Benoît SAGOT

Informations personnelles

E-mail benoit.sagot@inria.fr Site web http://pauillac.inria.fr/~sagot/

Expérience professionnelle

Recherche

2017– **Responsable de l'équipe-projet ALMAnaCH** (Automatic Language Modelling and Analysis & Computational Humanities), *Équipe-projet du Centre Inria de Paris spécialisée en traitement automatique des langues (TAL) et en humanités numériques (HN)*, Paris, France.
Directeur de Recherche Inria (2020–), précédemment Chargé de Recherche Inria
Titulaire d'une chaire dans l'institut PRAIRIE (2019–)

2014–2016 **Responsable de l'équipe-projet ALPAGE** (Analyse Linguistique Profonde à Grande Échelle), *équipe-projet commune entre le Centre Inria de Rocquencourt puis Paris et l'Université Paris-Diderot, spécialisée en TAL*, Paris et Rocquencourt, France.
Chargé de Recherche Inria (1ère classe)

2007–2013 **Chargé de Recherches Inria (2ème puis 1ère classe) au sein de l'équipe-projet ALPAGE**, Paris et Rocquencourt, France.

2006–2007 **Chargé de Recherches Inria (2ème classe)**, *Signes (Signes linguistiques, grammaire et sens : algorithmique logique de la langue), équipe Inria au sein de l'UMR LabRI*, Bordeaux, France.

2002–2006 **Ingénieur du Corps des Télécommunications en détachement à Inria**, *Atoll (Atelier d'outils logiciels pour le langage naturel), équipe-projet Inria*, Rocquencourt, France, En détachement en vue de la réalisation de ma thèse de doctorat, sous la direction de Laurence Danlos (Université Paris-Diderot).

2002 (4 mois) **Stage de DEA**, *Institut Jean-Nicod*, Paris, France, Stage en linguistique cognitive : réflexion théorique et étude comparative multilingue de quelques types de relations spatiales, sous la direction de Richard Carter (Institut Jean Nicod, EHESS).

2001 (6 mois) **Stage d'ingénieur-élève de l'École Polytechnique**, *IBM France*, Paris, France, Stage en traitement automatique des langues : réalisation d'une application de compréhension automatique du langage, sous la responsabilité de Claire Waast-Richard.

2000 (4 mois) **Stage de recherche**, *Johns Hopkins University*, Baltimore, États-Unis, Stage en astrophysique dans l'équipe scientifique du satellite FUSE de la NASA (astronomie UV).

Enseignement

2017– **Master 2**, *ENS Cachan*, Paris, France.
Cours de traitement automatique des langues et de la parole, avec Emmanuel Dupoux (en anglais), au sein du Master MVA (Mathématiques, Vision et Apprentissage (24h/an)

2009–2015 **Master 2**, *Université Paris-Diderot*, Paris, France, Cours d'analyse syntaxique du langage naturel dans le Master 2 de linguistique informatique (7 années, 24h/an).

2010–2011 **Licence 3**, *Université Paris-Diderot*, Paris, France, Cours d'introduction au traitement informatique des langues en Licence 3 d'informatique (2 années, 24h/an).

1998–2000 **Classes préparatoires (MP*)**, *École Sainte-Geneviève*, Versailles, France, Interrogations orales de mathématiques (2 années).

Diplômes et titres

2018 **Habilitation à diriger les recherches en informatique**, *Sorbonne Université*, Paris, France,
« Informatiser le lexique : Modélisation, développement et exploitation de lexiques morphologiques, syntaxiques et sémantiques »
Parrain : Laurent Romary (Inria, France).
Autres membres du jury : Philippe Blache (CNRS, rapporteur), James P. Blevins (University of Cambridge, rapporteur), Christiane D. Fellbaum (Princeton University), Ludovic Denoyer (Sorbonne Université), Anna Korhonen (University of Cambridge), Gertjan Van Noord (University of Groningen)

2006 **Doctorat en Informatique**, *Université Paris-Diderot (Paris 7)*, Paris, France,
« Analyse automatique du français : lexiques, formalismes, analyseurs », Mention très honorable avec les félicitations du jury.
Direction : Laurence Danlos (Professeur des Universités, Université Paris-Diderot) ; Co-direction : Éric Villemonte de La Clergerie (Chargé de Recherches, Inria)
Autres membres du jury : Philippe Blache (CNRS, rapporteur), Gérard Huet (Inria, rapporteur), John Carroll (University of Sussex), Pierre Boullier (Inria, invité)

2002 **DEA IARFA (Intelligence Artificielle, Reconnaissance de Formes et Applications)**, *Université Pierre et Marie Curie*, Paris, France.

2002 **Ingénieur du Corps des Télécommunications**, *Télécom ParisTech*, Paris, France, Spécialisation en informatique, algorithmique et traitement automatique des langues.

2000 **Ingénieur de l'École polytechnique**, *École polytechnique*, Palaiseau, France.

Activités de recherche

Responsabilités dans des projets de recherche

2022–	**Projet ANR CulturIA**	Responsable pour ALMAnaCH *Vers une histoire culturelle de l'intelligence artificielle*
2021–2022	**Initiative BigScience**	Co-responsable du working group sur la tokenisation *Développement d'un très grand modèle de langue génératif multilingue ouvert*
2019–2024	**Institut 3IA PRAIRIE (PaRis AI Research InstitutE)**	Titulaire d'une chaire (2019–2024) *Institut de recherche en intelligence artificielle dirigé par Isabelle Ryl*
2011–2019	**LabEx EFL (Empirical Foundations of Linguistics)**	Responsable (2011–2015) puis responsable adjoint (depuis 2015) de l'un des 7 axes de recherche (axe 6 « Ressources Linguistiques »), membre de 2011 à 2015 puis membre suppléant (depuis 2015) du Comité Scientifique Restreint, responsable de plusieurs opérations de recherche *LabEx dirigé par Jacqueline Vaissière (Université Paris 3) puis Christian Puech (Université Paris 3)*
2016–	**Projet ANR Profiterole**	Responsable de tâche *Développement de ressources linguistique et d'outils de traitement automatique pour différents états ancien de la langue française, notamment pour en modéliser l'évolution au cours des siècles*

2015–2017	**Projet FUI VerDI**	Responsable pour ALPAGE puis ALMAnaCH *Identification automatisée des dissimulations d'information, notamment dans les contenus journalistiques en ligne*
2012–2016	**Projet ANR ASFALDA**	Co-responsable de tâche *Développement d'un FrameNet du français et d'outils associé, porté par Marie Candito (Université Paris-Diderot)*
2011–2015	**Projet FUI PACTE**	Responsable pour ALPAGE *Correction automatique à grande échelle de sorties de systèmes de reconnaissance optique de caractères, porté par l'entreprise Numen Digital*
2010–2013	**Projet ANR EDyLex**	**Porteur du projet** *Enrichissement dynamique de ressources lexicales*
2010–2012	**Projet bi-latéral PROTEUS (France-Slovénie)**	**Co-porteur** *Développement et exploitation de corpus parallèles et comparables français-slovène pour le développement de wordnets. Co-dirigé avec Mojca Schalmberger Brezar (Université de Ljubljana) pour la partie Slovène*
2009–2011	**Projet ANR SEQUOIA**	Responsable pour ALPAGE *Analyse syntaxique probabiliste du français, projet porté par Alexis Nasr (Université de la Méditerrannée)*

Participation à des projets de recherche

2017–2023	**Projet ANR BASNAGE**	*Numérisation et exploitation automatique du* Dictionnaire Universel *de Basnage (1701). PI : Laurent Romary (Inria)*
2016–2022	**Projet ANR PARSITI**	*Analyse et traduction automatique de données textuelles bruitées telles que publiées sur l'internet et notamment sur les réseaux sociaux. PI : Djamé Seddah (Inria)*
2016–2021	**Projet franco-américain ANR-NSF MCM-NL**	*Exploration des corrélations entre données de neuro-imagerie (IRMf, EEG) et systèmes de traitement automatique des langues (analyseurs syntaxiques), à partir de données du* Petit prince, *lu en anglais et en français. PI : John Hale (Cornell University)*
2015–2021	**Projet ANR SoSweet**	*Analyse automatique de données textuelles bruitées issues de Twitter et analyses sociolinguistiques en lien avec la structure de graphe du réseau d'utilisateurs. PI : Jean-Philippe Magué (ENS Lyon)*
2012–	**Consortiums Corpus Écrits puis CORLI du TGIR Huma-Num**	Membre du comité de pilotage, participant dans différents groupes de travail. En particulier, participation au projet CoMéRé (Corpus Médié par les Réseaux), placé sous la responsabilité de Thierry Chanier (Université Blaise-Pascal) *Projets fédérateurs portés par l'Institut de Linguistique Française, sous la responsabilité de Franck Neveu (Université Paris-Sorbonne)*
2012–2014	**Projet IARPA "BABEL Program" (États-Unis)**	Expert externe *Construction rapide de systèmes de recherche d'informations dans des données de parole pour n'importe quelle langue. Expertise dans le cadre de la tâche d'analyse morphologique non-supervisée (projet* Lorelei*) placé sous la responsabilité de Owen Rambow et Nizar Habash (Columbia University)*
2009–2011	**Projet ANR/DFG franco-allemand PerGram**	*Développement de descriptions linguistiques et de ressources (lexique, grammaire) pour le persan, projet porté conjointement par Pollet Samvelian (Université Paris 3) et Stefan Müller (Freie Universität Berlin)*

2007–2009	**Projet ANR PASSAGE**	*Développement et exploitation automatiques de corpus arborés à grande échelle, porté par Éric de La Clergerie (Inria)*
2005–2007	**Projet ILF LexSynt**	*Projet sur les lexiques syntaxiques pour le français, porté par Sylvain Kahane (Université Paris X)*
2004, 2007	**Projet Technolangue EASy**	*Évaluation des analyseurs syntaxiques du français. Participant à la campagne.*

