Akenou-Breizh, un projet de plate-forme valorisant des ressources et outils informatiques et linguistiques pour le breton

Annie Foret[1] Valérie Bellynck[2] Christian Boitet[2]
(1) IRISA & Université Rennes 1, Campus de Beaulieu, F-35042 Rennes cedex
(2) LIG-GETALP, UdG &CNRS, Campus, 41 rue des Mathématiques, GRENOBLE cedex 9
Annie.Foret@irisa.fr, Valerie.Bellynck@imag.fr, Christian.Boitet@imag.fr

Résumé. Nous présentons un nouveau projet, *Akenou-Breizh*, qui vise (1) à mettre en place une plate-forme permettant d'étudier les influences d'une *langue d'héritage*, comme le breton, sur une *langue d'usage*, comme le français, et (2) à mettre à disposition de tous les intéressés des outils s'intégrant au "Web sémantique et multilingue", et proposant des accès proactifs aux connaissances sur le breton ainsi qu'une visualisation directe des correspondances sous-phrastiques dans des présentations bilingues alignées. Nous nous proposons non seulement d'utiliser les nombreuses ressources disponibles librement, en particulier celles de l'OPLB [1] et du projet APERTIUM, mais aussi d'en créer de nouvelles, comme des corpus bilingues alignés de bonne qualité, en utilisant le "Web collaboratif", et de construire sur le site dédié `lingwarium.org` des modules linguistiques améliorant ou étendant ceux qui existent, par exemple un analyseur-générateur morphologique. Nous décrivons aussi une expérience réalisée à partir d'un lexique réduit pour le breton, qui montre comment on peut enrichir un dictionnaire classique, en le reliant à un treillis de thèmes et à un système de gestion de contexte (ici CAMELIS), de façon à ce qu'on puisse l'interroger (par facettes sémantiques) et comparer différentes ressources.

Abstract.
***Akenou-Breizh*, a platform project to develop computational and linguistic resources and tools for breton**
We present a new project, *Akenou-Breizh*, that aims to (1) put in place a platform allowing to study the influences of an *heritage language*, such as Breton, on a *usage language*, such as French, and (2) to make available, to all interested persons, tools well integrated in the "semantic and multilingual web" and proposing proactive access to various kinds of knowledge concerning Breton, as well as direct visualisation of infrasentential correspondences in aligned bilingual presentations. We plan not only to use the numerous freely available resources, in particular those of OPLB and of the APERTIUM project, but also to create new ones, such as good quality bilingual aligned corpora, thereby using the "collaborative web", and to build on the dedicated lingwarium.org web site linguistic modules improving on or extending those that exist, for example a morphological analyzer-generator. We also describe an experiment set up starting from a reduced lexicon for Breton, that shows how it is possible to enrich a classical dictionary, by linking it to a lattice of topics and to a context management system (here CAMELIS), in such a way one can query it (along semantic facets) and compare different resources.

Mots-clés : breton, langue d'héritage, langue d'usage, outils et ressources, études contrastives.

Keywords: Breton, heritage language, usage language, tools and resources, contrastive studies.

1 Problématique et enjeux

On cite souvent Claude Hagège : *les langues sont les drapeaux des identités nationales*. C'est très vrai, et l'on voit des variantes de langue être érigées en langues distinctes pour cette raison, par exemple l'hindoustani en hindi et ourdou, ou le serbo-croate en serbe, croate et bosniaque. L'article 2 de la Constitution Française, dans sa révision de 1992, [2] dispose ainsi que *le français est la langue de la République*. En Inde, fédération d'états, les frontières de certains états ont bougé depuis l'indépendance (1947) pour s'ajuster à l'évolution des fontières linguistiques, et au moins un état nouveau a été créé, celui de Goa, sur la base du conkani.

1. Office Public de la Langue Bretonne, `www.fr.opab-oplb.org/`
2. `http://www.conseil-constitutionnel.fr/conseil-constitutionnel/francais/la-constitution/les-revisions-constitutionnelles/loi-constitutionnelle-n-92-554-du-25-juin-1992.138025.html`

La politique d'uniformisation linguistique en France a mené non seulement à imposer que tous utilisent le français, mais aussi, jusqu'à la dernière guerre au moins, à réprimer l'usage des langues dites "régionales", comme le breton, l'alsacien, l'occitan, le catalan ou le basque. Pourtant les langues sont intrinsèquement et d'abord liées aux divers patrimoines culturels, et les faire mourir reviendrait à détruire une partie importante de l'identité culturelle de nombreux citoyens.

De fait, Claude Hagège commence sa présentation de la linguistique générale [3] en disant que *le langage est une faculté définitoire de l'être humain*, et que *les langues sont la manifestation historique et sociale de cette faculté*. Les langues (et leurs systèmes d'écriture, avec les calligraphies associées) semblent donc être d'abord inséparables des *cultures* qui les ont engendrées et qu'elles ont nourries et nourrissent, et ensuite (et parfois) seulement être liées à une identité nationale.

Il est donc tout à fait normal que les citoyens et plus généralement les habitants d'un pays comme la France, ayant une langue officielle unique, aient un sentiment aigu d'appartenance à telle ou telle culture dont ils sont issus et à laquelle ils participent, et souhaitent approfondir leur connaissance de leur *langue d'héritage*, ou au minimum comprendre comment il se fait que leur façon d'utiliser leur *langue d'usage* soit influencée par cette langue d'héritage, sans même qu'ils la pratiquent, et qu'ils se reconnaissent immédiatement entre participants de la même *culture d'héritage*, alors même qu'ils se parlent dans la langue d'usage. C'est ce qui se passe pour les Bretons en France.

Le projet *Akenou-Breizh* que nous présentons ici est porté par des informaticiens et des informaticiens-linguistes qui éprouvent ce besoin de mieux comprendre la part de leur identité culturelle liée au breton, langue que la plupart ne parlent pas et n'ont que peu ou pas entendue. Quelques-uns voudraient devenir des auto-apprenants du breton et utiliser pour cela des outils et des ressources intégrés au Web contributif (2.0), et au Web sémantique (3.0), en contexte multilingue. Tous voudraient aussi concilier des aspects touchant à leur recherche, et des aspects génériques concernant le "soutien aux langues peu dotées". C'est dire que le projet, s'il commence concrètement sur le breton-français, se veut ouvert à des participants qui s'intéresseraient à d'autres couples *langue d'héritage – langue d'usage*, comme amazigh-arabe, arabe-français, comorien-français, français-anglais (au Canada ou en Acadie), irlandais-anglais, basque-français, etc.

