Accéder à et comprendre des contenus en portugais par des étrangers
Peut-on généraliser cette méthodologie à d'autres langues ?

Claudio Menezes

Département de Langues Étrangères et Traduction (LET), UNESCO Chair sur Politiques Linguistiques pour le Multilinguisme, Université de Brasília, Brésil

Résumé : Avec l'émergence des bibliothèques numériques, l'accessibilité à distance est devenue une pratique courante dans le cyberespace. Dans le domaine universitaire, un certain nombre d'initiatives sont connues, telles que le NDLTD – un réseau international de mémoires et de dépôts de thèses – ou la BDTD – Bibliothèque numérique de thèses et de mémoires au Brésil. Ce sont des interfaces qui permettent d'accéder au texte intégral des documents dans la langue originale du texte. De plus, dans le contexte actuel de mondialisation, l'offre d'inscription d'étudiants étrangers dans les universités est devenue plus importante, ce qui implique la nécessité d'un service de bibliothèque pour ce type d'utilisateur. Étant donné le volume important de contenu scientifique, la traduction complète de ces documents semblerait être une tâche herculéenne irréalisable. Dans ce contexte, cette recherche vise à identifier les obstacles rencontrés par les étrangers pour accéder à et comprendre le contenu scientifique. Elle a comme objectif de développer une méthodologie supportée par une application informatique qui pourrait améliorer la compréhension des étudiants en leur facilitant ces contenus. La méthodologie proposée peut être adaptée à n'importe quelles langues. On peut également envisager son adaptation aux langues des signes et à la communication orale. La communication proposée couvre divers aspects : – la place de la compétence culturelle et du multilinguisme dans les dispositifs d'apprentissage des langues ; – l'interaction entre l'apprentissage des langues et la socialisation des élèves / apprenants allophones ; – la question de l'autonomie dans les dispositifs d'apprentissage ; – la question de la formation des professeurs de langue seconde et de la langue de scolarisation ; – l'intégration sociale, culturelle et professionnelle des apprenants allophones – les langues accessibles sur le Web.

Mots-clés : résumé automatique ; bibliothèque numérique ; traduction ; alignement

1 Introduction

La création de documents accessibles à distance dans les bibliothèques numériques est devenue une pratique fréquente dans le cyberespace. Dans le domaine académique, un

certain nombre d'initiatives sont connues, y compris la NDLTD [1] – réseau international de dépôts de mémoire et de thèses – et la BDTD [2] – Bibliothèque numérique de thèses et dissertations – au Brésil. Ce sont des projets qui permettent l'accès au texte intégral des thèses de doctorat et des mémoires de maîtrise dans la langue originale de l'œuvre. En outre, dans le monde globalisé actuel, l'offre d'inscriptions aux étudiants étrangers par les universités est une politique fréquemment adoptée, ce qui implique le besoin d'un service de bibliothèque pour ce type d'utilisateur. Étant donné le volume élevé de contenu scientifique, la traduction complète de ces documents semblerait être une tâche herculéenne et difficile pour les différentes langues des utilisateurs étrangers. Dans ce contexte, la présente recherche vise à identifier les obstacles rencontrés par les étrangers pour accéder à et comprendre les contenus scientifiques, et à développer une méthodologie qui repose sur une application informatique qui pourrait faciliter leur compréhension par les étudiants francophones. La méthodologie proposée peut être adaptée à toutes les langues [3] (à partir d'une paire constituée par une langue source ou langue de départ et une langue cible ou langue d'arrivée).

2 Synthèse de la recherche

L'identification des obstacles pour accéder à et comprendre les textes scientifiques par des étudiants francophones de l'Université de Brasilia et de l'Université de Lille 3 (Université Charles de Gaulle) a été possible grâce à l'élaboration de questionnaires [4], pour obtenir des données sur l'expertise linguistique, la connaissance et l'usage de logiciels de traitement automatique des langues. D'une façon générale, deux phénomènes ont été identifiés : 1) le simple désistement renoncement à la de 1'usage de bibliographie scientifique en portugais par l'usager étranger dû à une maîtrise insuffisante de la langue portugaise ; 2) la méconnaissance des outils de traitement automatique de chaque langue pour faciliter l'accès et la compréhension de textes scientifiques disponibles dans les bibliothèques numériques.