Séjours de recherche

2013	**Columbia University, États-Unis**	*Séjour d'une semaine dans le cadre d'une collaboration suivie avec Owen Rambow et Nizar Habash* Morphologie formelle et computationnelle, dans le cadre du projet IARPA "BABEL Program" (cf. ci-dessus)
2008–2012	**Université de Ljubljana, Slovénie**	*Multiples séjours d'une semaine dans le cadre d'une collaboration suivie avec Darja Fišer* Développement de wordnets pour le français et le slovène, pour partie dans le cadre du projet bi-latéral PROTEUS mentionné ci-dessus
2007–2011	**Université de Vigo, Espagne, et Université de Nice, France**	*Trois séjours de quelques jours dans le cadre d'une collaboration suivie* Développement de ressources lexicales pour le français, l'espagnol et le galicien, en relation avec le projet galicien Victoria
2009	**Université de Padoue, Italie**	*Séjour d'une semaine dans le cadre d'une collaboration suivie avec Giorgio Satta* Propriétés formelles des langages faiblement sensibles au contexte
2007	**Académine polonaise des sciences, Varsovie, Pologne**	*Séjour de recherche de deux mois* Morphologie computationnelle du polonais, développement et extension de lexiques morphologiques, amélioration du corpus national polonais

Activités d'encadrement

Direction et co-direction de thèses

2023–	**You Zuo**	*Approches neuronales pour le traitement autoamtique des brevets* Thèse CIFRE en partenariat avec la start-up qatent *Encadrant en entreprise : Kim Gerdes*
2023–	**Wissam Antoun**	*Détection de contenus produits par des modèles de langue génératifs* *Encadrant principal : Djamé Seddah (ALMAnaCH)*
2021–	**Nathan Godey**	*Approches adversariales pour les modèles de langue* Thèse financée par la chaire PRAIRIE de B. Sagot *Autre encadrant : Éric Villemonte de La Clergerie (ALMAnaCH)*
2021–	**Lydia Nishimwe**	*Traduction automatique neuronale robuste* Thèse financée par la chaire PRAIRIE de R. Bawden *Autre encadrant : Rachel Bawden (ALMAnaCH)*
2021–	**Roman Castagné**	*Modèles de langues neuronaux informés linguistiquement* Thèse financée par la chaire PRAIRIE de B. Sagot *Autre encadrant : Éric Villemonte de La Clergerie (ALMAnaCH)*

2021– **Matthieu Futeral-Peter**
Modèles multimodaux texte-image
Thèse financée pour partie par Inria et pour partie par les chaires PRAIRIE de B. Sagot, C. Schmid et I. Laptev
Directeur de thèse : Ivan Laptev (WILLOW). Autres encadrants : Rachel Bawden (ALMAnaCH), Cordelia Schmid (WILLOW)

2021– **Paul-Ambroise Duquenne**
Study of vector spaces for sentence representation
Thèse CIFRE en partenariat avec le META Artificial Intelligence lab
Encadrant en entreprise : Holger Schwenk (META AI Paris)

2021– **Tu Anh Nguyen**
Acquisition non-supervisée de représentations linguistiques à partir de données de parole
Thèse CIFRE en partenariat avec le META Artificial Intelligence lab
Encadrant en entreprise : Emmanuel Dupoux (META AI Paris)

2019– **Robin Algayres**
Segmentation non-supervisée de données de parole brutes
Thèse financée par une bourse de l'école doctorale ED3C
Directeur de thèse : Emmanuel Dupoux (CoML).

2019–2022 **Clémentine Fourrier**
Neural models of language evolution
Thèse financée par Inria
Soutenance le 26 septembre 2022

2018–2022 **Benjamin Muller**
Traitement de la variation et de la diversité linguistiques dans les modèles de langue neuronaux
PhD thesis funded by the SoSweet and ParSiTi national (ANR) projects
Autre encadrant : Djamé Seddah (ALMAnaCH)
Soutenance le 17 novembre 2022

2018–2022 **Pedro Javier Ortiz Suárez**
Approches neuronales pour l'extraction d'information à partir de dictionnaires anciens
Thèse financée par le projet ANR BASNUM
Directeur de thèse : Laurent Romary (ALMAnaCH)
Soutenance le 27 juin 2022

2018–2021 **Louis Martin**
Simplification Automatique de Textes
Thèse CIFRE en partenariat avec le Facebook Artificial Intelligence Research lab
Thèse dirigée par Laurent Romary (ALMAnaCH) ; autre co-directeur : Éric Villemonte de La Clergerie.
Soutenance le 27 octobre 2021

2012–2015 **Marion Baranes**
Normalisation de textes bruités pour la fouille d'opinions
Thèse industrielle au sein de l'entreprise viavoo, en partenariat avec ALPAGE
Thèse dirigée par Laurence Danlos (Université Paris-Diderot).
Soutenance le 23 octobre 2015, mention très honorable

2011–2015 **Valérie Hanoka**
Construction et extension semi-automatique de réseaux lexicaux multilingues
Thèse CIFRE en partenariat entre ALPAGE et l'entreprise Verbatim Analysis – Vera
Thèse dirigée par Laurence Danlos (Université Paris-Diderot).
Soutenance le 6 juillet 2015, mention très honorable

2010-2013 **Pierre Magistry**
Unsupervised Word Segmentation and Wordhood Assessment. The case for Mandarin Chinese
Thèse financée par une allocation ministérielle
Thèse dirigée par Sylvain Kahane (Université Paris X) ; autre co-directeur : Marie-Claude Paris (Université Paris-Diderot).
Soutenance le 19 décembre 2013, mention très honorable avec les félicitations du jury

2009–2013 **Rosa Stern**
Identification automatique d'entités pour l'enrichissement de contenus textuels
Thèse CIFRE en partenariat entre ALPAGE et l'Agence France-Presse
Thèse dirigée par Laurence Danlos (Université Paris-Diderot).
Soutenance le 28 juin 2013, mention très honorable

Encadrement de post-docs

2019–2020 **Gaël Guibon**
Modélisation du lexique de l'ancien français
Projet ANR PROFITEROLE

2016–2017 **Hector Martínez Alonso**
Étiquetage morphosyntaxique neuronal ; identification automatique d'omissions dans les textes jouralistiques
Projet ANR-RAPID VerDI

2013–2015 **Kata Gábor**
Identification et correction d'entités nommées spécifiques au domaine dans les sorties de systèmes de reconnaissance automatique de caractères (OCR)
Projet FUI PACTE

2012–2013 **Damien Nouvel**
Extension dynamique de ressources lexicales à partir de corpus dynamiques (dépêches d'agence)
Projet ANR EDyLex

2012–2013 **Yves Scherrer**
Développement de lexique et étiquetage morphosyntaxique pour des langues non dotées à partir de ressources pour des langues proches
LabEx EFL

2010–2011 **Yayoi Nakamura–Delloye**
Extraction automatique de relations entre entités nommées
Projet ANR EDyLex

2010–2011 **Marianna Apidianaki**
Extension automatique de lexiques sémantiques (wordnets)
Projet ANR EDyLex

2009–2011 **Sattisvar Tandabany**
Algorithmique de l'analyse syntaxique non contextuelle
Projet ANR SEQUOIA

Encadrement d'ingénieurs de recherche

2022– **Anna Chepaikina**
Génération automatique de commentaires œnologiques
Partenariat bilatéral avec la start-up Winespace

2022–2023 **Wissam Antoun**
Détection de discours haineux sur les réseaux sociaux. Encadrant principal : Djamé Seddah
Financé par la chaire PRAIRIE de B. Sagot et par le projet européen H2020 CounteR

2022– **Rua Ismail**
Identification de la langue pour le développement de corpus bruts multilingues
Financé par la chaire PRAIRIE de B. Sagot

2021–2022 **You Zuo**
Classification fine de brevets. Co-encadrant : Kim Gerdes
Partenariat bilatéral avec l'INPI

2021 **Thomas Wang**
Modèles de langue neuronaux avec attention linéaire
Financé par la chaire PRAIRIE de B. Sagot

2021 **Arij Riabi**
TAL pour les variétés non-standard à faibles ressources, notamment l'arabe dialectal d'Afrique du Nord écrit en alphabet latin. Encadrant principal : Djamé Seddah
Financé par la chaire PRAIRIE de B. Sagot et par le projet européen H2020 CounteR

2021– **Julien Abadji**
Développement et extension de corpus multilingues de grande taille (le corpus OSCAR). Co-encadrant : Pedro Ortiz Suarez
Financé par la chaire PRAIRIE de B. Sagot

2021–2022 **Thibault Charmet** — *Détection de similarités dans les arrêts de la Cour de cassation. Co-encadrante : Rachel Bawden*
Collaboration financée par le gouvernement (LabIA)

Encadrement de mémoires de Master 2

2021–2022 **Camille Rey** — *Master Inalco en NLP. Co-encadré avec Rachel Bawden.* Améliorer la désambiguïsation lexicale en traduction automatique neuronale.

2021 **Roman Castagné** — *Master "Mathématiques, Vision, Apprentissage" (MVA) à l'École Normale Supérieure.* Quelle tokenisation pour les modèles de langue ?

2021 **Matthieu Futeral-Peter** — *Master "Mathématiques, Vision, Apprentissage" (MVA) à l'École Normale Supérieure. Co-encadré avec Cordelia Schmid (WILLOW), Rachel Bawden (ALMAnaCH) et Ivan Laptev (WILLOW).* Exploration des word embeddings multilingues et multimodaux

2019 **Charlotte Rochereau** — *Master en Sciences Cognitives de l'École Normale Supérieure (CogMaster). Co-encadré avec Emmanuel Dupoux (CoML).* Comparaison de jugements humains et de scores produits par des modèles de langue sur l'ordre des mots en allemand.