Plus concrètement, le projet *Akenou-Breizh* vise (1) à mettre en place une plate-forme permettant d'étudier les influences d'une *langue d'héritage*, comme le breton, sur une *langue d'usage*, comme le français, et (2) à proposer à l'intention de tous les intéressés des outils s'intégrant au "Web sémantique et multilingue", et proposant des accès proactifs aux connaissances sur le breton, ainsi qu'une visualisation directe des correspondances sous-phrastiques dans des présentations bilingues alignées (voir figure 1a). Nous nous proposons non seulement d'utiliser les nombreuses ressources disponibles librement, en particulier celles de l'OPLB et du projet Apertium, mais aussi d'en créer de nouvelles, comme des corpus bilingues alignés de bonne qualité, en utilisant le "Web collaboratif", et de construire sur le site dédié `lingwarium.org` [4] des modules linguistiques améliorant ou étendant ceux qui existent, par exemple un analyseur-générateur morphologique.

Nous décrivons aussi une expérience réalisée à partir d'un lexique réduit pour le breton, qui montre comment on peut enrichir un dictionnaire classique, en le reliant à un treillis de thèmes et à un système de gestion de contexte (ici CAMELIS), de façon à ce qu'on puisse l'interroger selon des facettes sémantiques variées, et comparer différentes ressources lexicales après les avoir enrichies de cette façon.

Nous commencerons bien sûr par présenter brièvement l'état de l'art, c'est à dire les ressources, les outils et les plates-formes à accès ouvert concernant spécifiquement le breton-français, ou génériques et utilisables dans ce contexte. Dans la section suivante, nous présentons la méthodologie prévue pour le projet *Akenou-Breizh* : (1) à quelles ressources s'intéresser, comment les utiliser, les étendre, les intégrer, et (2) quelles recherches mener durant le projet en utilisant ces ressources. Dans la dernière section, nous présentons en détail l'expérience d'enrichissement d'un dictionnaire du breton et les possibilités qu'elle ouvre pour faire de l'accès lexical selon des facettes sémantiques variées.

3. http ://claude.hagege.free.fr/

4. Le site `apertium.org` offre des outils de développement de systèmes de TA "à transfert de surface" pour des couples de langues à structures de surface très voisines, comme espagnol-catalan ou espagnol-galicien. Ces outils ne permettent cependant pas d'obtenir de bonnes traductions pour des couples éloignés comme russe-français ou français-breton (pour lesquels il faut disposer d'outils permettant de passer par des structures "abstraites"), ni d'augmenter la qualité par spécialisation heuristique à des sous-langages. Le site `http://www.lingwarium.org` créé par Vincent Berment est similaire, mais offre tous les langages spécialisés (ATEF, ROBRA, TRACOMPL, TRANSF/EXPANS, SYGMOR) du système Ariane-G5 créé par Bernard Vauquois et son équipe entre 1971 et 1985, et tous les modules et systèmes de TA réalisés avec eux entre 1972 et 2015, en source ouvert (ru-fr, fr/pr-en, en-my/th, UNL-fr, maquettes zh-fr/en/de/jp/ru et BEX-FEX). Il est prévu de lui ajouter d'autres outils obtenus par réingénierie, comme le langage des SYSTÈMES-Q d'Alain Colmerauer avec lequel le système TAUM-MÉTÉO fut réalisé en 1977 (après 1985, il fut réécrit en GRAMR par J. Chandioux).

2 État de l'art

2.1 Approches informatiques liées au projet

Les progrès de ces dernières années dans les technologies du Web permettent de proposer à tous les internautes des interfaces d'accès à des informations et à des outils répartis sur de nombreux serveurs, et cela à travers les navigateurs installés sur microordinateurs, tablettes et téléphones mobiles.

Accès et visualisation d'informations, utilisation proactive d'outils. La technique des *passerelles* (gateways) permet, d'enrichir l'interface usuelle de lecture d'un texte à travers le Web de diverses façons. L'une des plus intéressantes consiste à faire surgir une *palette* (relative à un mot ou à une phrase) quand le curseur passe au-dessus d'un mot ou d'une phrase. Une *palette de mot* peut contenir des informations morphosyntaxiques (par exemple, pour "irais" : lemme = "aller", cat = "VRB", mode = "COND", temps = "PRES", personne = "1 2", nombre = "SIN"...), ainsi éventuellement que des indications sur son ou ses sens (par exemple, pour "charme" comme NOM : sens = "#arbre #qualité-de-personne #propriété-de-quark"). Il est également possible de présenter les informations trouvées dans plusieurs dictionnaires et bases terminologiques multilingues sous la forme d'un article de dictionnaire construit à partir de ces informations. C'est le principe de l'outil ALEXANDRIA de Dominique Dutoit [5].

Une *palette de phrase* peut contenir une ou plusieurs traductions, une ou plusieurs analyses grammaticales, ou une traduction, avec ou sans une visualisation de l'alignement entre l'original et la traduction. La figure 1a montre un *amphigramme* (alignement hiérarchique entre une phrase source et sa traduction), et la figure 1b illustre la visualisation dynamique produite en SVG par la technique de Christophe Chenon (Chenon, 2005).

Un point important ici est qu'on peut précalculer toutes les informations potentiellement intéressantes pour les utilisateurs à l'avance, et donc ne jamais les faire attendre quand ils demandent une information, une visualisation ou un traitement. C'est une condition nécessaire pour pouvoir proposer des aides *proactives* à la lecture, à l'apprentissage, à l'annotation, ou à des études spécifiques. La *proactivité* est le fait que les fonctionnalités en question peuvent être activées sans que l'utilisateur n'ait à faire aucune action (autre que de régler les paramètres de proactivité bien sûr).

Avec les outils modernes de suivi oculaire par les webcams intégrées à de nombreux dispositifs, on pourrait même produire (dans la page en cours, dans une *palette de paragraphe* ou dans un nouvel onglet) une présentation proactive "à la Vocable" [6] du paragraphe sur lequel l'utilisateur porte son regard : une zone de *minidictionnaire local* serait alors remplie par l'information lexicale liée au paragraphe en question, comme si on avait passé le curseur dessus.

Construction contributive de ressources et d'outils. Nous sommes depuis presque 10 ans dans l'ère du *Web contributif* [7], qui a donné lieu à de nombreux projets de création contributive de ressources linguistiques. En ce qui concerne les **connaissances lexicales**, on peut mentionner les projets PAPILLON, ITOLDU, WIKTIONARY et JEUXDEMOTS.

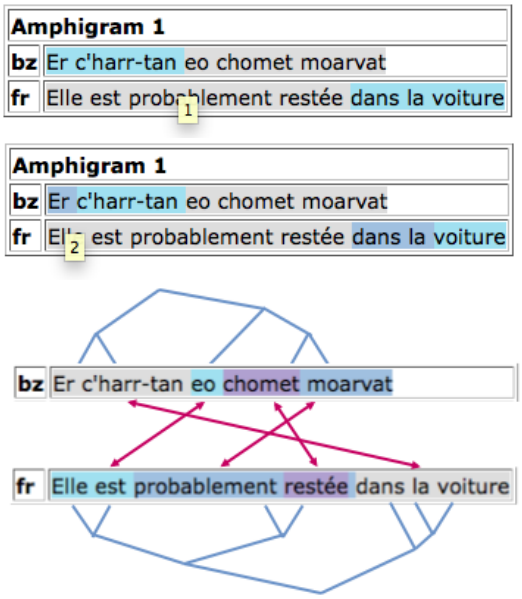
Lancée en 2001 par Mathieu Mangeot et Gilles Sérasset, la base lexicale multilingue en ligne PAPILLON-CDM regroupait en 2003 une vingtaine de dictionnaires électroniques bilingues ou multicible concernant 9 langues, et plus de 2M d'entrées. Son "socle logiciel" `Jibiki` permet de définir des bases lexicales de structures très variées, la définition d'une telle base consistant en la description de sa *macrostructure*, et de la *microstructure* de chacun de ses *volumes*.

