concept et analyse de productions néologiques récentes* (Lexica. Mots et Dictionnaire 4). Champion.

Santos, Diana. 2005. Relatório da Linguateca de 15 de Maio de 2004 a 14 de Maio de 2005. Tech. rep. Linguateca. http://www.linguateca.pt/documentos/RelatorioLinguatecaMaio2005.pdf.

Sarmento, Luís & Belinda Maia. 2003. Gestor de Corpora – Um ambiente Web integrado para Linguística baseada em Corpora. In José João Almeida (ed.), *Corpora Paralelos, Aplicações e Algoritmos Associados (CP3A)*, 25–30.

Sarmento, Luís, Belinda Maia & Diana Santos. 2004. The Corpógrafo - a Web-based environment for corpora research. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, 449–452.

Sarmento, Luís, Belinda Maia, Diana Santos, Ana Pinto & Luís Cabral. 2006. Corpógrafo V3 : From Terminological Aid to Semi-automatic Knowledge Engine. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik & Daniel Tapias (eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, 1502–1505.

Schiller, Jochen. 2003. *Mobile Communications*. Harlow (GB) : Pearson Education Limited 2nd edn.

Schiller, Jochen. 2008. Wireless LANs. Cette version consultée le 15/09/2008 n'est plus disponible ; une version remaniée non datée est disponible à https://www.iith.ac.in/~tbr/teaching/docs/wireless_lans.pdf.

CONTACTS

Françoise Bacquelaine
Faculdade de Letras, Universidade do Porto
franba@letras.up.pt