2014 **Sarah Beniamine** — *Master en linguistique informatique de l'Université Paris-Diderot* Vers un traitement linguistiquement motivé des unités multi-tokens du français dans l'analyse syntaxique en constituants

Participation à des jurys de thèse et d'HDR

2022 **Léo Laugier** — *Président du jury*
Doctorat en informatique, Télécom Paris, France
Analysis and Control of Online Interactions through Neural Natural Language Processing
Directeurs de thèse : Thomas Bonald (Télécom Paris), Lucas Dixon (Google)

2021 **Hicham El Boukkouri** — *Rapporteur*
Doctorat en informatique, Université Paris-Saclay, France
Domain Adaptation of Word Embeddings Through the Exploitation of In-domain Corpora and Knowledge Bases
Directeurs de thèse : Pierre Zweigenbaum, Olivier Ferret, Thomas Lavergne

2021 **Adelle Abdallah** — *Rapporteur*
Doctorat en informatique, Université Paris 8 Vincennes à Saint-Denis, France & Université Libanaise, Lebanon
Catégorisation sémantique et information grammaticale en arabe
Directeurs de thèse : Gilles Bernard, Mohammad Hajjar

2020 **Silvia García Mendéz** — *Rapporteur*
Doctorat en informatique, Universidad de Vigo, Spain
Contribution to Natural Language Generation for Spanish
Directeurs de thèse : Enrique Costa Montenegro, Milagros Fernández Gavilanes

2020 **Alice Millour** — *Rapporteur ("rapporteur")*
Doctorat en informatique, Sorbonne Université, France
Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées
Directeurs de thèse : Claude Montacié (Sorbonne Université), Karën Fort (Sorbonne Université)

2020 **Jack Bowers** — *Examinateur*
Doctorat en linguistique, EPHE, France
Language Documentation and Standards in Digital Humanities : TEI and the documentation of Mixtepec-Mixtec
Directeur de thèse : Laurent Romary (Inria)

2020 **Jacobo Levy Abitbol**
Rapporteur ("rapporteur")
Doctorat en informatique, Université de Lyon, France
Computational detection of socioeconomic inequalities
Directeurs de thèse : Éric Fleury (Inria), Márton Karsai (Central European University & Inria)

2019 **Hazem Al Saied**
Examinateur
Doctorat en informatique, Université de Lorraine, France
Analyse automatique par transitions pour l'identification des expressions polylexicales
PhD supervisors : Matthieu Constant (Université de Lorraine), Marie Candito (Université Paris-Diderot)

2019 **Tamara Álvarez López**
Rapporteur ("rapporteur")
Doctorat en informatique, Université de Vigo, Espagne
Análisis de Opinión en Redes Sociales mediante Técnicas de Procesamiento del Lenguaje Natural
Directeurs de thèse : Enrique Costa Montenegro, Milagros Fernández Gavilanes (University of Vigo)

2019 **Nourredine Alliane**
Rapporteur ("rapporteur")
Doctorat en informatique, Université Paris 8, France
Evaluation des représentations vectorielles de mots
Directeur de thèse : Gilles Bernard (Université Paris 8)

2018 **Sébastien Delecraz**
Rapporteur ("rapporteur")
Doctorat en informatique, Aix-Marseille Université, France
Approches jointes texte/image pour la compréhension multimodale de documents
Directeurs de thèse : Frédéric Béchet, Alexis Nasr (Aix-Marseille Université)

2018 **Kim Gerdes**
Président du jury
HDR en linguistique, Université Paris X Nanterre, France
Same same but different : Paradigms in syntax
Parrain d'HDR : Sylvain Kahane (Université Paris X Nanterre)

2015 **Wajdi Zaghouani**
Examinateur
Thèse d'informatique de l'Université Paris X Nanterre
Directeur de thèse : Sylvain Kahane (Université Paris X Nanterre)

2014 **Édouard Grave**
Examinateur
Thèse d'informatique de l'Université Pierre et Marie Curie
Directeurs de thèse : Francis Bach (Inria) et Guillaume Obozinski (École des Ponts)

2011 **Rania Voskaki**
Examinateur
Thèse d'informatique linguistique de l'Université Paris-Est Marne-la-Vallée
Directrice de thèse : Tita Kyriakopoulou

2010 **Claire Mouton**
Examinateur
Thèse d'informatique de l'Université Paris Sud
Directrice de thèse : Anne Vilnat (Université Paris Sud).
Co-directeur : Gaël de Chalendar (CEA)

2010 **Lionel Nicolas**
Examinateur
Thèse d'informatique de l'Université de Nice
Directeur de thèse : Jacques Farré

2009 **Juan Otero Pombo**
Rapporteur
Thèse d'informatique linguistique de l'Universidade de Vigo, Espagne
Directeurs de thèse : Manuel Vilares Ferro (Universidade de Vigo) et Jorge Graña Gil (Universidade de A Coruña)

2008 **Laurence Delort**
Examinateur
Thèse de linguistique de l'Université Paris-Diderot
Directrice de thèse : Laurence Danlos

Activités pour la communauté

2018	**Rédacteur en chef invité**	*Co-éditeur, avec Olivier Bonami, du numéro de la revue* Morphology *sur les approches computationnelles de la morphologie*
2012	**Rédacteur en chef invité**	*Co-éditeur, avec Núria Bel, du numéro de la revue* Traitement Automatique des Langues *sur les ressources linguistiques*
2016–	**Responsabilité chez Inria**	*Membre du Bureau du Comité des Projets d'Inria Paris*
2011–2019	**Responsabilité chez Inria**	*Membre du groupe de travail Relations Internationales du Comité d'Orientation Scientifique et Technologique d'Inria*
2015–	**Conseil Scientifique**	*Membre du Conseil Scientifique de l'EquipEx Ortolang (Infrastructure nationale pour les ressources linguistiques)*
2012–	**Conseil Scientifique**	*Membre du Comité de Pilotage des consortiums Corpus Écrits puis CORLI, au sein de la TGIR Huma-Num*
2011–2019	**Conseil Scientifique**	*Membre du Conseil Scientifique Restreint (CSR) du LabEx EFL et responsable de l'axe 6 « Ressources Linguistiques » (2011–2015), actuellement responsable adjoint de cet axe et membre suppléant du CSR*
2016	**Organisation scientifique**	*Workshop « Computational methods for descriptive and theoretical morphology » de l'édition 2016 l'International Morphology Meeting (co-organisation avec Olivier Bonami)*
2011	**Organisation scientifique**	*WoLeR, International Workshop on Lexical Resources, atelier de l'école d'été ESSLLI 2011*
2016	**Expertise**	*Expert pour l'HCERES, membre du comité d'évaluation de l'UMR ATILF*
2011–	**Expertise**	*Expert pour l'ANR, chargé de l'évaluation de propositions de projets dans différents domaines (non-thématique, STIC, humanités et sciences sociales)*
2005–	**Comités scientifiques**	*Membre de comités de lecture pour de nombreuses conférences et journaux, dont ACL, EMNLP, NAACL, EACL, EMNLP, CoLing, IJCNLP,* Computational Linguistics, Language Resources and Evaluation, Natural Language Engineering, Journal of Language Modelling *et* Traitement Automatique des Langues
2017–	**Société savante**	*Membre du Bureau, ancien Trésorier Adjoint et actuel Administrateur de la Société de Linguistique de Paris*
2005–2016	**Société savante**	*Membre de l'ATALA (Association pour le traitement automatique des langues). Ancien membre élu du Conseil d'Administration (2007-2016) et ancien Secrétaire (2011-2013).*
2014–2016	**Société savante**	*Membre du Comité Permanent des conférences TALN et RECITAL en tant que représentant du Conseil d'Administration de l'ATALA*

Langues

Maternelle	**français**
Courant	**anglais**
Lu, écrit, parlé	**allemand**
Conversationnel	**slovaque**

Publications

La liste de mes publications est acccessible sur la plateforme HAL. Elle comporte entre autres :

- 2 chapitres de livres,
- 16 articles dans des revues internationales, dont trois dans *Language Resources and Evaluation*, deux dans *Linguisticæ Investigationes*, quatre dans *Traitement Automatique des Langues*, un dans le *Journal of Language Modelling*, un dans *Indogermanische Forschungen* et un, à paraître, dans les *Münchener Studien zur Sprachwissenschaft*,
- la co-édition de numéros thématiques pour les revues *Traitement Automatique des Langues* (2011) et *Morphology* (2018),
- 133 articles dans des actes de conférences internationales avec comité de lecture, dont trois articles présentés à ACL et deux articles présentés à CoLing.

Articles dans des revues avec comité de lecture

[1] Robin Algayres, Tristan Ricoul, Julien Karadayi, Hugo Laurençon, Salah Zaiem, Abdelrahman Mohamed, Benoît Sagot, and Emmanuel Dupoux. DP-Parse : Finding Word Boundaries from Raw Speech with an Instance Lexicon. *Transactions of the Association for Computational Linguistics*, 10 :1051–1065, September 2022.

[2] Marianna Apidianaki and Benoît Sagot. Data-driven Synset Induction and Disambiguation for Wordnet Development. *Language Resources and Evaluation*, 48(4) :655–677, November 2014.

[3] Núria Bel and Benoît Sagot. Free Language Resources. *Revue TAL*, 52(3), 2011. Benoît Sagot and Núria Bel (eds.).

[4] Sacha Beniamine, Olivier Bonami, and Benoît Sagot. Inferring inflection classes with description length. *Journal of Language Modelling*, 5(3) :465–525, February 2018.

[5] Olivier Bonami and Benoît Sagot. Computational methods for descriptive and theoretical morphology : a brief introduction. *Morphology*, 27(4) :1–7, 2017.

[6] Pierre Boullier and Benoît Sagot. Analyse syntaxique profonde à grande échelle : SxLFG. *Revue TAL*, 46(2) :65–89, 2005.

[7] Thierry Chanier, Céline Poudat, Benoît Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi, and Djamé Seddah. The CoMeRe corpus for French : structuring and annotating heterogeneous CMC genres. *Journal for language technology and computational linguistics*, 29(2) :1–30, 2014. Final version to Special Issue of JLCL (Journal of Language Technology and Computational Linguistics (JLCL, http ://jlcl.org/) : BUILDING AND ANNOTATING CORPORA OF COMPUTER-MEDIATED DISCOURSE : Issues and Challenges at the Interface of Corpus and Computational Linguistics (ed. by Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel).

[8] Laurence Danlos and Benoît Sagot. Constructions pronominales dans Dicovalence et le lexique-grammaire – Intégration dans le Lefff. *Lingvisticae Investigationes*, 32(2) :293–304, 2009.

[9] Pascal Denis and Benoît Sagot. Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4) :721–736, 2012.

[10] Darja Fišer and Benoît Sagot. Constructing a poor man's wordnet in a resource-rich world. *Language Resources and Evaluation*, 49(3) :601–635, 2015.

[11] Romain Garnier and Benoît Sagot. A shared substrate between Greek and Italic. *Indogermanische Forschungen*, 122(1) :29–60, September 2017.

[12] Romain Garnier and Benoît Sagot. Metathesis of Proto-Indo-European Sonorants. *Münchener Studien zur Sprachwissenschaft*, 73(1) :29–53, 2019.

[13] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile

Dlamini, Nisansa de Silva, Sakine Çabuk Balli, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a Glance : An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10 :50–72, January 2022.