Destiné à des élèves-ingénieurs devant apprendre le vocabulaire technique anglais de leurs spécialités, le service Web ITOLDU (Bellynck *et al.*, 2005) a permis en 2003-04 de collecter 17.000 entrées en anglais→français, correctes à plus de 90%, avec 250 étudiants répartis en 15 classes, et 6 enseignants ; 30% de la note d'anglais était donnée par le système.

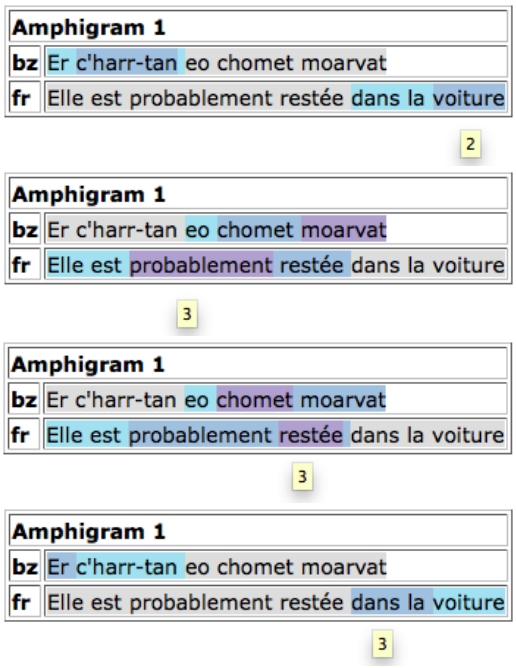
Le projet WIKTIONARY a été lancé en 2002, et contient un dictionnaire d'environ 33.600 mots pour le breton (au 12/4/2015). Voir `http://www.wiktionary.org/` et `http://en.wikipedia.org/wiki/Wiktionary` et `https://br.wiktionary.org/wiki/Wikeriadur:Degemer`.

Il est assez difficile de susciter de nombreuses contributions lexicales de nombreuses personnes. Il y a le plus souvent un petit nombre de "mordus", mais les autres ne contribuent presque rien. Utiliser la myriadisation (Amazon-Turk) ne va pas

5. Une bonne présentation se trouve à l'url `http://www.tv5monde.com/TV5Site/alexandria/entretien.php` : *Comment peut-on définir le logiciel Alexandria ?*
Alexandria appartient à plusieurs familles de logiciel. Une première famille est celle des logiciels d'aide contextuelle ; en effet, Alexandria fournit un service d'aide à la compréhension (définitions, traductions...) quand un lecteur le sollicite. D'autre part, Alexandria appartient à la famille des agents "intelligents" : ce service est disponible en chaque lieu où l'agent a été installé. Aujourd'hui, les lieux où l'on trouve les réplications les plus courantes de cet agent sont les pages html du Web. C'est le cas pour TV5MONDE.org. Mais Alexandria peut prendre aussi d'autres formes techniques.
6. Voir par exemple `http://www.vocable.fr/pdfreader/magazines/vocD632o.pdf`
7. aussi appelé de façon cryptique *Web 2.0*.



(a) Exemple d'amphigramme breton-français. En fait, l'amphigramme est le bi-arbre n-aire obtenu par *tassement* minimal des deux arbres binaires obtenus à partir des scores des *coupures*, voir (Chenon, 2005).



(b) Visualisation dynamique, exemple en français-breton

FIGURE 1: Visualisation de correspondances sous-phrastiques hiérarchiques (amphigrammes).
On voit deux inversions d'ordre, et aussi une correspondance entre un mot (*er*) et deux mots (*dans la*). Les six rectangles montrent comment la visualisation évolue au fur et à mesure qu'on déplace le curseur, dont la position est indiquée par le petit rectangle jaune numéroté. La couleur la plus foncée correspond au groupe le plus petit contenant le curseur.

non plus : même si on paye très peu, ce serait très cher car on vise plus d'un million de mots (on en a plus dans Wiktionary pour l'anglais et le français), et surtout la qualité des informations serait bien trop basse en moyenne, comme plusieurs expériences l'ont prouvé. [8]

Une idée très intéressante est de *motiver les contributeurs potentiels par un aspect ludique*. C'est ce que fait le système JEUXDEMOTS (voir `http://www.jeuxdemots.org`). Créé par Mathieu Lafourcade et Gilles Sérasset en 2006, ce système a ensuite été développé et considérablement étendu par Mathieu Lafourcade. L'objet de ce projet est la construction de ressources lexicales, et plus précisément d'un réseau lexical du français, pouvant servir à diverses applications du TALN. Ces données sont le produit de l'activité des joueurs de JeuxDeMots. Le réseau lexical du français de JDM contient plusieurs centaines de milliers de mots et de relations, telles que la synonymie ou quasi-synonymie, l'antonymie, l'intensification, et plusieurs fonctions lexicosémantiques (FLS) de Mel'tchuk. Il existe des versions de JeuxDeMots pour d'autres langues que le français ; le projet *Akenou-Breizh* se propose d'en créer une pour le breton.

En ce qui concerne les **corpus**, des outils comme IMAG/SECTRA (Boitet *et al.*, 2010) ont déjà permis de créer de bons *corpus parallèles* par post-édition contributive de résultats de TA. Bien que ces résultats soient souvent très mauvais, s'ils sont jugés par des traducteurs ou des linguistes, ils ont une très grande qualité d'usage, si on les utilise pour faire de la post-édition. Ainsi, deux stagiaires Chinois ont pu, durant un stage d'été en 2013, post-éditer environ 500 pages (10.000 segments) de supports de cours d'informatique, à raison d'environ 10 mn/page. Le résultat n'est pas de qualité professionnelle, mais il est comparable à ce que produirait un traducteur junior en 1 h/page environ, la connaissance du domaine et de la terminologie compensant la différence de niveau en langue source. Il semble que même une traduction "pidgin" [9] puisse permettre de post-éditer une page en 25 à 30 mn dans ces conditions.

8. Voir par exemple `http://www.liberation.fr/societe/2015/05/07/miracles-et-mirages-du-crowdsourcing_1297262` et les autres références de Karën Fort & al. sur le sujet.

