META NET

White Paper Series Collection de Livres Blancs

THE FRENCH LANGUAGE IN THE DIGITAL AGE

LA LANGUE FRANÇAISE À L'ÈRE DU NUMÉRIQUE

Joseph Mariani
Patrick Paroubek
Gil Francopoulo
Aurélien Max
François Yvon
Pierre Zweigenbaum

Springer

White Paper Series Collection de Livres Blancs

THE FRENCH LANGUAGE IN THE DIGITAL AGE

LA LANGUE FRANÇAISE À L'ÈRE DU NUMÉRIQUE

Joseph Mariani IMMI-CNRS & LIMSI-CNRS
Patrick Paroubek LIMSI-CNRS
Gil Francopoulo IMMI-CNRS & TAGMATICA
Aurélien Max LIMSI-CNRS & U. Paris Sud 11
François Yvon LIMSI-CNRS & U. Paris Sud 11
Pierre Zweigenbaum LIMSI-CNRS

Georg Rehm, Hans Uszkoreit
(Éditeurs, editors)

Editors
Georg Rehm
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: georg.rehm@dfki.de

Hans Uszkoreit
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: hans.uszkoreit@dfki.de

ISSN 2194-1416 ISSN 2194-1424 (electronic)
ISBN 978-3-642-30760-7 ISBN 978-3-642-30761-4 (eBook)
DOI 10.1007/978-3-642-30761-4
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012946759

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

PRÉFACE PREFACE

Ce livre blanc fait partie d'une collection qui a pour objectif de faire connaître le potentiel des technologies de la langue. Il s'adresse en particulier aux journalistes, politiciens, communautés linguistiques, enseignants mais aussi à tous. La disponibilité et l'utilisation des technologies de la langue variant grandement d'une langue à l'autre en Europe, les actions nécessaires au soutien des activités de recherche et développement peuvent être très différentes en fonction des langues selon des facteurs multiples, comme leur complexité intrinsèque ou leur nombre de locuteurs.

Le Réseau d'Excellence META-NET, financé par la Commission Européenne, a analysé l'état des ressources et des technologies de la langue dans cette collection de livres blancs (p. 95). Cette étude, qui a concerné les 23 langues officielles de l'Union Européenne ainsi que des langues nationales et régionales de l'Europe, montre qu'il y a des déficits énormes en termes de soutien technologique et des lacunes significatives en recherche selon les langues. L'analyse de la situation actuelle permettra de maximiser l'impact des futures recherches.

En mars 2012, META-NET regroupe 54 centres de recherche de 33 pays européens (p. 91) et travaille avec les acteurs de l'économie (sociétés informatiques, fournisseurs de technologies, utilisateurs), les agences gouvernementales et non gouvernementales, les organismes de recherche, les communautés linguistiques et les universités européennes, pour établir une vision commune des technologies de la langue et un échéancier stratégique de recherche dans la vision de l'Europe multilingue de 2020.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others. The availability and use of language technology in Europe vary between languages. Consequently, the actions that are required to further support research and development of language technologies also differ. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series (p. 95). The analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of additional research.

As of March 2012, META-NET consists of 54 research centres from 33 European countries (p. 91). META-NET is working with stakeholders from economy (software companies, technology providers, users), government agencies, research organizations, non-governmental organizations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

Les auteurs de ce document remercient les auteurs du Livre Blanc allemand de leur avoir permis de réutiliser des éléments génériques de leur document [1].

La production de ce Livre Blanc a été financée par le septième Programme-Cadre et le Programme d'appui stratégique en Technologies de l'Information et de la Communication (TIC) de la Commission Européenne dans le cadre des contrats T4ME (convention d'aide 249 119), CESAR (convention d'aide 271 022), METANET4U (convention d'aide 270 893) et META-NORD (convention d'aide 270 899).

The authors of this document are grateful to the authors of the White Paper on German for permission to re-use selected language-independent materials from their document [1].

The development of this White Paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249 119), CESAR (Grant Agreement 271 022), METANET4U (Grant Agreement 270 893) and META-NORD (Grant Agreement 270 899).

TABLE DES MATIÈRES TABLE OF CONTENTS

LA LANGUE FRANÇAISE À L'ÈRE DU NUMÉRIQUE

THE FRENCH LANGUAGE IN THE DIGITAL AGE

RÉSUMÉ EXÉCUTIF

Le multilinguisme est une donnée essentielle de la construction Européenne. Il est primordial d'assurer à chaque citoyen européen la possibilité d'utiliser sa langue maternelle et à chaque Etat européen la capacité de préserver sa culture, tout comme il est essentiel de permettre la communication entre les citoyens pour franchir la barrière des langues dans l'espace informationnel ou commercial communautaire. Ce même besoin existe de fait à l'échelle de la planète.