[14] Tu Anh Nguyen, Benoît Sagot, and Emmanuel Dupoux. Are discrete units necessary for Spoken Language Modeling ? *IEEE Journal of Selected Topics in Signal Processing*, August 2022.

[15] Benoît Sagot. Ressources lexicales libres pour le français. *Culture et recherche*, 124 :53, 2011.

[16] Benoît Sagot. Représentation de l'information sémantique lexicale : le modèle wordnet et son application au français. *Revue Française de Linguistique Appliquée*, XXII, 2017.

[17] Benoît Sagot and Pierre Boullier. From Raw Corpus to Word Lattices : Robust Pre-parsing Processing with SxPipe. *Archives of Control Sciences*, 15(4) :653–662, 2005.

[18] Benoît Sagot and Pierre Boullier. SxPipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts. *Revue TAL*, 49(2) :155–188, 2008.

[19] Benoît Sagot and Laurence Danlos. Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire - Constructions impersonnelles et expressions verbales figées. *Cahiers du Cental*, 5 :107–126, 2008.

[20] Benoît Sagot, Karën Fort, and Fabienne Venant. Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes. *Lingvisticae Investigationes*, 32(2) :305–315, January 2009.

[21] Benoît Sagot and Elsa Tolone. Exploitation des tables du Lexique-Grammaire pour l'analyse syntaxique automatique. *Arena Romanistica - Journal of Romance studies*, 4 :302–312, September 2009. The 28th Conference on Lexis and Grammar.

[22] Benoît Sagot and Éric Villemonte de La Clergerie. Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. *Revue TAL*, 49(1) :41–60, 2008.

[23] Géraldine Walther and Benoît Sagot. Modélisation et implémentation de phénomènes flexionnels non-canoniques. *Revue TAL*, 52(2), 2011.

[24] Géraldine Walther and Benoît Sagot. Modélisation et implémentation de phénomènes non-canoniques. *Revue TAL*, 52(2/2011) :91–122, December 2011. Vers la morphologie et au-del&agrave ;.

Articles dans des actes de colloques avec comité de lecture

[25] Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. In *Thirteenth Language Resources and Evaluation Conference - LREC 2022*, Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, June 2022. 12 pages, 6 figures, 2 tables.

[26] Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. Ungoliant : An Optimized Pipeline for the Generation of a Very Large-Scale Multilingual Web Corpus. In *CMLC 2021 - 9th Workshop on Challenges in the Management of Large Corpora*, Limerick / Virtual, Ireland, July 2021.

[27] Gilles Adda, Benoît Sagot, Karen Fort, and Joseph Mariani. Crowdsourcing for Language Resource Development : Critical Analysis of Amazon Mechanical Turk Overpowering Use. In *5th Language and Technology Conference*, Poznan, Poland, November 2011.

[28] Jesujoba O Alabi, Lydia Nishimwe, Benjamin Muller, Camille Rey, Benoît Sagot, and Rachel Bawden. Inria-ALMAnaCH at the WMT 2022 shared task : Does Transcription Help Cross-Script Machine Translation ? In *EMNLP 2022 - Seventh Conference on Machine Translation (WMT22 - Workshop on Statistical Machine Translation)*, Abu Dhabi, United Arab Emirates, December 2022.

[29] Robin Algayres, Adel Nabli, Benoît Sagot, and Emmanuel Dupoux. Speech Sequence Embeddings using Nearest Neighbors Contrastive Learning. In *Interspeech 2022 - 23rd INTERSPEECH Conference*, Incheon, South Korea, September 2022.

[30] Robin Algayres, Mohamed Salah Zaiem, Benoît Sagot, and Emmanuel Dupoux. Evaluating the reliability of acoustic speech embeddings. In *INTERSPEECH 2020 - Annual Conference of the International Speech Communication Association*, Shanghai / Vitrtual, China, October 2020.

[31] Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. ASSET : A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations. In *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States, July 2020.

[32] Marianna Ma Apidianaki and Benoît Sagot. Applying cross-lingual WSD to wordnet development. In *LREC 2012 - Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, May 2012.

[33] Srinivas Bangalore, Pierre Boullier, Alexis Nasr, Owen Rambow, and Benoît Sagot. MICA : A Probabilistic Dependency Parser Based on Tree Insertion Grammars. In *NAACL 2009 - North American Chapter of the Association for Computational Linguistics (Short Papers)*, Boulder, Colorado, United States, 2009.

[34] Marion Baranes and Benoît Sagot. A Language-Independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014.

[35] Marion Baranes and Benoît Sagot. Normalisation de textes par analogie : le cas des mots inconnus. In *TALN - Traitement Automatique du Langage Naturel*, pages 137–148, Marseille, France, July 2014.

[36] Alexandre Bartz, Juliette Janes, Laurent Romary, Philippe Gambette, Rachel Bawden, Pedro Ortiz Suarez, Benoît Sagot, and Simon Gabay. Expanding the content model of annotationBlock. In *Next Gen TEI, 2021 - TEI Conference and Members' Meeting*, Virtual, United States, October 2021.

[37] Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. Automatic Normalisation of Early Modern French. In *LREC 2022 - 13th Language Resources and Evaluation Conference*, pages 3354–3366, Marseille, France, June 2022. European Language Resources Association.

[38] Frédéric Béchet, Benoît Sagot, and Rosa Stern. Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées. In *TALN'2011 - Traitement Automatique des Langues Naturelles*, Montpellier, France, June 2011.

[39] Sacha Beniamine, Olivier Bonami, and Benoît Sagot. Information-theoretic inflectional classification. In *1st International Quantitative Morphology Meeting*, Belgrade, Serbia, July 2015.

[40] Sacha Beniamine and Benoît Sagot. Segmentation strategies for inflection class inference. In *Décembrettes 9, Colloque international de morphologie*, Toulouse, France, December 2015. Université de Toulouse.

[41] Christophe Benzitoun, Karen Fort, and Benoît Sagot. TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *JEP-TALN 2012 - Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*, pages 99–112, Grenoble, France, June 2012.

[42] Helena Blancafort San José, Gaëlle Recourcé, Javier Couto, Benoît Sagot, Rosa Stern, and Denis Teyssou. Traitement des inconnus : une approche systématique de l'incomplétude lexicale. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Canada, July 2010.

[43] Pierre Boullier, Lionel Clément, Benoît Sagot, and Éric Villemonte de La Clergerie. Chaînes de traitement syntaxique. In *TALN 05*, pages 103–112, Dourdan, France, 2005.

[44] Pierre Boullier, Lionel Clément, Éric Villemonte de La Clergerie, and Benoît Sagot. Simple comme EASy. In *TALN 05*, pages 57–60, Dourdan, France, 2005.

[45] Pierre Boullier, Alexis Nasr, and Benoît Sagot. Constructing parse forests that include exactly the n-best PCFG trees. In *IWPT'09 - 11th International Conference on Parsing Technologies*, Paris, France, October 2009.

[46] Pierre Boullier and Benoît Sagot. Efficient LFG parsing : SxLfg. In *International Workshop on Parsing Technologies*, pages 1–10, Vancouver, Canada, 2005.

[47] Pierre Boullier and Benoît Sagot. Multi-Component Tree Insertion Grammars. In *FG 2009 - 14 th Conference on Formal Grammars*, Bordeaux, France, 2009.

[48] Pierre Boullier and Benoît Sagot. Parsing Directed Acyclic Graphs with Range Concatenation Grammars. In *International Conference on Parsing Technologies (IWPT 2009)*, Paris, France, 2009.

[49] Pierre Boullier, Benoît Sagot, and Lionel Clément. Un analyseur LFG efficace pour le français : SxLfg. In *Ttraitement Automatique des Langues Naturelles*, pages 403–408, Dourdan, France, 2005.

[50] Marie Candito, Pascal Amsili, Lucie Barque, Farah Benamara, Gaël de Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, Benoît Sagot, and Laure Vieu. Developing a French FrameNet : Methodology and First results. In *9th International Conference on Language Resources and Evaluation (LREC 2014)*, volume L14-1, pages 1372–1379, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[51] Thibault Charmet, Inès Cherichi, Matthieu Allain, Urszula Czerwinska, Amaury Fouret, Benoît Sagot, and Rachel Bawden. Complex Labelling and Similarity Prediction in Legal Texts : Automatic Analysis of France's Court of Cassation Rulings. In *LREC 2022 - 13th Language Resources and Evaluation Conference*, Marseille, France, June 2022.

[52] Jūratė Čingienė, Dimitri Tcherniak, and Benoît Sagot. Sentiment analysis of write-in comments related to organisational change. In *17th Congress of the European Association of Work and Organizational Psychology*, Oslo, Norway, May 2015. European Association of Work and Organizational Psychology.

[53] Lionel Clément, Bernard Lang, and Benoît Sagot. Morphology based automatic acquisition of large-coverage lexica. In *LREC 04*, pages 1841–1844, Lisbonne, Portugal, 2004.

[54] Laurence Danlos and Benoît Sagot. Constructions pronominales dans Dicovalence et le lexique-grammaire–intégration dans le Lefff. In *Proceedings of the 27th Lexicon-Grammar Conference*, L'Aquila, Italy, 2008.

[55] Laurence Danlos and Benoît Sagot. Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français. In *Proceedings of the workshop "Lexicographie et informatique : bilan et perspectives"*, Nancy, France, January 2008.

[56] Laurence Danlos and Benoît Sagot. Ponctuations fortes abusives. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Canada, July 2010.

[57] Laurence Danlos, Benoît Sagot, and Rosa Stern. Analyse discursive des incises de citation. In *2ème Congrès Mondial de Linguistique Française - CMLF 2010*, La Nouvelle Orléans, United States, July 2010. Institut de Linguistique Française.

[58] Pascal Denis and Benoît Sagot. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Pacific Asia Conference on Language, Information and Computation*, Hong Kong, China, 2009.

[59] Pascal Denis and Benoît Sagot. Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. In *Traitement automatique des langues naturelles*, Montréal, Canada, July 2010. Association pour le Traitement Automatique des Langues.

[60] Pascal Denis and Benoît Sagot. Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Canada, July 2010.

[61] Paul-Ambroise Duquenne, Hongyu Gong, Benoît Sagot, and Holger Schwenk. T-Modules : Translation Modules for Zero-Shot Cross-Modal Machine Translation. In *EMNLP 2022 - 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, December 2022.