À partir de cette constatation, on a entamé une recherche exhaustive des outils informatiques pouvant aider l'usager étranger à accéder à et comprendre des contenus scientifiques disponibles dans un corpus de thèses et mémoires de l'Université de Brasilia, faisant partie de la Bibliothèque Numérique de Thèses et Dissertations (BDTD), projet coordonné par l'IBICT (Institut Brésilien d'Information en Science et Technologie).

3 Description de la méthodologie

L'idée clé de la proposition est de construire une représentation sémantique des textes scientifiques afin d'éviter la nécessité d'une traduction complète des thèses et des mémoires originaux. Pour ce faire, quatre technologies de calcul ont été utilisées :

1. http://www.ndltd.org/
2. http://bdtd.ibict.br/vufind
3. Le logiciel d'appui à cette méthodologie est disponible à http://multilingua.cdtc.unb.br:8080/
4. Le pré-test a été réalisé avec des étudiants francophones du Centre d'enseignement et de recherche portugais (NEPPE), de l'UnB. Une deuxième collecte de données a été réalisée avec des étudiants en Sciences de l'information de l'Université Charles de Gaulle (Lille 3).

Published in : Antonella Corvaglia and Françoise Wolf-Mandroux (dir.), *Formation linguistique des apprenants allophones et pédagogies innovantes*, Éditions des archives contemporaines, Coll. «PLIDAM», France, ISBN : 9782813003621,
doi : https ://doi.org/10.17184/eac.9782813003621

le filtrage, la synthèse automatique, la traduction automatique et l'alignement des phrases, comme expliqué ci-dessous.

Puisque les documents disponibles dans les bibliothèques numériques sont en général au format *pdf* et contiennent des chapitres ou des sections sans intérêt sémantique majeur, la première ressource de calcul utilisée est un filtre. Cette fonctionnalité produit un nouveau document au format *txt*, débogué des sections sans intérêt sémantique (remerciements, présentation, index, bibliographie par exemple). Ce nouveau document ne contient que les chapitres de la thèse ou de la dissertation à étudier. La conversion du format en *txt* est due à la nécessité d'utiliser des programmes dans lesquels les fichiers d'entrée sont obligatoires dans ce format.

La deuxième composante de l'outil informatique – la synthèse automatique – permet de produire un texte plus petit formé par les phrases les plus pertinentes du document original. Il existe plusieurs techniques et critères pour créer des résumés pertinents. Pour fournir une plus grande flexibilité à l'utilisateur, les paramètres d'entrée sont les numéros de page de début et de fin du texte à résumer. Il est toutefois recommandé de choisir comme paramètres la première page du premier chapitre et la dernière page du dernier chapitre d'une thèse. L'utilisateur doit également fournir le taux de compression souhaité, lequel paramètre indiquera la taille du résumé à créer et à traduire ultérieurement. Deux outils de synthèse automatique du portugais ont été testés : GISTSUMM et GENSIM. Cependant, il est à noter que dans une des expériences réalisées, la précision et la couverture sont calculées en utilisant le programme ROUGE [5] pour avoir une idée de la qualité du résumé produit par GISTSUMM.

Le troisième composant de l'application – la traduction automatique – permet d'obtenir du texte traduit dans la langue cible. Dans notre cas, nous travaillons avec les deux langues (FR, PT), mais le logiciel choisi a la capacité de traduire en six langues : portugais, français, espagnol, allemand, anglais et japonais.

Le quatrième composant – l'alignement automatique de phrases – produit un bi-texte en portugais et dans la langue cible. Dans la recherche ciblée, des expériences ont été réalisées avec les deux langues PT – FR. Les figures 1 et 2 montrent un exemple de bi-texte produit par cet outil informatique et une version formatée du bi-test avec identification de la bibliothèque centrale de l'UnB.