9. comme celle produite pour le lao-français sur le site `laosoftware.com` : on donne la ou les traductions possibles de chaque mot, dans l'ordre du texte source, avec éventuellement des annotations utiles pour quelqu'un ayant quelques rudiments de la langue source, comme par exemple l'existence

Possibilité d'amélioration contributive des ressources pendant une activité différente. La post-édition de traductions brutes est faite assez volontiers par des internautes accédant des pages Web "prétraduites" dans leur langue, du moment que l'interface est "sans couture" (sans changement de contexte, i.e. sans nouvel onglet ni nouvelle fenêtre).

Il faut également mentionner la possibilité d'améliorer des annotations de tous types de façon contributive, en parallèle à une activité voulue, comme la lecture. Il peut s'agir de sens des mots, d'arbres syntaxiques, d'alignements, ou de graphes sémantiques. La condition essentielle pour y arriver est que l'utilisateur dispose d'une interface intuitive, proactive, à manipulation directe [10].

2.2 Outils et ressources propres au breton

Plusieurs sites, relevant de diverses initiatives (institutionnelles, associatives ou individuelles), regroupent des liens pour le breton, nous en citons quelques-uns : `http://www.fr.opab-oplb.org/35-outils-linguistiques.htm`, par l'office public de la langue bretonne (OPLB), et `http://www.lexilogos.com`. Nous indiquons ci-dessous des outils et ressources liés à notre projet sur le breton.

2.2.1 Logiciels pour le breton

Traduction (APERTIUM et OPLB). Apertium propose une plateforme libre de droits pour développer la traduction automatique entre des paires de langues. Pour le breton, le site de l'OPLB propose un outil en ligne de traduction automatique dans le sens breton→français. L'OPLB développe ce traducteur, en collaboration avec APERTIUM. L'outil `glosbe` repris plus loin propose certaines traductions, dans les deux sens entre le français et le breton.

Détection de langue. l'identification de la langue est parfois une première étape ; une méthode pour les langues celtiques est proposée dans (Minocha & Tyers, 2014). Il y a peu de détecteurs accessibles, comme `openxerox.com` ou `G2LI`, qui gèrent le breton.

Correcteur grammatical. `https://www.languagetool.org/` propose des correcteurs grammaticaux, en open-source, avec un test en ligne, pour plusieurs en langues dont le breton. L'association `http://www.drouizig.org` propose un correcteur pour le breton, `An Drouizig Difazier`, pouvant être intégré à `Office`.

Il faut noter que les langues celtiques présentent certaines particularités, comme le phénomène des mutations (variations de la consonne initiale d'un mot, voir plus bas pour des détails sur le breton), dont une approche possible est décrite par (Poibeau, 2014). La liste d'outils en ligne ci-dessus n'est pas exhaustive, mais nous constatons certains manques, par exemple concernant l'analyse morphologique.

2.2.2 Ressources pour le breton

Dictionnaires en ligne

`Meurgorf` est un dictionnaire historique du breton en ligne (interrogeable) proposé par l'OPBL. Il contient actuellement 55.747 entrées [11], et continue à s'enrichir (`Meurgorf` est un projet).

APERTIUM et l'OPLB proposent plusieurs dictionnaires concernant le breton disponibles dans APERTIUM (Tyers *et al.*, 2009; Forcada *et al.*, 2011) (dans un format XML spécifique). Il s'agit d'un dictionnaire morphologique par langue (pour le breton, pour le français), et d'un dictionnaire bilingue portant des indications grammaticales pour les mots des deux langues.

Francis Favereau est l'auteur de plusieurs dictionnaires, avec des versions en ligne `http://www.arkaevraz.net/dicobzh/index.php` (volumétrie : fr : 37.401 ; bz : 33.440) permettant d'ajuster la recherche (par exemple, avec mutation : "vugale" trouve "bugale").

Un autre lien est : `http://www.agencebretagnepresse.com/cgi-bin/dico.cgi`.

Tomaz Jacquet est l'auteur du dictionnaire `Freelang`, dans les deux sens entre le français et le breton, voir `http://www.freelang.com/enligne/breton.php?lg=fr` (volumétrie : 37.800 entrées).

`Brezhoneg 21` `http://www.brezhoneg21.com` est une ressource scolaire, en sciences et techniques.

de cas en russe, pour du russe-français.

10. comme par exemple le logiciel *Annot*^{ED} de Johan Ségura (Ségura, 2012).

11. `http://meurgorf.opab-oplb.org/page/index/pr__sentation_du_projet`, consulté le 8/4/2015.

Geriadur, disponible à http://www.geriadur.com/, traduit du français vers le breton (volumétrie : traduction de 22.302 mots français).

Logos gère une ressource libre et contributive, avec des dictionnaires monolingues ou multilingues, interrogeables (à http://www.logosdictionary.org/index.php). Il comprend une version monolingue pour le breton.

http://br.wiktionary.org/ : le projet WIKTIONARY de dictionnaires descriptifs et libres contient un sous-projet pour le breton.

https://fr.glosbe.com/br/fr : ce site se présente comme un dictionnaire multilingue. Il propose des traductions (par exemple entre le français et le breton), et utilise des *mémoires de traductions*. Par exemple, en réponse à la question "pajennoù", le site glosbe répondra que le mot est inconnu mais que l'expression similaire "pajennoù melen" est connue (avec l'équivalent : "pages jaunes").

Autres ressources

Le site ARBRES (http://arbres.iker.cnrs.fr/index.php/Arbres:Le_site_de_grammaire_du_breton) est un site d'informations sur la grammaire du breton, et un centre de ressources pour la recherche en syntaxe formelle sur la langue bretonne (Jouitteau, 2005).

Mentionnons aussi l'atlas linguistique ALBB (http://sbahuaud.free.fr/ALBB/), la base de données toponymique KerOfis de l'OPLB (http://www.fr.opab-oplb.org/40-kerofis.htm), et la base de données TermOfis du centre de terminologie de l'OPLB, avec 62.794 termes (http://www.fr.opab-oplb.org/36-termofis.htm.

Enfin, on peut consulter l'inventaire fait par par l'ELDA, en 2014 [12].

3 Méthodologie

La méthodologie du projet *Akenou-Breizh* s'articule autour de deux pôles : *ressources*, et *recherches*. Les ressources en question sont bien sûr celles qu'on estime nécessaires pour mener les recherches qui, *in fine*, motivent le projet. Cependant, on ne peut pas disposer des ressources très détaillées ou très volumineuses requises par certaines recherches, même si ce sont les plus intéressantes, avant de disposer de ressources moins détaillées ou moins volumineuses. C'est pourquoi la méthode suivie par le projet consistera à élaborer des ressources, des annotations, des outils d'accès ou des traitements au fur et à mesure des possibilités, et à lancer les recherches souhaitées quand les ressources et outils minimaux nécessaires seront disponibles.