[62] Emmanuel Eckard, Lucie Barque, Alexis Nasr, and Benoît Sagot. Dictionary-Ontology Cross-Enrichment Using TLFi and WOLF to enrich one another. In Michael Zock and Reinhard Rapp, editors, *CogALex-III - 3rd Workshop on Cognitive Aspects of the Lexicon*, Mumbai, India, December 2012. Curran Associates, Inc.

[63] Murielle Fabre, Pedro Javier Ortiz Suárez, Benoît Sagot, and Éric Villemonte de La Clergerie. French Contextualized Word-Embeddings with a sip of CaBeRnet : a New French Balanced Reference Corpus. In *CMLC-8 - 8th Workshop on the Challenges in the Management of Large Corpora*, Marseille, France, May 2020.

[64] Darja Fišer and Benoît Sagot. Combining multiple resources to build reliable wordnets. In *TSD 2008 - Text Speech and Dialogue*, Brno, Czech Republic, 2008.

[65] Darja Fišer and Benoît Sagot. Avtomatska razširitev in čiščenje sloWNeta. In *Devete konference Jezikovne Tehnologije / Ninth Language Technologies Conference*, Ljubljana, Slovenia, October 2014.

[66] Karen Fort and Benoît Sagot. Influence of Pre-annotation on POS-tagged Corpus Development. In *The Fourth ACL Linguistic Annotation Workshop*, pages 56–63, Uppsala, Sweden, July 2010.

[67] Clémentine Fourrier, Rachel Bawden, and Benoît Sagot. Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task ? In *ACL-IJCNLP 2021 - Findings of the Association for Computational Linguistics*, Bangkok, Thailand, August 2021.

[68] Clémentine Fourrier and Benoît Sagot. Comparing Statistical and Neural Models for Learning Sound Correspondences. In *LT4HALA 2020 - First Workshop on Language Technologies for Historical and Ancient Languages*, Marseille, France, May 2020. Due to the COVID-19 pandemic, the workshop will not take place. However, the proceedings are published online.

[69] Clémentine Fourrier and Benoît Sagot. Methodological Aspects of Developing and Managing an Etymological Lexical Resource : Introducing EtymDB 2.0. In *LREC 2020 - 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020. Due to COVID19 pandemic, the 12th edition is cancelled. The LREC 2020 Proceedings are available at http ://www.lrec-conf.org/proceedings/lrec2020/index.html.

[70] Clémentine Fourrier and Benoît Sagot. Probing Multilingual Cognate Prediction Models. In *ACL 2022 - Findings of the Association for Computational Linguistics*, Dublin, Ireland, May 2022.

[71] Simon Gabay, Rachel Bawden, Philippe Gambette, Jonathan Poinhos, Eleni Kogkitsidou, and Benoît Sagot. Le changement linguistique au XVIIe s. : nouvelles approches scriptométriques. In *CMLF 2022 - 8e Congrès Mondial de Linguistique Française*, volume 138 of *SHS Web of conferences*, pages 02006.1–14, Orléans, France, July 2022. EDP Sciences.

[72] Simon Gabay, Rachel Bawden, Benoît Sagot, and Philippe Gambette. Vers l'étude linguistique sur données artificielles. In *Variation(s) en français*, Nancy, France, November 2022. ATILF.

[73] Simon Gabay, Philippe Gambette, Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, and Benoît Sagot. Variation graphique dans les documents d'Ancien Régime : Nouvelles approches scriptométriques. In *Journée d'étude : " Pour une histoire de la langue 'par en bas' : textes privés et variation des langues dans le passé "*, Paris, France, September 2021.

[74] Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. From FreEM to D'AlemBERT. In *13th Language Resources and Evaluation Conference - LREC 2022*, Proceedings of the 13th Language Resources and Evaluation Conference, pages 3367–3374, Marseille, France, June 2022. European Language Resources Association. 8 pages, 2 figures, 4 tables.

[75] Simon Gabay, Pedro Ortiz Suarez, Rachel Bawden, Alexandre Bartz, Philippe Gambette, and Benoît Sagot. Le projet FREEM : ressources, outils et enjeux pour l'étude du français d'Ancien Régime. In Yannick Estève, Tania Jiménez, Titouan Parcollet, and Marcely Zanon Boito, editors, *TALN 2022 - Traitement Automatique des Langues Naturelles*, pages 154–165, Avignon, France, June 2022. ATALA.

[76] Kata Gábor, Marianna Apidianaki, B Sagot, and Éric Villemonte de La Clergerie. Boosting the coverage of a semantic lexicon by automatically extracted event nominalizations. In *International Conference on Language Resources and Evaluation*, Istanbul, Turkey, January 2012.

[77] Kata Gábor, Marianna Ma Apidianaki, Benoît Sagot, and Éric Villemonte de La Clergerie. Boosting the Coverage of a Semantic Lexicon by Automatically Extracted Event Nominalizations. In *LREC 2012 - Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, May 2012.

[78] Kata Gábor and Benoît Sagot. Automated Error Detection in Digitized Cultural Heritage Documents. In *EACL 2014 Workshop on Language Technology for Cultural Heritage*, Göteborg, Sweden, April 2014.

[79] Romain Garnier and Benoît Sagot. Could Greek and Italic share a same Indo-European substratum ? In *22nd International Conference on Historical Linguistics*, Naples, Italy, July 2015.

[80] Romain Garnier and Benoît Sagot. New results on a centum substratum in Greek : the Lydian connection. In *International Colloquium on Loanwords and Substrata in Indo-European languages*, Limoges, France, June 2018.

[81] Antoine Gérard, Benoît Sagot, and Emilie Pons. Le Traitement Automatique des Langues au service du vin. In *Dataquitaine 2021 - IA, Recherche Opérationnelle & Data Science*, Bordeaux / Virtual, France, February 2021.

[82] Nathan Godey, Roman Castagné, Eric Villemonte de La Clergerie, and Benoît Sagot. MANTa : Efficient Gradient-Based Tokenization for Robust End-to-End Language Modeling. In *EMNLP 2022 - The 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, December 2022.

[83] Omer Goldman, Francesco Tinner, Hila Gonen, Benjamin Muller, Victoria Basmov, Shadrack Kirimi, Lydia Nishimwe, Benoît Sagot, Djamé Seddah, Reut Tsarfaty, and Duygu Ataman. The MRL 2022 Shared Task on Multilingual Clause-level Morphology. In *1st Shared Task on Multilingual Clause-level Morphology*, Abu Dhabi, United Arab Emirates, December 2022.

[84] Loïc Grobol, Mathilde Regnault, Pedro Ortiz Suarez, Benoît Sagot, Laurent Romary, and Benoît Crabbé. BERTrade : Using Contextual Embeddings to Parse Old French. In *13th Language Resources and Evaluation Conference*, Marseille, France, June 2022. European Language Resources Association.

[85] Gaël Guibon and Benoît Sagot. OFrLex : A Computational Morphological and Syntactic Lexicon for Old French. In *LREC 2020 - 12th Language Resources and Evaluation Conference*, pages 3217–3225 (updated version), Marseille, France, May 2020. Due to COVID19 pandemic, the 12th edition is cancelled. The LREC 2020 Proceedings are available at http ://www.lrec-conf.org/proceedings/lrec2020/index.htmlThe version 2 of the paper is an updated version with regard to the originally published version (minor corrections).

[86] Valérie Hanoka and Benoît Sagot. Wordnet creation and extension made simple : A multilingual lexicon-based approach using wiki resources. In *LREC 2012 : 8th international conference on Language Resources and Evaluation*, page 6, Istanbul, Turkey, May 2012.

[87] Valérie Hanoka and Benoît Sagot. YaMTG : An Open-Source Heavily Multilingual Translation Graph Extracted from Wiktionaries and Parallel Corpora. In *Language Resources and Evaluation Conference*, Reykjavik, Iceland, May 2014. European Language Resources Association.

[88] Ganesh Jawahar, Benjamin Muller, Amal Fethi, Louis Martin, Éric Villemonte de La Clergerie, Benoît Sagot, and Djamé Seddah. ELMoLex : Connecting ELMo and Lexicon features for Dependency Parsing. In *CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium, October 2018.

[89] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language ? In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019.

[90] Joseph Le Roux, Benoît Sagot, and Djamé Seddah. Statistical Parsing of Spanish and Data Driven Lemmatization. In *ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012)*, page 6 pages, Jeju, South Korea, July 2012.

[91] Pierre Magistry and Benoît Sagot. Segmentation et induction de lexique non-supervisées du mandarin. In *TALN'2011 - Traitement Automatique des Langues Naturelles*, Montpellier, France, June 2011. ATALA.

[92] Pierre Magistry and Benoît Sagot. Unsupervized Word Segmentation : the case for Mandarin Chinese. In *ACL - Annual Meeting of the Association for Computational Linguistics - 2012*, Jeju, South Korea, July 2012. ACL.

[93] Pierre Magistry and Benoît Sagot. Can MDL Improve Unsupervised Chinese Word Segmentation ? In *Sixth International Joint Conference on Natural Language Processing : Sighan workshop*, page 2, Nagoya, Japan, October 2013.

[94] Louis Martin, Angela Fan, Éric Villemonte de la Clergerie, Antoine Bordes, and Benoît Sagot. MUSS : Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In *LREC 2022 - 13th Language Resources and Evaluation Conference*, Marseille, France, June 2022.

[95] Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Antoine Bordes, Éric Villemonte de La Clergerie, and Benoît Sagot. Reference-less Quality Estimation of Text Simplification Systems. In *1st Workshop on Automatic Text Adaptation (ATA)*, Tilburg, Netherlands, November 2018.

[96] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoan Dupont, Laurent Romary, Eric Villemonte de La Clergerie, Benoît Sagot, and Djamé Seddah. Les modèles de langue contextuels Camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entrainement. In Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, and Stéphane Schneider, editors, *JEP-TALN-RECITAL 2020 - 33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 54–65, Nancy / Virtuel, France, June 2020. ATALA.

[97] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT : a Tasty French Language Model. In *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States, July 2020.

[98] Louis Martin, Éric Villemonte de La Clergerie, Benoît Sagot, and Antoine Bordes. Controllable Sentence Simplification. In *LREC 2020 - 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020. Due to COVID19 pandemic, the 12th edition is cancelled. The LREC 2020 Proceedings are available at http ://www.lrec-conf.org/proceedings/lrec2020/index.html.