5. ROUGE : Recall-Oriented Understudy for Gisting Evaluation

Published in : Antonella Corvaglia and Françoise Wolf-Mandroux (dir.), *Formation linguistique des apprenants allophones et pédagogies innovantes*, Éditions des archives contemporaines, Coll. «PLIDAM», France, ISBN : 9782813003621, doi : https ://doi.org/10.17184/eac.9782813003621

Mémoire de maîtrise par Ana Cláudia Vieira Braga

Norme linguistique et prétendue oralité dans la traduction de Persépolis [6]

De certa forma, há uma ampliação dos recursos que serão aplicados na tradução : o autor utilizou a imagem que deve ser considerada pelo tradutor como chave de leitura.

|||

D'une certaine façon, il y a une augmentation des ressources qui seront appliquées dans la traduction : l'auteur a utilisé l'image qui doit être considérée par le traducteur comme la clé de la lecture.

- - - - - - - - - -

A partir dos elementos presentes na linguagem dos quadrinhos, o trabalho do tradutor dessas histórias não se restringe ao texto escrito é preciso levar em conta outros desafios que vão desde uma linguagem icônica até as várias normas linguísticas incidentes no texto.

|||

À partir des éléments présents dans la langue de la bande dessinée, le travail du traducteur de ces histoires ne se limite pas au texte écrit, il faut tenir compte d'autres défis allant d'un langage iconique aux différentes normes linguistiques incidentes dans le texte.

- - - - - - - - - -

No gênero quadrinhos, os balões expressam a fala de cada personagem e os autores usam os recursos gráficos para expressar a oralidade na escrita, o que leva o pesquisador (e, por conseguinte, o tradutor) a rever os conceitos de língua falada e língua escrita bem como suas inter-relações.

|||

Dans le genre comique, les ballons expriment le discours de chaque personnage et les auteurs utilisent les ressources graphiques pour exprimer l'oralité par écrit, ce qui conduit le chercheur (et donc le traducteur) à revoir des concepts de langue parlée et de langue écrite en plus de leurs interrelations.

Figure 1 – Bi-texte en PT-FR

6. Titre em portugais : Dissertação de Mestrado de Ana Cláudia Vieira Braga. Norma Linguística e Oralidade Fingida na Tradução de Persépolis

Published in : Antonella Corvaglia and Françoise Wolf-Mandroux (dir.), *Formation linguistique des apprenants allophones et pédagogies innovantes*, Éditions des archives contemporaines, Coll. «PLIDAM», France, ISBN : 9782813003621, doi : https ://doi.org/10.17184/eac.9782813003621

Universidade de Brasília

Biblioteca Central

Serviço de acesso a conteúdos digitais em português para estudantes estrangeiros
Título do Documento Original : Norma linguística e oralidade fingida na tradução de Persépolis
Tese de Mestrado de Ana Cláudia Vieira Braga
Língua do Original : Português
Língua Alvo : Francês
Nome do Arquivo : nancy.these.français.docx

Figure 2 – Exemple de bi-texte avec l'identification de la Bibliothèque Centrale de l'UnB

Parágrafo Fonte

De certa forma, há uma ampliação dos recursos que serão aplicados na tradução : o autor utilizou a imagem que deve ser considerada pelo tradutor como chave de leitura

A partir dos elementos presentes na linguagem dos quadrinhos, o trabalho do tradutor dessas histórias não se restringe ao texto escrito é preciso levar em conta outros desafios que vão desde uma linguagem icónica até as várias normas linguísticas incidentes no texto.