3.1 Construction ou collecte de ressources, facilitation d'accès, enrichissement

Corpus parallèles avec visualisation des alignements sous-phrastiques. La pemière tâche concrète du projet sera la construction d'une base de données (BDcorp) contenant des *corpus parallèles* (phrase à phrase en regard) français-breton, munis d'accès interactifs à des dictionnaires en ligne (ce qu'on sait déjà bien faire, grâce à des outils comme ALEXANDRIA ou IMAG/SECTRA) et montrant par des couleurs les correspondances sous-phrastiques (voir figure 1 ci-dessus), l'infrastructure de la BDcorp devant permettre d'ajouter des annotations variées (comme des arbres linguistiques, des annotations discursives, etc.), et de mener des études visant à construire et exploiter ces annotations.

Pour construire des corpus parallèles (alignés au niveau des phrases), et cela de façon essentiellement *contributive (et bénévole)*, on commencera par créer les mémoires de traductions associées, et à améliorer les traductions, si nécessaire, en utilisant une passerelle IMAG/SECTRA d'accès multilingue interactif, qui permet à des contributeurs organisés ou occasionnels de "post-éditer" les traductions, segment par segment (Boitet *et al.*, 2010) [13]. Il est aussi possible de créer de nouvelles traductions en utilisant cette même interface : on peut partir de traductions "pidgin" (cf. supra), reposant sur un analyseur morphologique, un dictionnaire bilingue, et un générateur morphologique. Ces trois composants existent déjà pour le breton, en source ouvert (dans le projet APERTIUM), ce qui fournit un point de départ immédiatement utilisable.

12. Rapport disponible à http://www.culturecommunication.gouv.fr/content/download/106817/1248227/version/1/file/Rapport_dglflf_05112014.pdf

13. http://service.aximag.fr/xwiki/bin/view/imag/home

Il y a ici une possibilité intéressante, qui n'est pas de la recherche, mais une aide potentielle importante à l'auto-apprentissage ou à la simple découverte du breton. Elle consiste à produire dynamiquement une vue bilingue parallèle avec visualisation des correspondances sous-phrastiques, ou, pour les plus avancés, une vue monolingue annotée *à la Vocable*, c'est-à-dire avec affichage proactif d'un minidictionnaire associé à la page, au paragraphe ou au segment (phrase ou titre) en cours de lecture. De même que pour les traductions, il est possible de corriger interactivement les alignements, par manipulation directe depuis le contexte de lecture (Ségura, 2012).

Quand on disposera de suffisamment de bons bisegments pour un sous-langage donné [14] (par exemple, des sites Web ou des romans sur la Bretagne, ou des sections de journaux), on pourra développer des systèmes "empiriques" de TA (statistiques ou fondés sur les exemples) en utilisant des outils comme Moses (en version "factorielle" à cause de la richesse des morphologies flexionnelles du breton et du français, en intégrant les mutations dans la morphologie flexionnelle).

Intégration du breton dans une base lexicale multilingue organisée par acceptions interlingues. Une seconde tâche essentielle est de construire une base lexicale plus "sémantique" que ce qui existe, bien sûr en commençant par réutiliser l'existant, et si possible en le faisant par accès aux ressources en ligne, de façon à bénéficier immédiatement des apports dont elles-mêmes bénéficient constamment.

La première étape serait sans doute d'importer les dictionnaires libres de droits et existant en format XML dans la base lexicale multilingue Papillon-CDM créée et gérée par Mathieu Mangeot (Mangeot, 2001) [15]. La seconde consisterait à enrichir ces données au niveau sémantique, en les reliant à un treillis de domaines et à un ensemble de dénotations de sens, comme les lexèmes interlingues (dits UW) d'UNL [16] (voir `http://www.undl.org`).

Amélioration et extension d'outils pour l'analyse et la génération morphologique. Pour pouvoir accéder aux informations lexicales à partir des mots des textes, il faut en faire l'analyse morphologique. Cela peut se faire par un analyseur, activé pour chaque forme, ou par consultation d'une liste de formes accompagnées de leurs attributs morphosyntaxiques, produite par génération morphologique.

Le projet *Akenou-Breizh* cherchera à étendre les outils existants, tant en termes de couverture que de puissance : il s'agit de couvrir tous les mots simples des dictionnaires existants, de traiter la morphologie dérivationnelle et la morphologie compositionnelle (pour les mots composés), et enfin de construire une "grammaire du mot inconnu" (un "devineur") pour le breton, comme celle réalisée en ATEF [17] par J. Ph. Guilbaud pour le français (Guilbaud & Boitet, 1997).

Désambiguïsation lexicale et préparation à l'extraction de contenu. On utilisera les ressources lexico-sémantiques produites pour transposer au breton les techniques de désambiguïsation lexicale (WSD) qui fonctionnent déjà pour les langues "bien dotées" (le domaine a bien progressé ces dernières années, en particulier grâce à l'organisation des campagnes CLEF). Cela permettra d'intégrer le breton à un extracteur multilingue de contenu construit sur le modèle proposé par le projet ANR OMNIA (Falaise *et al.*, 2010). [18].

3.2 Recherches utilisant ces ressources

Les recherches envisagées par le projet *Akenou-Breizh* devraient concourir à l'étude de l'influence des langues d'héritage sur les langues d'usage. On peut distinger celles qui portent sur les phénomènes linguistiques, à plusieurs niveaux d'interprétation, et à plusieurs degrés de complexité, de celles qui toucheront aux aspects plus pragmatiques et culturels des rapports entre breton et français.

14. Pour un sous-langage assez restreint, partir de 2.000 bisegments de 20 mots, soit environ 100 pages standard, est suffisant si on fait l'apprentissage sur un corpus annoté (par les valeurs des "facteurs", i.e. des attributs morphosyntaxiques).

15. Voir `http://www.papillon.org`. L'import se fait en quelques minutes, une fois qu'on a précisé par des chemins `Xpath` où est l'information correspondant à chaque balise CDM (*Common Dictionary Markup*).

16. UNL = Universal Networking Language. C'est à la fois un projet initié par l'UNU en 1996 et un formalisme de graphes "anglo-sémantiques" utilisable tant pour la traduction que pour le résumé et l'extraction de contenu.

17. ATEF = Analyse de Textes en États Finis. C'est un langage créé par J. Chauché en 1975 (voir Microfiche ACL de 1975) pour écrire des analyseurs morphologiques. C'est le tout premier langage spécialisé fondé sur le modèle des transducteurs finis. Il contient des extensions permettant la programmation heuristique, la modification dynamique de la chaîne en entrée, l'analyse de mots composés, et un vrai traitement des mots inconnu.