Peut-on accepter de voir disparaître des langues européennes, et les cultures dont elles font partie ? Du seul fait de la barrière des langues, peut-on accepter de se borner à constater la faiblesse de la croissance du marché européen ? De ne pas avoir accès à la richesse culturelle des autres pays ? De ne pas connaître à leur source les informations qui forgent l'Europe ?

Le multilinguisme a un coût, important, qui fait que progressivement les langues disparaissent au profit des langues majoritaires. Sur les quelques 6500 langues qui existent sur la planète, il est estimé que la moitié auront disparu à la fin de ce siècle. De nombreuses langues européennes ont déjà disparu, ou ont failli disparaître et n'ont été sauvées que grâce à une volonté politique.

Comment traiter les 48 heures de vidéos qui arrivent toutes les minutes sur YouTube, dans toutes les langues ? Comment faire en sorte que les brevets européens soient accessibles pour les entreprises européennes autres que celles qui parlent anglais, français ou allemand ? Comment permettre à un enseignant de faire un cours à des élèves qui ne parlent pas sa langue ? A un chercheur de ne pas avoir à rédiger ses articles dans une langue unique, en délaissant la sienne ? Comment faire en sorte qu'une langue continue de s'enrichir de termes nouveaux au rythme de l'accroissement des connaissances ? Comment éviter que sa langue maternelle soit juste bonne à commander un café, mais que l'on doive passer à une autre langue pour suivre un cours dans l'amphithéâtre d'une université ?

L'arrivée des technologies du numérique, et des technologies de la langue en particulier, change la donne. La toile électronique facilite la production et la consultation des contenus d'information et de connaissance pour tous. Wikipedia existe dans 300 langues environ. Les réseaux sociaux impliquent l'utilisation des langues de chacun. Facebook existe dans 80 langues, et Twitter dans une vingtaine. Les progrès scientifiques ont conduit à la réalisation et à la diffusion de technologies de la langue, moteurs de recherche, systèmes de reconnaissance et synthèse vocales, traduction automatique et traduction vocale,... pour un nombre croissant de langues. Ainsi Google Translate fonctionne pour une soixantaine de langues, dont une vingtaine sur support vocal, Apple Siri pour quatre langues, Jibbigo, système de traduction vocale embarqué, pour une dizaine. Cependant ces technologies ne sont disponibles, de plus à des degrés très variables de qualité et donc d'utilisabilité, que pour une soixantaine de langues, soit 1% des langues parlées dans le monde. De nouveaux systèmes apportent des fonctionnalités plus avancées, comme le système IBM Watson de réponses aux questions qui a remporté le jeu télévisé Jeopardy aux Etats-Unis en 2011, mais qui ne fonctionne que pour la langue anglaise

alors que la connaissance humaine ne saurait se réduire à celle qui a été codée dans une seule langue, et qui est le reflet d'une seule culture.

L'apport de ces technologies diminue le coût que représente le multilinguisme et, ainsi, le permet. C'est même la seule façon de le permettre. Et ce faisant, certaines d'entre elles, comme les systèmes de sous-titrage automatique avec traduction ou les correcteurs orthographiques, facilitent aussi l'apprentissage des langues.

Mais peut-on accepter que, dans le meilleur des cas, ces technologies nous soient fournies par des entreprises américaines au prix d'une gratuité qui pourrait un jour nous coûter très cher du fait de la perte de notre indépendance et de notre souveraineté ? Comment comprendre qu'une communauté d'Etats qui aimeraient pouvoir partager la richesse de leurs cultures et qui constatent que la barrière linguistique est un obstacle à leurs échanges, n'investissent pas, ne s'unissent pas, pour valoriser cette richesse et surmonter cet obstacle, sauf à penser qu'ils ne traitent pas les questions essentielles à leur union ?

Convaincre de la nécessité de développer ces technologies est cependant chose difficile. Aucun grand groupe industriel ne mettra le multilinguisme au premier rang de ses priorités, que ce soit dans les secteurs de l'automobile, de l'aéronautique, des télécommunications, de l'électronique grand public, de l'informatique, du médical ou de l'audiovisuel. Mais chacun de ces secteurs en a besoin à divers titres, et c'est la somme de ces petites priorités qui est, elle, très importante, et fait du multilinguisme une priorité majeure. Mais qui va la calculer ? Qui va l'expliquer ? Qui va réunir les acteurs pour la porter ? Seule une volonté politique communautaire peut le faire et montrer que les technologies de la langue ne sont pas qu'un thème de recherche et développement parmi d'autres, ne sont pas que des données noyées dans beaucoup d'autres, mais qu'elles sont un élément essentiel de la construction européenne, partagé par la plupart des secteurs de la Commission et par la totalité des Etats Membres.

META-NET, l'Alliance Technologique pour une Europe multilingue, est un réseau d'excellence soutenu par la Commission Européenne. Il comprend actuellement plus de 50 laboratoires de recherche parmi les meilleurs dans le domaine des sciences et technologies de la langue, dans une trentaine de pays. Il a pris l'initiative de rédiger un ensemble de Livres Blancs sur chacune des langues de ces pays, chacun rédigé dans la langue correspondante et en anglais.