No gênero quadrinhos, os balões expressam a fala de cada personagem e os autores usam os recursos gráficos para expressar a oralidade na escrita, o que leva o pesquisador (e, por conseguinte, o tradutor) a rever os conceitos de língua falada e língua escrita bem como suas inter-relaçoes

Source : élaboré par l'auteur

Paragraphe Cible

D'une certaine façon, il y a une augmentation des ressources qui seront appliquées dans la traduction : l'auteur a utilisé l'image qui doit être considérée par le traducteur comme une clé de lecture.

À partir des éléments présents dans la langue de la bande dessinée, le travail du traducteur de ces histoires ne se limite pas au texte écrit, il faut tenir compte d'autres défis allant d'un langage iconique aux différentes normes linguistiques incidentes dans le texte.

Dans le genre conique, les ballons expriment le discours de chaque personnage et les auteurs utilisent les ressources graphiques pour exprimer l'oralité par écrit, ce qui conduit le chercheur (et donc le traducteur) à examiner les concepts de langue parlée et de langue écrite ainsi que leurs interrelations

Published in : Antonella Corvaglia and Françoise Wolf-Mandroux (dir.), *Formation linguistique des apprenants allophones et pédagogies innovantes*, Éditions des archives contemporaines, Coll. «PLIDAM», France, ISBN : 9782813003621, doi : https ://doi.org/10.17184/eac.9782813003621

Le fluxogramme ci-après (Figure 3) illustre les relations entre chaque composant de l'outil informatique utilisé [7]

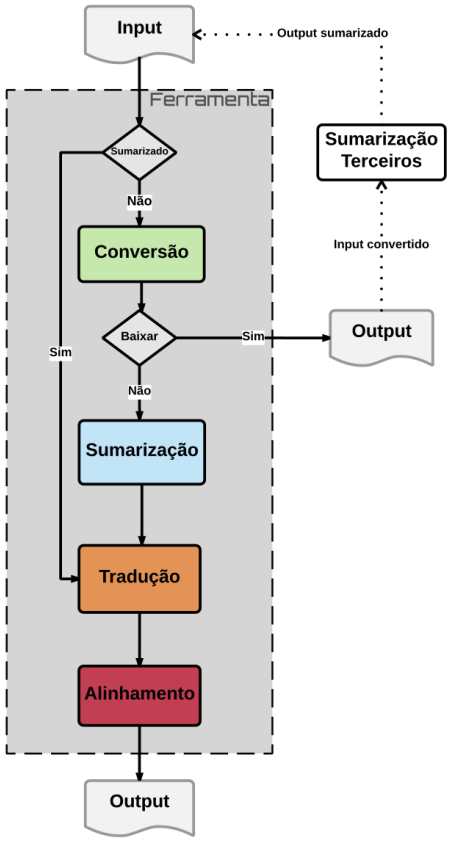


Figure 3 – Fluxogramme des composantes de l'outil informatique

4 Extension de la méthodologie à d'autres groupes/corpus de langues

La recherche a été réalisée avec des étudiants francophones. Cependant, les quatre composantes de la méthodologie proposée peuvent être adaptées à d'autres langues. Naturellement, il sera nécessaire d'utiliser des filtres, des synthétiseurs automatiques, des programmes de traduction et d'alignement de phrases capables de fonctionner avec les langues choisies. L'extension de la méthodologie proposée à d'autres langues est donc un travail d'expansion qui peut être développé dans le cadre d'autres projets similaires.

7. Accessible à `http://multilingua.cdtc.unb.br:8080/login/`

Published in : Antonella Corvaglia and Françoise Wolf-Mandroux (dir.), *Formation linguistique des apprenants allophones et pédagogies innovantes*, Éditions des archives contemporaines, Coll. «PLIDAM», France, ISBN : 9782813003621, doi : https ://doi.org/10.17184/eac.9782813003621

5 Conclusion

Que ce soit dans les bibliothèques ou directement par ses utilisateurs *via* le Web, l'utilisation d'outils de traitement automatique des langues du langage naturel pour médiatiser l'accès à la compréhension du contenu dans une autre langue est encore très embryonnaire. On espère que la proposition méthodologique de cette recherche sera adoptée et améliorée par d'autres groupes de chercheurs intéressés par le sujet, dans le but d'élargir le multilinguisme dans le cyberespace et de promouvoir la vitalité linguistique d'un plus grand nombre de langues dans le monde numérique.