18. Ce modèle est en plusieurs étapes. (1) On transforme une phrase ou un texte (ou la partie textuelle d'une requête) en un graphe donnant les segmentations et les analyses morphologiques possibles. (2) On enrichit ce graphe en attachant à chaque lemme tous les UW (lexèmes interlingues UNL) lui correspondant dans la base lexicale. (3) On utilise un module de désambiguïsation lexicale qui attache des scores aux UW. (4) On construit ou on met à jour (automatiquement) un alignement entre la "préontologie" formée par l'ensemble des UW et l'ontologie (ou la base de connaissances) vers laquelle on veut extraire le contenu "pertinent". (5) À l'aide d'une connaissance minimale de la langue du texte (grammaires locales), on construit et on score les groupes élémentaires (*chunks*) formé d'UW alignés avec des concepts ou des attributs de l'ontologie, et on produit un *descripteur* (une liste attributs-valeurs, ou une liste de petits graphes UNL). (6) On transforme ce descripteur en une *description* dans le langage de l'ontologie. Si c'est une donnée, on la range dans la T-box (conteneur des *termes*). Si c'est une requête, on lance l'agorithme de production d'une réponse.

3.2.1 Recherches à plusieurs niveaux et à plusieurs degrés de complexité

Il s'agit de

- recherches simples basées sur les textes parallèles ;
- recherches nécessitant aussi des dictionnaires (donc sans doute de la lemmatisation) :
  - . extraction de mots ou termes hors dictionnaire,
  - . extraction de couples de mots ou de termes "candidats" à l'équivalence (en traduction),
  - . étude des registres d'expression (par exemple, formes polies d'un côté, directes de l'autre).
- recherches nécessitant aussi des analyses sous forme d'arbres syntaxiques :
  - . oppositions actif/passif,
  - . expressions de la modalité et de la politesse (tournures, modes...),
  - . étude contrastive des "mots composés" ou "tournures".

3.2.2 Recherches autour de la pragmatique du breton (approche statistique et linguistique)

Au travers de présupposés et sous-entendus portés par des locutions choisies, on vise

- à étudier les associations de termes et leurs variations dans le temps et l'espace,
- à comparer avec le français et à mettre en évidence des différences et influences entre les deux langues.

4 Un cas d'étude

Nous décrivons une expérience qui met en œuvre un petit lexique et un système d'information permettant de le manipuler.

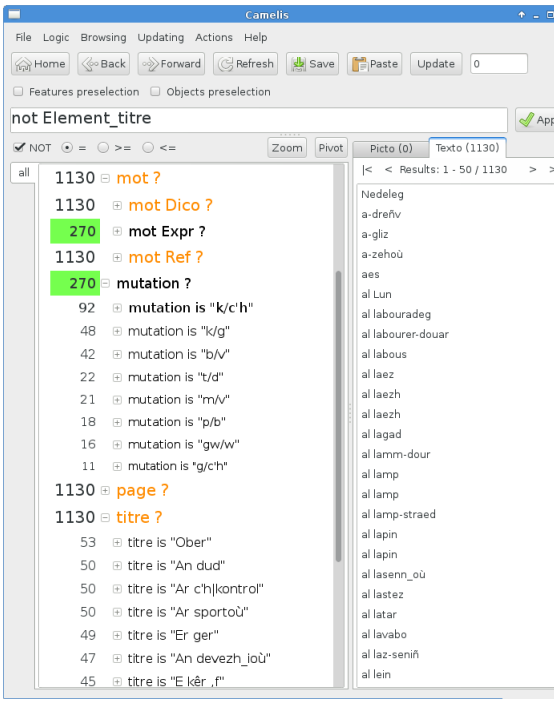
4.1 Ressources et outils utilisés

Le lexique choisi suit l'album (Kergoat *et al.*, 2007) qui fournit le vocabulaire fondamental du breton (1000 mots), conçu au départ pour les enfants (pour plus de dix langues). Le lexique a été saisi par thèmes. Il est organisé en conservant le rattachement des mots et expressions à des thèmes correspondant à des scènes de la vie quotidienne ("à la maison", "à l'école", etc.).

Les noms sont indiqués avec un déterminant, et leur genre (dans l'album aussi). Des synonymes sont proposés dans l'album ; nous les avons aussi indiqués, avec un marquage. Le breton partage avec d'autres langues celtiques un phénomène connu sous le nom de *mutation* : il s'agit d'une modification de la consonne intiale (comme $k > c'h$), régie par plusieurs sortes de déclencheurs [19]. Nous avons choisi d'inclure en plus dans le lexique une indication de mutation (par exemple, l'expression *an daol* pour "une table" sera indiquée par : *an d|taol*, le lemme du nom étant *taol*, et la mutation avec cet article *an* étant $d > t$. Cette information supplémentaire permettra ensuite d'interroger le *contexte* du lexique selon ce critère (fréquence d'une mutation particulière, son contexte, etc.), éventuellement combiné à d'autres critères.

Le système d'information choisi est un *contexte* CAMELIS, un tel contexte étant défini par un ensemble fini d'objets, avec pour chaque objet un ensemble fini de descriptions (formules logiques). CAMELIS (version 1, accessible à `http://www.irisa.fr/LIS/ferre/camelis/`) est un *système de gestion de contexte logique* permettant plusieurs formes de manipulations flexibles par *facettes* : interrogation/navigation sans connaissance *a priori* (par clics et sélections successives), mais guidé par un index contextuel de propriétés et de manipulations plus expertes (écriture de requêtes, mise à jour). Ce logiciel est basé sur l'analyse de concept logique (LCA) définie dans (Ferré & Ridoux, 2004), qui propose une extension de l'analyse de concept formel (FCA, voir (Ganter & Wille, 1999)). Un *concept logique*, noté c, est un couple formé d'une extension $ext(c)$ (un ensemble d'objets) et d'une intension $int(c)$ (une formule), tel que les éléments de $ext(c)$ sont exactement ceux qui vérifient $int(c)$. Ces concepts forment un treillis auquel correspond l'*arbre de navigation logique* et incrémentale dans la fenêtre gauche du logiciel. Le logiciel CAMELIS est aussi prévu pour gérer des ensembles d'objets de types différents et plusieurs facettes (graduelles) dans l'arbre de navigation.

19. `http://arbres.iker.cnrs.fr/index.php?title=Les_mutations_consonantiques`, consulté le 01/4/2015



(a) Contexte initial, sans les objets pour les titres

(b) Contexte après nouvelle sélection dans l'index

FIGURE 2: Contexte monolingue

4.2 Construction de contextes et usages

Nous illustrons quelques usages du lexique transformé en contexte à explorer : le lexique breton, d'abord pris isolément, ensuite complété avec d'autres ressources. Ce type de scénarios illustre une recherche d'information possible, mais aussi une forme d'évaluation d'une ressource en fonction d'une autre.