[99] Héctor Alonso Martínez, Djamé Seddah, and Benoît Sagot. From Noisy Questions to Minecraft Texts : Annotation Challenges in Extreme Syntax Scenarios. In *2nd Workshop on Noisy User-generated Text (W-NUT) at CoLing 2016*, Osaka, Japan, December 2016.

[100] Héctor Martínez Alonso, Amaury Delamaire, and Benoît Sagot. Annotating omission in statement pairs. In *11th Linguistic Annotation Workshop*, pages 41–45, Valencia, Spain, April 2017.

[101] Seyed Abolghasem Mirroshandel, Alexis Nasr, and Benoît Sagot. Enforcing Subcategorization Constraints in a Parser Using Sub-parses Recombining. In *NAACL 2013 - Conference of the North American Chapter of the Association for Computational Linguistics*, Atlanta, United States, June 2013.

[102] Miguel A. Molinero, Benoît Sagot, and Lionel Nicolas. A morphological and syntactic wide-coverage lexicon for Spanish : The Leffe. In *RANLP 2009 - Recent Advances in Natural Language Processing*, Borovets, Bulgaria, September 2009.

[103] Miguel A. Molinero, Benoît Sagot, and Lionel Nicolas. Building a morphological and syntactic lexicon by merging various linguistic resources. In *NODALIDA 2009 - the 17th Nordic Conference of Computational Linguistics*, Odense, Denmark, May 2009.

[104] Amir More, Özlem Çetinoğlu, Çağri Çöltekin, Nizar Habash, Benoît Sagot, Djamé Seddah, Dima Taji, and Reut Tsarfaty. CoNLL-UL : Universal Morphological Lattices for Universal Dependency Parsing. In *11th Language Resources and Evaluation Conference*, Miyazaki, Japan, May 2018.

[105] Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. When Being Unseen from mBERT is just the Beginning : Handling New Languages With Multilingual Language Models. In *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Mexico City, Mexico, June 2021.

[106] Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. Quand être absent de mBERT n'est que le commencement : Gérer de nouvelles langues à l'aide de modèles de langues multilingues. In Yannick Estève, Tania Jiménez, Titouan Parcollet, and Marcely Zanon Boito, editors, *TALN 2022 - 29° conférence sur le Traitement Automatique des Langues Naturelles*, pages 450–451, Avignon, France, June 2022. ATALA. Code available at https ://github.com/benjamin-mlr/mbert-unseen-languages.git Apprentissage par transfert.

[107] Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. First Align, then Predict : Understanding the Cross-Lingual Ability of Multilingual BERT. In *EACL 2021 - The 16th Conference of the European Chapter of the Association for Computational Linguistics*, Kyiv / Virtual, Ukraine, April 2021.

[108] Benjamin Muller, Benoît Sagot, and Djamé Seddah. Enhancing BERT for Lexical Normalization. In *The 5th Workshop on Noisy User-generated Text (W-NUT)*, Hong Kong, China, November 2019.

[109] Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. Generative Spoken Dialogue Language Modeling. In *SLT-2022 - IEEE Spoken Language Technology Workshop*, Doha-Qatar, Qatar, January 2023.

[110] Lionel Nicolas, Miguel A. Molinero, Benoît Sagot, Nieves Fernández Formoso, and Vanesa Vidal Castro. Creating and maintaining language resources : the main guidelines of the Victoria project. In *Workshop on Language Resources : From Storyboard to Sustainability and LR Lifecycle Management (LREC 2010 workshop)*, Valletta, Malta, May 2010.

[111] Lionel Nicolas, Miguel A. Molinero, Benoît Sagot, Elena Sánchez Trigo, Éric Villemonte de La Clergerie, Miguel Alonso Pardo, Jacques Farré, and Joan Miquel-Vergès. Construcciòn y extensiòn de un léxico morfológico y sintáctico para el Español : el Leffe. In *Proceedings of SEPLN 09*, San Sebastian, Spain, Spain, 2009.

[112] Lionel Nicolas, Miguel A. Molinero, Benoît Sagot, Elena Sánchez Trigo, Éric Villemonte de La Clergerie, Miguel Alonso Pardo, Jacques Farré, and Joan Miquel-Vergès. Towards efficient production of linguistic resources : the Victoria Project. In *Proceedings of the International Conference RANLP-2009*, pages 318–323, Borovets, Bulgaria, Bulgaria, 2009. Association for Computational Linguistics.

[113] Lionel Nicolas, Miguel A. Molinero, Benoît Sagot, Elena Sánchez Trigo, Éric Villemonte de La Clergerie, M.A. Pardo, Jacques Farré, and J. Miquel Vergés. Producción eficiente de recursos lingüísticos : el proyecto Victoria. In *SEPLN 09 - 25th edition of the Annual Conference of the Spanish Society for Natural Language Processing*, Donostia, Spain, September 2009.

[114] Lionel Nicolas, Benoît Sagot, Miguel A. Molinero, Jacques Farré, and Éric Villemonte de La Clergerie. Mining Parsing Results for Lexical Correction : Toward a Complete Correction Process of Wide-Coverage Lexicons. In Zygmunt Vetulani and Hans Uszkoreit, editors, *LTC 2007 - Third Language and Technology Conference*, volume 5603 of *Lecture Notes in Computer Science*, pages 178–191, Poznan, Poland, October 2007. Springer.

[115] Lionel Nicolas, Benoît Sagot, Miguel A. Molinero, Jacques Farré, and Éric Villemonte de La Clergerie. Computer aided correction and extension of a syntactic wide-coverage lexicon. In *Coling 2008*, CD ROM, pages pp 604–611, Manchester, United Kingdom, August 2008.

[116] Lionel Nicolas, Benoît Sagot, Miguel A. Molinero, Jacques Farré, and Éric Villemonte de La Clergerie. Extensión y corrección semi-automática de léxicos morfo-sintácticos. In *24th edition of the conference of the Spanish Society for Natural Language Processing (SEPLN 2008)*, Madrid, Spain, September 2008. El Advanced Database research group, LaBDA.

[117] Lionel Nicolas, Benoît Sagot, Miguel A. Molinero, Jacques Farré, and Éric Villemonte de La Clergerie. Trouver et confondre les coupables : un processus sophistiqué de correction de lexique. In *16ème conférence sur le Traitement Automatique des Langues Naturelles : TALN'09*, Senlis, France, June 2009. ATALA ; LIPN.

[118] Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary, and Benoît Sagot. Establishing a New State-of-the-Art for French Named Entity Recognition. In *LREC 2020 - 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020. Due to COVID19 pandemic, the 12th edition is cancelled. The LREC 2020 Proceedings are available at http ://www.lrec-conf.org/proceedings/lrec2020/index.html.

[119] Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. Preparing the Dictionnaire Universel for Automatic Enrichment. In *10th International Conference on Historical Lexicography and Lexicology (ICHLL)*, Leeuwarden, Netherlands, June 2019.

[120] Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States, July 2020.

[121] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen, and Caroline Iliadi, editors, *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom, July 2019. Leibniz-Institut für Deutsche Sprache.

[122] Arij Riabi, Benoît Sagot, and Djamé Seddah. Can Character-based Language Models Improve Downstream Task Performance in Low-Resource and Noisy Language Scenarios ? In *Seventh Workshop on Noisy User-generated Text (W-NUT 2021, colocated with EMNLP 2021)*, Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Punta Cana, Dominican Republic, January 2022.

[123] Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta cana, Dominican Republic, November 2021.

[124] Benoît Sagot. Building a Morphosyntactic Lexicon and a Pre-syntactic Processing Chain for Polish. In Zygmunt Vetulani and Hans Huszkoreit, editors, *Language and Technology Conference*, volume 5603 of *Lecture Notes in Computer Science*, Poznań, Poland, 2007. Springer.

[125] Benoît Sagot. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, May 2010.

[126] Benoît Sagot. Comparing Complexity Measures. In *Computational approaches to morphological complexity*, Paris, France, February 2013. Surrey Morphology Group.

[127] Benoît Sagot. DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German. In *Language Resources and Evaluation Conference*, Reykjavik, Iceland, May 2014. European Language Resources Association.

[128] Benoît Sagot. Les catégories prédicatives dans le Lefff. In *Journée d'étude " CAtégories Prédicatives et Traitement Automatique des Langues " (CAPTAL)*, Lille, France, February 2014.

[129] Benoît Sagot. Commentary on Jim Blevins, Implicational Morphology. In *AnaMorphoSys*, Lyon, France, June 2016.

[130] Benoît Sagot. Étiquetage multilingue en parties du discours avec MElt. In *23ème Conférence sur le Traitement Automatique des Langues Naturelles*, Paris, France, July 2016.

[131] Benoît Sagot. Construction automatique d'une base de données étymologiques à partir du wiktionary. In *Traitement Automatique des Langues Naturelles 2017*, Orléans, France, June 2017.

[132] Benoît Sagot. Extracting an Etymological Database from Wiktionary. In *Electronic Lexicography in the 21st century (eLex 2017)*, pages 716–728, Leiden, Netherlands, September 2017.

[133] Benoît Sagot. A multilingual collection of CoNLL-U-compatible morphological lexicons. In *Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018.

[134] Benoît Sagot. A new PIE root *h1er '(to be) dark red, dusk red' : drawing the line between inherited and borrowed words for 'red(ish)', 'pea', 'ore', 'dusk' and 'love' in daughter languages. In *International Colloquium on Loanwords and Substrata in Indo-European languages*, Limoges, France, June 2018.

[135] Benoît Sagot. Développement d'un lexique morphologique et syntaxique de l'ancien français. In *26ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Toulouse, France, July 2019.

[136] Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie, and Pierre Boullier. Vers un méta-lexique pour le français : architecture, acquisition, utilisation. In *Journée ATALA sur l'interface lexique-grammaire*, Paris, France, 2005.

[137] Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie, and Pierre Boullier. The Lefff 2 syntactic lexicon for French : architecture, acquisition, use. In *LREC 06*, pages 1–4, Gênes, Italy, 2006.

[138] Benoît Sagot and Laurence Danlos. Verbes de citation et Tables du Lexique-Grammaire. In *International Conference on Lexis and Grammar*, Belgrade, Serbia, September 2010.

[139] Benoît Sagot and Laurence Danlos. Merging syntactic lexica : the case for French verbs. In *LREC'12 Workshop on Merging Language Resources*, Istanbul, Turkey, May 2012.

[140] Benoît Sagot, Laurence Danlos, and Margot Colinet. Sous-catégorisation en pour et syntaxe lexicale. In *Traitement Automatique du Langage Naturel 2014*, Marseille, France, July 2014.