Bibliographie

BOJAR, Ondrej *et al.* "Findings of the 2016 conference on machine translation (wmtl6)". *Proceedings of the First Conference on Machine Translation*, v. 2 : Shared Task Papers, p. 131—198, Berlin, Germany, Aug. 11-12, 2016.

BRANCO, António *et al. The Portuguese Language in the Digital Age.* Berlin, Springer, 2012.

FRIAS-MARTINEZ, Enrique *et al.* "Automated user modeling for personalized digital libraries", *International Journal of Information Management,* v. 26, n. 3, p. 234-248, 2006. doi:10.1016/j.ijinfomgt.2006.02.006

GALE, William A. ; CHURCH, Kenneth W. "A program for aligning sentences in bilingual corpora", *Computational* linguistics, v. 19, n. 1, p. 75-102, 1993.

GAUCH, Susan *et al.* User profiles for personalized information access. In : BRUSILOVSKY, P. ; KOBSA, A. ; NEJDL, W. (Eds.) : *The Adaptive Web,* LNCS 4321, 2007. p. 54-89. http://link.springer.com/chapter/10.1007/978-3-540-72079-9_2 (consulté le 15/ 06/2017). doi:10.1007/978-3-540-72079-9_2

GUPTA, Vishal ; LEHAL, Gurpreet Singh. "A survey of text summarization extractive techniques", *Journal of emerging technologies in Web intelligence*, v. 2, n. 3, p. 258-268, 2010.

GUTT, Ernst-august. Translation and Relevance. 1989. 309 f. Thesis (Ph. D. in Linguistic) – Department of Phonetics and Linguistics, University College London, London, 1989. http://discovery.ucl.ac.uk/1317504/1/241978.pdf. (consulté le 15/11/ 2015)

LAZZARIN, Fabiana Aparecida *et al.* "Da informação à compreensão : reflexões sobre Arquitetura da Informação, Usabilidade e Acessibilidade no campo da Ciência da Informação", Biblioline, João Pessoa, v. 8, n., p. 231-244, jan. 2012.

LEE-JAHNKE, Hanelore ; DELISLE, Jean ; CORNIER, Monique C. (dir.) ; *Terminologia da Tradução,* Traduction et adptation par le portugais par Alvaro Faleiros e Cláudia Xatara — Brasília : Editora Universidade de Brasília, 2013, 172p, ISBN 978-85-230-1071-3.

LEVACOV, Marília. "Tornando a informação disponível : o acesso expandido e a reinvenção da biblioteca", In : MARCONDES, Carlos H. et al (Org.). Bibliotecas digitais : saberes e práticas. 2. ed. Brasília : Ibict, 2006. p. 205-221.

LLORET, Elena *et al.* "Compendium : a text summarisation tool for generating summaries of multiple purposes, domains, and genres", Natural Language Engineering, v. 19, n. 2, p. 147-186, 2013, doi : 10.1017/S1351324912000198

MARCONDES, Carlos H. *et al.* (Org.). *Bibliotecas digitais : saberes e práticas. 2.* ed. Salvador : Ufba, 2006.

MÁRDERO, Arellano Àngel. "Serviços de referência virtual", Ciência da Informação", Brasilia, v. 30, p. 1-15, 2001.