La langue française est une grande langue internationale, avec une estimation de 220 millions de locuteurs de par le monde, auxquels il faut ajouter plus de 100 millions d'apprenants. Elle est une des langues officielles de l'Union Européenne et d'une trentaine de pays, ainsi que de grandes organisations internationales. Elle a longtemps figuré comme la langue préférée pour la diplomatie ou la culture, mais l'anglais l'a progressivement remplacée dans tous ces rôles. Elle est très présente sur l'internet, où elle figure au huitième rang des langues pratiquées par les internautes, devancée parmi les langues européennes par l'anglais, mais aussi par l'espagnol, le portugais et l'allemand. Langue du savoir, elle apparaît au troisième rang des langues de Wikipédia, derrière l'anglais et l'allemand. D'autres langues régionales, plus d'une soixantaine, sont également parlées en France métropolitaine comme dans les territoires d'outre-mer.

Il existe des technologies de la langue pour le traitement automatique du français, que cela concerne la langue écrite ou parlée, ou encore la langue des signes pour les malentendants. Elles regroupent les correcteurs de texte, les moteurs de recherche sur la toile, les systèmes de réponse aux questions, la reconnaissance et la synthèse automatique de la parole, le dialogue oral, la traduction automatique et la traduction vocale, mais aussi la reconnaissance du locuteur ou de la langue parlée, l'extraction d'information ou le résumé automatique.

La recherche française a bénéficié de programmes dans ce domaine, comme le programme francophone des industries de la langue (FRANCIL) de l'Association des Universités Francophones (AUF), ou le programme TechnoLangue soutenu par plusieurs ministères. Aujourd'hui, le grand programme franco-allemand Quaero sur le traitement des documents multilingues et multimédias rassemble une trentaine de partenaires industriels et académiques autour de huit projets applicatifs et du développement d'une trentaine de technologies de traitement de la langue écrite et parlée, de l'image, de la vidéo et de la musique. Il est entièrement structuré autour de l'évaluation systématique des progrès des technologies, et de la production des données nécessaires au développement et au test de ces technologies.

Tous ces projets ont permis d'investir pour produire les données nécessaires au développement des technologies pour la langue française. Cela lui permet de se placer à une excellente place dans le concert des langues européennes disposant de technologies, au sein d'un peloton qui rassemble l'allemand, l'espagnol, l'italien et le néerlandais, mais qui se trouve loin derrière l'anglais, aucune langue ne disposant par ailleurs encore de l'éventail complet des technologies de la langue à un niveau de qualité suffisant, ni des données permettant de les développer.

Les campagnes d'évaluation internationales montrent de manière objective et quantitative que les laboratoires de recherche français et les technologies qu'ils développent se situent parmi les meilleurs au monde.

Les entreprises françaises tout comme les entreprises européennes sont cependant pour la quasi-totalité des PME qui ont bien du mal à rivaliser avec les géants américains que sont Google, Apple, IBM, Microsoft ou Nuance, qui ont investi massivement dans ces technologies. Et paradoxalement, beaucoup des chercheurs de ces sociétés américaines ont été formés dans les laboratoires de recherche européens.

La situation est semblable dans les autres grands pays industrialisés où la langue française est très pratiquée, Belgique, Suisse ou Canada.

Le financement de la recherche et de l'innovation sur les technologies de la langue manque de continuité, avec des programmes coordonnés de courte durée interrompus par des périodes de financement faible ou épars, et la coordination est manquante avec les programmes existant dans d'autres Etats de l'Union Européenne ou à la Commission Européenne, alors que ce thème de recherche semble idéalement placé pour faire l'objet d'un effort transnational partagé. La situation est similaire à la Commission où la priorité accordée à ce domaine fluctue au fil des ans, et où il bénéficie tour à tour d'une attention particulière, avec un Commissaire, une Unité et une ligne de programme attitrés, puis se trouve noyé dans des agglomérats de différentes natures alors que sa spécificité dans la construction européenne est pourtant clairement identifiée.

Une directive européenne comme il en existe pour l'accès des handicapés à l'information, exprimant l'importance de lever la barrière des langues et stipulant que tout citoyen européen, quelle que soit la langue qu'il parle, doit pouvoir avoir accès à toute information produite dans l'Union Européenne, livre, journal, émission de télévision ou de radio, film, etc. quelle que soit la langue dans laquelle elle a été produite, donnerait une impulsion déterminante à ce secteur.

Un grand programme coordonné sur les Technologies de la Langue dans le cadre du prochain programme européen pour la recherche et l'innovation permettrait le multilinguisme et aiderait à sauver la langue française, dans toutes ses dimensions, tout comme les autres langues, nationales et régionales, et à faciliter les échanges culturels et commerciaux, en Europe et ailleurs.