[141] Benoît Sagot, Laurence Danlos, and Susanne Salmon-Alt. French frozen verbal expressions : from lexicon-grammar tables to NLP applications. In *Colloque Lexique et Grammaire 2006*, Palerme, Italy, 2006.

[142] Benoît Sagot, Laurence Danlos, and Rosa Stern. A Lexicon of French Quotation Verbs for Automatic Quotation Extraction. In *7th international conference on Language Resources and Evaluation - LREC 2010*, Valetta, Malta, May 2010.

[143] Benoît Sagot and Darja Fišer. Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Morocco, May 2008.

[144] Benoît Sagot and Darja Fišer. Construction d'un wordnet libre du français à partir de ressources multilingues. In *TALN 2008 -Traitement Automatique des Langues Naturelles*, Avignon, France, June 2008.

[145] Benoît Sagot and Darja Fišer. Extending wordnets by learning from multiple resources. In *LTC'11 : 5th Language and Technology Conference*, Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland, November 2011.

[146] Benoît Sagot and Darja Fišer. Automatic Extension of WOLF. In *GWC2012 - 6th International Global Wordnet Conference*, Matsue, Japan, January 2012. Global Wordnet Association + Toyohashi University of Technology + National Institute of Japanese Language and Linguistics.

[147] Benoît Sagot and Darja Fišer. Cleaning noisy wordnets. In *LREC 2012 - Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, May 2012.

[148] Benoît Sagot and Karen Fort. Améliorer un lexique syntaxique à l'aide des tables du Lexique-Grammaire : Adverbes en -ment. In *26e Colloque International sur le Lexique et la grammaire 2007*, Bonifacio, France, October 2007.

[149] Benoît Sagot and Karen Fort. Description et analyse des verbes désadjectivaux et dénominaux en -ifier et -iser. In *28ème Colloque international sur le lexique et la grammaire (LGC'09)*, volume 4, pages 102–109, Bergen, Norway, September 2009.

[150] Benoît Sagot, Karen Fort, Gilles Adda, Joseph Mariani, and Bernard Lang. Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. In *TALN'2011 - Traitement Automatique des Langues Naturelles*, Montpellier, France, June 2011.

[151] Benoît Sagot, Karen Fort, and Fabienne Venant. Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes. In *27e Colloque international sur le lexique et la grammaire 2008*, page 0, L'Aquila, Italy, September 2008.

[152] Benoît Sagot, Karen Fort, and Fabienne Venant. Extending the Adverbial Coverage of a French WordNet. In *NODALIDA 2009 workshop on WordNets and other Lexical Semantic Resources*, page 0, Odense, Denmark, May 2009.

[153] Benoît Sagot and Kata Gábor. Détection et correction automatique d'entités nommées dans des corpus OCRisés. In *Traitement Automatique du Langage Naturel 2014*, Marseille, France, July 2014.

[154] Benoît Sagot and Héctor Martínez Alonso. Improving neural tagging with lexical information. In *15th International Conference on Parsing Technologies*, pages 25–31, Pisa, Italy, September 2017.

[155] Benoît Sagot, Damien Nouvel, Virginie Mouilleron, and Marion Baranes. Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel. In *TALN - Traitement Automatique du Langage Naturel*, pages 407–420, Les sables d'Olonne, France, June 2013.

[156] Benoît Sagot, Marion Richard, and Rosa Stern. Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. In Georges Antoniadis, Hervé Blanchon, and Gilles Sérasset, editors, *Traitement Automatique des Langues Naturelles (TALN)*, volume 2 - TALN of *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, Grenoble, France, June 2012.

[157] Benoît Sagot, Laurent Romary, Rachel Bawden, Pedro Javier Ortiz Suárez, Kelly Christensen, Simon Gabay, Ariane Pinche, and Jean-Baptiste Camps. Gallic(orpor)a : Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue. In *DataLab de la BnF : Restitution des travaux 2022*, Paris, France, December 2022. DataLab de la BnF.

[158] Benoît Sagot and Giorgio Satta. Optimal rank reduction for Linear Context-Free Rewriting Systems with Fan-Out Two. In *48th Annual Meeting of the Association for Computational Linguistics - ACL 2010*, Uppsala, Sweden, July 2010.

[159] Benoît Sagot and Rosa Stern. Aleda, a free large-scale entity database for French. In *LREC 2012 : eighth international conference on Language Resources and Evaluation*, page 4 pages, Istanbul, Turkey, May 2012.

[160] Benoît Sagot and Elsa Tolone. Intégrer les tables du Lexique-Grammaire à un analyseur syntaxique robuste à grande échelle. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, page electronic version (10 pp.), Senlis, France, June 2009.

[161] Benoît Sagot and Éric Villemonte de La Clergerie. Error mining in parsing results. In *The 21st International Conference of the Association for Computational Linguistics (ACL 2006)*, pages 329–336, Sydney, Australia, July 2006.

[162] Benoît Sagot and Géraldine Walther. A morphological lexicon for the Persian language. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, La Valette, Malta, May 2010.

[163] Benoît Sagot and Géraldine Walther. Développement de ressources pour le persan : lexique morphologique et chaîne de traitements de surface. In *TALN 2010*, Montréal, Canada, July 2010.

[164] Benoît Sagot and Géraldine Walther. Non-canonical inflection : data, formalisation and complexity measures. In *Systems and Frameworks for Computational Morphology*, volume 100 of *Communications in Computer and Information Science*, pages 23–45, France, 2011. Springer. Systems and Frameworks for Computational Morphology .

[165] Benoît Sagot and Géraldine Walther. Non-Canonical Inflection : Data, Formalisation and Complexity Measures. In Cerstin Mahlow and Michael Piotrowski, editors, *SFCM 2011 - The Second Workshop on Systems and Frameworks for Computational Morphology*, volume 100 of *Communications in Computer and Information Science*, pages 23–45, Zürich, Switzerland, August 2011. Springer.

[166] Benoît Sagot and Géraldine Walther. Implementing a formal model of inflectional morphology. In Cerstin Mahlow and Michael Piotrowski, editors, *Third International Workshop on Systems and Frameworks for Computational Morphology*, volume 380 of *Communications in Computer and Information Science*, pages 115–134, Berlin, Germany, September 2013. Humboldt-Universität, Springer.

[167] Benoît Sagot, Géraldine Walther, Pegah Faghiri, and Pollet Samvelian. A new morphological lexicon and a POS tagger for the Persian Language. In *International Conference in Iranian Linguistics*, Uppsala, Sweden, 2011.

[168] Benoît Sagot, Géraldine Walther, Pegah Faghiri, and Pollet Samvelian. Développement de ressources pour le persan : le nouveau lexique morphologique PerLex 2 et l'étiqueteur morphosyntaxique MElt-fa. In *TALN 2011 - Traitement Automatique des Langues Naturelles*, Montpellier, France, June 2011.

[169] Benoît Sagot, Géraldine Walther, Pegah Faghiri, and Pollet Samvelian. Développement de ressources pour le persan : le nouveau lexique morphologique PerLex 2 et l'étiqueteur morphosyntaxique $\mathrm{MElt}_{\mathrm{fa}}$. In *Actes de TALN 2011*, Montpellier, France, 2011.

[170] Benoît Sagot, Géraldine Walther, Pegah Faghiri, and Pollet Samvelian. Développement de ressources pour le persan : PerLex2, nouveau lexique morphologique et MElt_fa, étiqueteur morphosyntaxique. In *TALN 2011*, Montpellier, France, 2011.

[171] Pollet Samvelian, Laurence Danlos, and Benoît Sagot. On the predictability of light verbs. In *30th International Conference on Lexis and Grammar*, Nicosia, Cyprus, 2011.

[172] Yves Scherrer and Benoît Sagot. Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche. In *Atelier TALARE, TALN 2013*, Les Sables d'Olonne, France, June 2013. ATALA.

[173] Yves Scherrer and Benoît Sagot. Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources. In *RANLP Workshop on Adaptation of language resources and tools for closely related languages and language variants*, Hissar, Bulgaria, September 2013.

[174] Yves Scherrer and Benoît Sagot. A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages. In *Language Resources and Evaluation Conference*, Reykjavik, Iceland, May 2014. European Language Resources Association.

[175] Sebastian Schuster, Éric Villemonte de La Clergerie, Marie D Candito, Benoît Sagot, Christopher D Manning, and Djamé Seddah. Paris and Stanford at EPE 2017 : Downstream Evaluation of Graph-based Dependency Representations. In *EPE 2017 - The First Shared Task on Extrinsic Parser Evaluation*, Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation, pages 47–59, Pisa, Italy, September 2017.

[176] Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. Building a User-Generated Content North-African Arabizi Treebank : Tackling Hell. In *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States, July 2020.

[177] Djamé Seddah, Joseph Le Roux, and Benoît Sagot. Data Driven Lemmatization and Parsing of Italian. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *EVALITA 2011 - Evaluation of NLP and Speech Tools for Italian*, volume 7689 of *Lecture Notes in Computer Science*, pages 249–256, Rome, Italy, January 2012. Springer.

[178] Djamé Seddah, Joseph Le Roux, and Benoît Sagot. Data Driven Lemmatization for Statistical Constituent Parsing of Italian. In *Proceedings of EVALITA 2011*, Roma, Italy, Italy, January 2012. Springer.

[179] Djamé Seddah and Benoît Sagot. Modeling and Analysis of Elliptic Coordination by Dynamic Exploitation of Derivation Forests in LTAG Parsing. In *Proceedings of TAG+8 : The Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms*, Sydney, Australia, 2006.

[180] Djamé Seddah and Benoît Sagot. Modélisation et analyse des coordinations elliptiques par l'exploitation dynamique des forêts de dérivation. In *Proceedings of TALN 2006 : Traitement Automatique des Langues Naturelles*, pages 609–618, Leuven, Belgium, 2006.

[181] Djamé Seddah, Benoît Sagot, and Marie Candito. The Alpage Architecture at the SANCL 2012 Shared Task : Robust Pre-Processing and Lexical Bridging for User-Generated Content Parsing. In *SANCL 2012 - First Workshop on Syntactic Analysis of Non-Canonical Language , an NAACL-HLT'12 workshop*, Montréal, Canada, June 2012.