MENEZES, Cláudio ; BAPTISTA, Dulce Maria. "Metodologia de Acesso a Dissertações de Mestrado de Tradução por Estrangeiros : Uma abordagem preliminar", Revista Iberoamericana de Ciência da Informação, Brasilia, v. 10, n. 1, p. 154-163, jan./jul. 2017. http://periodicos.unb.br/index.php/RICI/articIe/view/16462/18074 consulté le 16/10/2017). doi:10.26512/rici.v10.n1.2017.2501

MENEZES, Francisco Cláudio Sampaio de, "O Multilinguismo e as Novas Tecnologias das Línguas no Século XXI", Belas Infiéis, Brasilia, v. 4, n. 12015, p. 85-98, 01 jun. 2015. http://periodicos.unb.br/index.php/belasinfieis/issue/view/1175/showToc (consulté le 15/11/2015) doi : 10.26512/belasinfieis.v4.n1.2015.11314

Published in : Antonella Corvaglia and Françoise Wolf-Mandroux (dir.), *Formation linguistique des apprenants allophones et pédagogies innovantes*, Éditions des archives contemporaines, Coll. «PLIDAM», France, ISBN : 9782813003621, doi : https ://doi.org/10.17184/eac.9782813003621

MIHALCEA, Rada ; TARAU, Paul. *TextRank : Bringing order into texts*. Association for Computational Linguistics. EECS News, jul. 2004. https://Web.eecs.umich.edu/mihalcea/papers/mihalcea.emnlp04.pdf (consulté le 19/06/2017)

PESSOA, Patrícia ; CUNHA, Murilo Bastos da, "Perspectivas dos serviços de referência digital", Informação & Sociedade, João Pessoa, v. 17, n. 3, p. 69-82, set. 2007. https://search.proquest.com/openview/cb039fcb82b6ed3bcf7a646002d559f5/1?pq-origsite=gscholar&cbl=2030753 (consulté le 15/06/2017)

RIBEIRO, Larissa 0. *Sumarização Automática : uma ferramenta para o tratamento automático das línguas naturais*. 2016. 43 f., i1. Trabalho de conclusão de curso (Bacharelado em Línguas Estrangeiras Aplicadas) Universidade de Brasília, Brasília, 2016.

RINO, Lúcia Helena Machado *et al.* "A Comparison of Automatic Summarizers of Texts in Brazilian Portuguese". In : 17TH BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE, 1., 2004, São Luis. Proceedings of the 17th Brazilian Symposium on Artificial Intelligence. São Luís : Springer-verlag 2004. v. 1, p. 235-244. https://www.researchgate.net/pub1ication/220974768 (consulté le 29/11/2004) doi: 10.1007/978-3-540-28645-5*24

SANTOS, Liliane Simões dos. Análise dos softwares Gistsummarizer e Google Tradutor : ferramentas de suporte à sumarização de textos científicos e à tradução automática. 2014. 47 f. Trabalho de conclusão de curso (Bacharelado em Línguas Estrangeiras Aplicadas – LEA/MSI) — Universidade de Brasília, Instituto de Letras. 2014. http://bdm.unb.br/bitstream/10483/9576/1/2014_LilianeSimoesDosSantos.pdf (consulté le 16/11/ 2015)

SILVA, Patrick Pereira. ExtraWeb : um sumarizador de documentos Web baseados em etiquetas HTML e ontologia, 2006. 168 f. Dissertação (Mestrado em Ciências Exatas e da Terra) – Universidade Federal de São Carlos, São Carlos, 2006.

SOUZA, C.F.R. ; NUNES, M.G.V. *Avaliação de Algoritmos de Sumarização Extrativa de Textos em Português*, Relatórios Técnicos do ICMC-USP. NILC-TR- 01-09, Novembre 2001.

UNESCO, *Recommandation sur la Promotion et l'Usage du Multilinguisme et l'Accès Universel au Cyberespace*, http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/official_documents/Fre%20 - (consulté le 05/10/2017)

UNESCO, A Decade in Promoting Multilingualism in Cyberspace, http://unesdoc.unesco.org/images/0023/002327/232743e.pdf (consulté le 05/11/2017)

Published in : Antonella Corvaglia and Françoise Wolf-Mandroux (dir.), *Formation linguistique des apprenants allophones et pédagogies innovantes*, Éditions des archives contemporaines, Coll. «PLIDAM», France, ISBN : 9782813003621, doi : https ://doi.org/10.17184/eac.9782813003621