Rôles des fenêtres CAMELIS par rapport à un contexte. L'outil CAMELIS, chargé avec un contexte initial, présente trois fenêtres relatives à un contexte courant, qui évolue au fil des sélections dans ces fenêtres. Pour les figures 2a et 2b , le contexte initial ("home", pour "all" dans la fenêtre supérieure) contient deux sortes d'objets, d'une part pour les expressions illustrées dans l'album, et d'autre part pour les titres. Il s'agit au départ d'informations monolingues. Ce contexte "home" sera augmenté par la suite, pour intégrer des informations multilingues. [20]

Fenêtre d'objets : la partie droite présente les objets du contexte courant, par leur label [21]

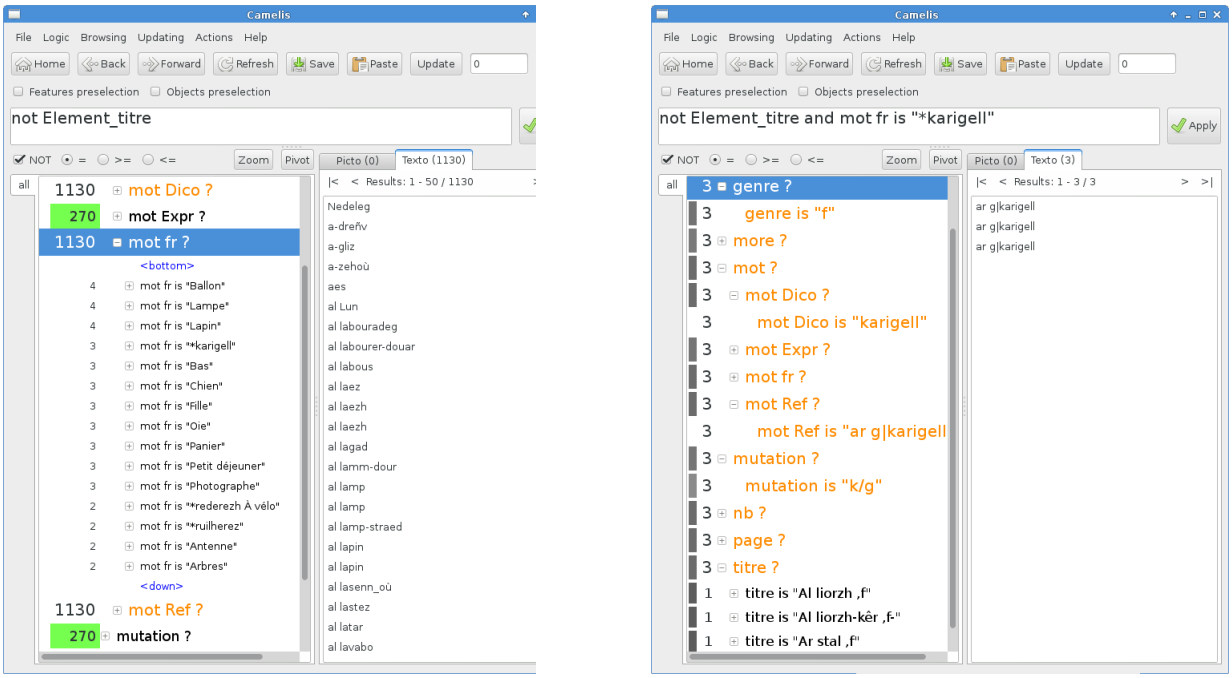
Fenêtre de propriétés : la partie gauche indique les propriétés, organisées en arbres selon les relations entre les propriétés. Il s'agit aussi d'un index cliquable qui permet de passer d'un contexte à un autre.

Fenêtre de requête : la partie du haut contient une requête caractérisant le contexte courant : c'est une propriété satisfaite par tous les objets du contexte courant ; elle n'a pas besoin d'être saisie puisqu'elle est mise à jour automatiquement selon les sélections dans les deux autres fenêtres. L'utilisation ne nécessite pas de connaissance *a priori*, mais il est aussi possible de rédiger directement les requêtes.

Sur le lexique initial

Les thèmes : nous voyons dans l'index de navigation à gauche, la facette `titre` qui indique, *pour chaque objet* mentionné à droite, le titre auquel il est rattaché (dans l'album et dans le contexte). Après sélection à gauche d'un

20. Un tel contexte peut être hétérogène, il peut contenir plus de sortes d'objets et de propriétés, selon les préférences et les usages prévus.
21. ils pourraient être présentés avec une photo, par l'onglet "Picto", selon le type de contexte



(a) Contexte breton et français

(b) Contexte après nouvelle sélection dans l'index

FIGURE 3: Contexte multilingue

`titre` particulier (ou de plusieurs titres, comme dans la figure 2b), les trois fenêtres seront synchonisées pour représenter le nouveau contexte, de façon que les objets à droite soient ceux rattachés à ce `titre` et que la requête en haut représente la propriété choisie, vérifiée par ces objets.

Les mutations : nous voyons aussi plus haut à gauche, la facette `mutation` qui indique, *pour certains objets*, la forme de mutation associée (dans le contexte [22]). Les mutations sont ordonnées ici selon leurs fréquences (on constate que les mutations en "k" dominent). Les couleurs à gauche servent à indiquer des propriétés satisfaites par un même ensemble d'objets : ici, les objets présentant une mutation ont ainsi été annotés pour indiquer leur mot de référence (`mot Ref is "..."`)

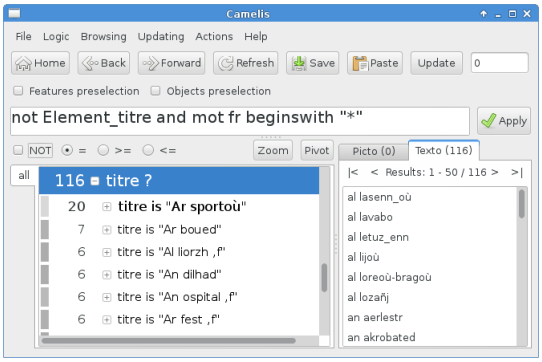
Quelques autres propriétés sont indiquées. Ainsi, la facette `AvecArticle` (sous `Classe` par un "axiome CAMELIS") rend visible le fait que la plupart des expressions contiennent un article, grâce à l'affichage automatique de son cardinal.

Avec intégration de ressources APERTIUM. Le contexte "home" est maintenant augmenté, pour intégrer des informations multilingues provenant d'APERTIUM. Pour chaque langue ajoutée, les mots qui sont inconnus d'APERTIUM sont marqués par * (au début). Cela peut servir de point de départ à diverses évaluations, comme (1) une analyse de couverture du traducteur et des suggestions de complétion, et (2) une caractérisation des cas sans traduction [23]. Certaines incohérences peuvent aussi être mises en évidence.