[182] Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. Building a treebank of noisy user-generated content : The French Social Media Bank. In *TLT 11 - The 11th International Workshop on Treebanks and Linguistic Theories*, Lisbonne, Portugal, November 2012. Cet article constitue une version réduite de l'article "The French Social Media Bank : a Treebank of Noisy User Generated Content" (mêmes auteurs).

[183] Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. The French Social Media Bank : a Treebank of Noisy User Generated Content. In *COLING 2012 - 24th International Conference on Computational Linguistics*, Mumbai, India, December 2012. Kay, Martin and Boitet, Christian.

[184] Djamé Seddah, Benoît Sagot, and Laurence Danlos. Control Verbs, Argument Cluster Coordination and MCTAG. In *10th International Conference on Tree Adjoining Grammars and Related Formalisms (TAG+10)*, page 0, New Haven, United States, 2010.

[185] Djamé Seddah, Éric Villemonte de La Clergerie, Benoît Sagot, Hector Martinez Alonso, and Marie Candito. Cheating a Parser to Death : Data-driven Cross-Treebank Annotation Transfer. In *Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018.

[186] Rosa Stern and Benoît Sagot. Détection et résolution d'entités nommées dans des dépêches d'agence. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Canada, July 2010.

[187] Rosa Stern and Benoît Sagot. Resources for Named Entity Recognition and Resolution in News Wires. In *Entity 2010 Workshop at LREC 2010*, Valletta, Malta, May 2010.

[188] Rosa Stern and Benoît Sagot. Population of a Knowledge Base for News Metadata from Unstructured Text and Web Data. In *AKBC-WEKEX 2012 - The Knowledge Extraction Workshop at NAACL-HLT 2012*, pages –, Montréal, Canada, June 2012.

[189] Rosa Stern, Benoît Sagot, and Frédéric Béchet. A Joint Named Entity Recognition and Entity Linking System. In *EACL 2012 Workshop on Innovative hybrid approaches to the processing of textual data*, pages –, Avignon, France, April 2012.

[190] Jana Strnadová and Benoît Sagot. Construction d'un lexique des adjectifs dénominaux. In *TALN'2011 - Traitement Automatique des Langues Naturelles, Montpellier*, pages 69–74, Montpellier, France, June 2011.

[191] Elsa Tolone and Benoît Sagot. Using Lexicon-Grammar Tables for French Verbs in a Large-Coverage Parser. In Zygmunt Vetulani, editor, *LTC 2009 - 4th Language and Technology Conference*, volume 6562 of *Lecture Notes in Artificial Intelligence*, pages 183–191, Poznań, Poland, November 2009. Springer.

[192] Elsa Tolone, Benoît Sagot, and Éric Villemonte de La Clergerie. Evaluating and improving syntactic lexica by plugging them within a parser. In *LREC 2012 - 8th International Conference on Language Resources and Evaluation*, page electronic version (8 pp.), Istambul, Turkey, May 2012.

[193] Elsa Tolone, Éric Villemonte de La Clergerie, and Benoît Sagot. Évaluation de lexiques syntaxiques par leur intégration dans l'analyseur syntaxique FRMG. In *LGC'11 - 30ème Colloque international sur le Lexique et la Grammaire*, pages pages 267–274, Nicosie, Cyprus, October 2011.

[194] Éric Villemonte de La Clergerie, Benoît Sagot, Lionel Nicolas, and Marie-Laure Guénot. FRMG : évolutions d'un analyseur syntaxique TAG du français. In Villemonte de la Clergerie, Éric, Paroubek, and Patrick, editors, *Journée de l'ATALA sur : Quels analyseurs syntaxiques pour le français ?*, Paris, France, October 2009. ATALA. Journée de l'ATALA organisée conjointement à la conférence IWPT 2009.

[195] Éric Villemonte de La Clergerie, Benoît Sagot, and Djamé Seddah. The ParisNLP entry at the ConLL UD Shared Task 2017 : A Tale of a #ParsingTragedy. In *Conference on Computational Natural Language Learning*, Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies, pages 243–252, Vancouver, Canada, August 2017.

[196] Éric Villemonte de La Clergerie, Benoît Sagot, Rosa Stern, Pascal Denis, Gaëlle Recourcé, and Victor Mignot. Extracting and Visualizing Quotations from News Wires. In Zygmunt Vetulani, editor, *LTC 2009 - 4th Language and Technology Conference*, volume 6562 of *Lecture Notes in Artificial Intelligence*, pages 522–532, Poznań, Poland, November 2009. Springer.

[197] Géraldine Walther, Guillaume Jacques, and Benoît Sagot. Uncovering the inner architecture of Khaling verbal morphology. In *3rd Workshop on Sino-Tibetan Languages of Sichuan*, Paris, France, September 2013.

[198] Géraldine Walther, Guillaume Jacques, and Benoît Sagot. The Opacity-Compactness Tradeoff : Morphomic Features for an Economical Account of Khaling Verbal Inflection. In *16th International Morphology Meeting (IMM 16)*, Budapest, Hungary, May 2014.

[199] Géraldine Walther and Benoît Sagot. Developing a Large-Scale Lexicon for a Less-Resourced Language : General Methodology and Preliminary Experiments on Sorani Kurdish. In *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)*, Valetta, Malta, 2010.

[200] Géraldine Walther and Benoît Sagot. Problèmes d'intégration morphologique d'emprunts d'origine anglaise en français. In *30th International Conference on Lexis and Grammar*, Nicosia, Cyprus, October 2011.

[201] Géraldine Walther and Benoît Sagot. Speeding up corpus development for linguistic research : language documentation and acquisition in Romansh Tuatschin. In *Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 89 – 94, Vancouver, Canada, August 2017.

[202] Géraldine Walther and Benoît Sagot. Morphological complexities. In *16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Florence, Italy, August 2019.

[203] Géraldine Walther, Benoît Sagot, and Karen Fort. Fast Development of Basic NLP Tools : Towards a Lexicon and a POS Tagger for Kurmanji Kurdish. In *International Conference on Lexis and Grammar*, Belgrade, Serbia, September 2010.

[204] You Zuo, Houda Mouzoun, Samir Ghamri Doudane, Kim Gerdes, and Benoît Sagot. Patent Classification using Extreme Multi-label Learning : A Case Study of French Patents. In *SIGIR 2022 - PatentSemTech workshop - 3rd Workshop on Patent Text Mining and Semantic Technologies*, Madrid, Spain, July 2022.

Mémoires

[205] Benoît Sagot. *Analyse automatique du français : lexiques, formalismes, analyseurs* . Thèse de doctorat, Université Paris-Diderot, April 2006.

[206] Benoît Sagot. *Informatiser le lexique : Modélisation, développement et exploitation de lexiques morphologiques, syntaxiques et sémantique*. Habilitation à diriger des recherches, Sorbonne Université, June 2018.

Autres publications

[207] Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. Tackling Ambiguity with Images : Improved Multimodal Machine Translation and Contrastive Evaluation. working paper or preprint, February 2023.

[208] Yu Lu Liu, Rachel Bawden, Thomas Scialom, Benoît Sagot, and Jackie Chi Kit Cheung. MaskEval : Weighted MLM-Based Evaluation for Text Summarization and Simplification. working paper or preprint, October 2022.

[209] Louis Martin, Angela Fan, Eric Villemonte de La Clergerie, Antoine Bordes, and Benoît Sagot. Multilingual Unsupervised Sentence Simplification. working paper or preprint, January 2021.

[210] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT : a Tasty French Language Model. Web site : https ://camembert-model.fr, October 2019.

[211] Louis Martin, Benoît Sagot, Éric Villemonte de La Clergerie, and Antoine Bordes. Controllable Sentence Simplification. Code and models : https ://github.com/facebookresearch/access, October 2019.

[212] Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. Between words and characters : A Brief History of Open-Vocabulary Modeling and Tokenization in NLP. 15 page preprint, January 2022.

[213] Benjamin Muller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. When Being Unseen from mBERT is just the Beginning : Handling New Languages With Multilingual Language Models. working paper or preprint, October 2020.

[214] Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. First Align, then Predict : Understanding the Cross-Lingual Ability of Multilingual BERT. Accepted at EACL 2021, March 2021.

[215] Benjamin Muller, Benoît Sagot, and Djamé Seddah. Can Multilingual Language Models Transfer to an Unseen Dialect ? A Case Study on North African Arabizi. working paper or preprint, March 2021.

[216] Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. Generative Spoken Dialogue Language Modeling : preprint version. working paper or preprint, October 2022.

[217] Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Silvie Cinková, Çağri Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-catherine De Marneffe, Valeria De Paiva, Arantza Diaz De Ilarraza, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo K. Gojenola, Memduh Gökirmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič Jr., Linh Hà My, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşikara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, John Lee, Phuong Lê Hong, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan Mcdonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Gunta Nešpore-bērzkalne, Luong Nguyen Thi, Huyen Nguyen Thi Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Robert Östling, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz L. Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel Van Niekerk, Gertjan Van Noord, Viktor Varga, Éric Villemonte de La Clergerie, Veronika Vincze, Lars Wallin, Jonathan North Washington, Mats Wirén, Tak-sum Wong, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. Universal Dependencies 2.1, November 2017. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University - Corpus - Project code : 15-10472S ; Project name : Morphologically and Syntactically Annotated Corpora of Many Languages.

[218] Charlotte Rochereau, Benoît Sagot, and Emmanuel Dupoux. Modeling German Verb Argument Structures : LSTMs vs. Humans. working paper or preprint, December 2019.

[219] Benoît Sagot. External Lexical Information for Multilingual Part-of-Speech Tagging. Research Report RR-8924, Inria Paris, June 2016.

[220] Benoît Sagot, Laurent Romary, Rachel Bawden, Pedro Ortiz Suarez, Kelly Christensen, Simon Gabay, Ariane Pinche, and Jean-Baptiste Camps. Gallic(orpor)a : Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue, December 2022. Présentation des travaux des projets lauréats de l'Appel à projet conjoint BnF DataLab/Huma-Num.

[221] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina Mcmillan-Major, Iz Beltagy, Huu Nguyen, Lucile

Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco de Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh Hajihosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael Mckenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel de Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-Aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. BLOOM : A 176B-Parameter Open-Access Multilingual Language Model. working paper or preprint, November 2022.

[222] Thomas Scialom, Louis Martin, Jacopo Staiano, Eric Villemonte de La Clergerie, and Benoît Sagot. Rethinking Automatic Evaluation in Sentence Simplification. working paper or preprint, April 2021.