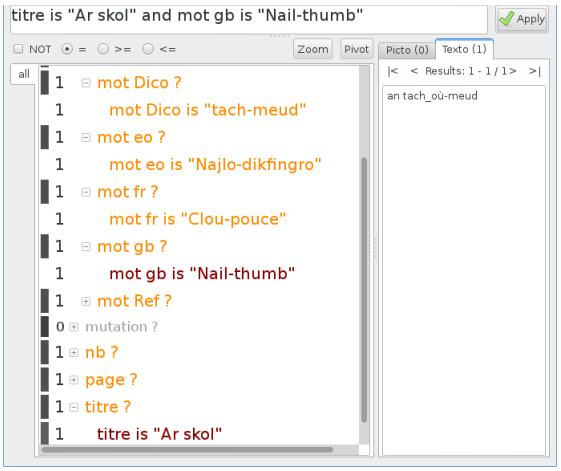
Facettes multilingues. les informations pour le français sont de deux sortes : une propriété *nb* =... indique le nombre d'occurrences dans un fichier dictionnaire pour la paire de langue breton-français ; une propriété `mot fr is ...`, comme dans l'image 3a, indique la traduction fournie par APERTIUM pour le `mot Dico is ...` associé à un objet/expression en breton. Par la suite, la traduction en espéranto avec une facette `mot eo is ...`, puis celle en anglais avec `mot gb is ...` sont aussi ajoutées.

Mise en évidence de manques. L'image 3b illustre un cas de mot breton présent dans plusieurs pages/thèmes de l'album et pourtant absent d'APERTIUM. Une requête plus fine, comme celle de l'image 4a, permet de constater le nombre

22. ajout par rapport à l'album
23. De cette façon, des erreurs dans une première version du source ont déjà pu être repérées.



(a) Contexte breton et français ; mots inconnus



(b) Contexte br, eo, fr, gb ; thème "école"

FIGURE 4: Contexte multilingue et Apertium

de mots de l'album non connus dans APERTIUM, avec les thématiques principalement concernées (ici les sports : "Ar sportoù"). Globalement, un peu plus de 100 mots parmi ceux de l'album n'ont pas de traduction. [24].

Autres défauts. Des problèmes particuliers peuvent apparaître avec des mots composés. En ajoutant les autres langues, comme dans l'image 4b, les manières d'interroger se multiplient, et une mise en rapport de facettes sans passer par le breton permet d'autres repérages, comme entre le français "clou-pouce" et l'anglais "thumb-nail", associés au même objet.

Note. Toutes les fonctionnalités de CAMELIS ne sont pas exploitées ici, comme les actions et la mise à jour.

5 Conclusion

Nous avons présenté dans cet article un nouveau projet, *Akenou-Breizh*, qui concerne directement le TALN appliqué au breton et au couple breton-français. Plus précisément, il vise (1) à mettre en place une plate-forme permettant d'étudier les influences d'une *langue d'héritage*, comme le breton, sur une *langue d'usage*, comme le français, et (2) à mettre à disposition de tous les intéressés des outils s'intégrant au "Web sémantique et multilingue", et proposant des accès proactifs aux connaissances sur le breton ainsi qu'une visualisation directe des correspondances sous-phrastiques dans des présentations bilingues alignées.

Après avoir présenté un état de l'art assez complet des outils et ressources concernant le breton, puis des nouvelles possibilités apportées par le Web en termes d'interfaces non seulement d'accès enrichi, mais aussi de contribution pour l'enrichissement ou la création de ressources et d'outils permettant des applications et des recherches à divers niveaux, de la morphologie à la pragmatique, nous avons décrit la méthodologie prévue dans le projet.

Notre premier but est de valoriser les ressources et les outils pour le breton (écrit), avec des accès proactifs. Nous souhaitons ensuite les évaluer et les enrichir ; une étude a été menée en ce sens pour permettre un accès lexical avec des facettes sémantiques variées. Il est prévu d'associer à ce projet d'autres partenaires spécialistes du breton qui produisent ou soutiennent des développements pour le breton. À terme, d'autres langues pourraient être aussi concernées, comme celles de la famille celtique.

24. Quelques erreurs dans l'album ne sont pas exclues non plus, cette méthode peut aider à les repérer aussi

Remerciements

Nous remercions l'Office public de la langue bretonne et des collègues de Rennes-2, pour les discussions utiles et leur soutien pour ce projet.

Références

BELLYNCK V., BOITET C. & KENWRIGHT J. (2005). *ITOLDU, a Web Service to Pool Technical Lexical Terms in a Learning Environment and Contribute to Multilingual Lexical Databases*. In A. GELBUKH, Ed., *Computational Linguistics and Intelligent Text Processing (Proc. CICLING-2005)*, number 3406 in LNCS, p. 319–327 : Springer.

BOITET C., HUYNH C.-P., NGUYEN H.-T. & BELLYNCK V. (2010). *The iMAG concept : multilingual access gateway to an elected Web site with incremental quality increase through collaborative post-edition of MT pretranslations*. In *Actes de TALN 2010, Montréal, Canada*.

CHENON C. (2005). *Vers une meilleure utilisabilité des mémoires de traductions, fondée sur un alignement sous-phrastique*. PhD thesis, UJF.

FALAISE A., ROUQUET D., SCHWAB D., BOITET C. & BLANCHON H. (2010). *Ontology-driven content extraction using interlingual annotation of texts in the OMNIA project*. In *Proc. CLIA workshop of COLING-2010* : ACL.

FERRÉ S. & RIDOUX O. (2004). *Introduction to logical information systems*. *Inf. Process. Manage.*, **40**(3), 383–419.

FORCADA M. L., GINESTÍ-ROSELL M., NORDFALK J., O'REGAN J., ORTIZ-ROJAS S., PÉREZ-ORTIZ J. A., SÁNCHEZ-MARTÍNEZ F., RAMÍREZ-SÁNCHEZ G. & TYERS F. M. (2011). *Apertium : a free/open-source platform for rule-based machine translation*. *Machine translation*, **25**(2), 127–144.

GANTER B. & WILLE R. (1999). *Formal concept analysis - mathematical foundations*. Springer.

GUILBAUD J.-P. & BOITET C. (1997). *Comment rendre une morphologie robuste du français encore plus robuste en traitant finement les mots inconnus avec les données disponibles*. In *Actes de TALN-97, Grenoble*, p. 12 p. : CLIPS, UJF.

JOUITTEAU M. (2005). *La syntaxe comparée du breton, une enquête sur la périphérie gauche de la phrase bretonne*. PhD thesis, Nantes, France.

KERGOAT L., AMERY H. & CARTWRIGHT S. (2007). *Les 1000 premiers mots en breton*. Skol an Emsav, 8 edition.

MANGEOT M. (2001). *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. PhD thesis, UJF.

MINOCHA A. & TYERS F. (2014). *Subsegmental language detection in Celtic language text*. In *Proceedings of the First Celtic Language Technology Workshop*, p. 76–80, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

POIBEAU T. (2014). *Processing Mutations in Breton with Finite-State Transducers*. In *Proceedings of the First Celtic Language Technology Workshop*, p. 28–32, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

SÉGURA J. (2012). *Mémoires partagées d'alignements sous-phrastiques bilingues*. PhD thesis, LIRMM, Université de Montpellier II.

TYERS F. M., DUGAST L. & PARK J. (2009). *Rule-based augmentation of training data in Breton–French statistical machine translation*. *Proceedings of the 13th Annual Conference of the European Association of Machine Translation, EAMT09*, p. 213–218.