

Personal information

E-mail benoit.sagot@inria.fr

Website <http://pauillac.inria.fr/~sagot/>

Professional experience

Research

- 2017–2020 **Head of Inria Paris' ALMA_nCH project-team** (Automatic Language Modelling and Analysis & Computational Humanities), *Inria Paris' research team specialised in Natural Language Processing (NLP) and Digital Humanities (DH)*, Paris, France.
Directeur de Recherche Inria (2020–), formerly Chargé de Recherche Inria (1st class, then normal class) (2017–2020)
Holder of a chair in the PRAIRIE research institute (2019–2024)
- 2014–2016 **Head of Inria Paris' ALPAGE project-team** (Analyse Linguistique Profonde à Grande Échelle), *joint research team specialised in NLP involving Inria and Université Paris-Diderot*, Paris and Rocquencourt, France.
Chargé de Recherche Inria (1st class)
- 2007–2013 **Researcher in Inria Paris' ALPAGE project-team**, Paris and Rocquencourt, France.
Chargé de Recherche Inria (2nd then 1st class)
- 2006–2007 **Researcher in Inria Futurs' SIGNES project-team** (Signes linguistiques, grammaire et sens: algorithmique logique de la langue), *an Inria research team within the UMR LabRI*, Bordeaux, France.
Chargé de Recherche Inria (2nd class)
- 2002–2006 **Ingénieur du Corps des Télécommunications seconded at Inria**, *Atoll (Atelier d'outils logiciels pour le langage naturel)*, *Inria research team*, Rocquencourt, France.
Seconded to prepare my PhD, under the supervision of Laurence Danlos (Université Paris-Diderot)
- 2002 (4 months) **DEA (Master's) work placement**, *Institut Jean-Nicod*, Paris, France.
Internship in cognitive linguistics: theoretical and comparative study of several spatial relations, under the supervision of Richard Carter (Institut Jean Nicod, EHESS)
- 2001 (6 months) **Research placement**, *IBM France*, Paris, France.
Work placement in NLP: development of a natural language understanding application, under the supervision of Claire Waast-Richard
- 2000 (4 months) **Research placement**, *Johns Hopkins University*, Baltimore, USA.
Work placement in astrophysics within the scientific team of NASA's FUSE probe (UV astronomy)

Teaching

- 2017– **Graduate (5th year)**, *ENS Cachan*, Paris, France.
Course in NLP in the Mathematics, Vision and Machine Learning curriculum (24hrs/yr, with Emmanuel Dupoux)
- 2009–2015 **Graduate (5th year)**, *Université Paris-Diderot*, Paris, France.
Course in NLP in the computational linguistic curriculum (24hrs/yr)

- 2010–2011 **Undergraduate (3rd year)**, *Université Paris-Diderot*, Paris, France.
Introductory course to natural language processing in the computer science curriculum (24hrs/yr)
- 1998–2000 **“Classes préparatoires” (2nd year)**, *École Sainte-Geneviève*, Versailles, France.
Weekly oral exams in mathematics in the highly selective course (option “MP*”) in preparation for the competitive admission examinations to French “Grandes Écoles d’Ingénieur”

Degrees

- 2018 **Habilitation à diriger les recherches in Computer Science**, *Sorbonne Université*, Paris, France,
“*Informatiser le lexique: Modélisation, développement et exploitation de lexiques morphologiques, syntaxiques et sémantiques*” [Computerising the lexicon: Modelling, development and use of morphological, syntactic and semantic lexicons],
Mentor: Laurent Romary (Inria).
Other committee members: Philippe Blache (CNRS, reviewer), James P. Blevins (University of Cambridge, reviewer), Christiane D. Fellbaum (Princeton University), Ludovic Denoyer (Sorbonne Université), Anna Korhonen (University of Cambridge), Gertjan Van Noord (University of Groningen)
- 2006 **PhD in computer science**, *Université Paris-Diderot*, Paris, France,
“*Analyse automatique du français: lexiques, formalismes, analyseurs*” [Automatic analysis of French: lexicons, formalisms, parsers], *Mention très honorable avec les félicitations du jury* [highest distinction]
Supervision: Laurence Danlos (Full professor, Université Paris-Diderot); Co-supervision: Éric Villemonte de La Clergerie (Chargé de Recherches, Inria).
Other committee members: Philippe Blache (CNRS, reviewer), Gérard Huet (Inria, reviewer), John Carroll (University of Sussex), Pierre Boullier (Inria, invited)
- 2002 **DEA (Master 2) in artificial intelligence**, *Université Pierre et Marie Curie*, Paris, France, DEA IARFA (Artificial Intelligence, Pattern Recognition and Applications).
- 2002 **Engineering degree from Télécom ParisTech as a member of the Corps des Télécommunications**, *Télécom ParisTech*, Paris, France, Specialisation in computer science, algorithmics and NLP.
- 2000 **Engineering degree from École polytechnique**, *École polytechnique*, Palaiseau, France, Specialisation in theoretical physics.

Research activities

Responsibilities within research projects

- 2019–2024 **PRAIRIE (PaRis AI Research InstitutE)** Chair (2019–2024)
Research Institute on Artificial intelligence headed by Isabelle Ryl
- 2011–2019 **LabEx EFL (Empirical Foundations of Linguistics)** Head (2011–2015) then deputy head (since 2015) of one of the 7 research strands (strand 6 “Language Resources”), member (2011–2015) then deputy member (since 2015) of the Executive Board, head of several research operations.
LabEx headed by Jacqueline Vaissière (Université Paris 3, 2011–2014) then Christian Puech (Université Paris 3, since 2014)
- 2016– **National ANR project Profiterole** Workpackage leader
Development of language resources and NLP tools for various stages of French with a focus on Old and Middle French, in order to model the diachronic evolution of the language

2015–2017	National FUI project VerDI	Local head for ALPAGE then ALMAAnaCH <i>Automatic identification of information concealment, especially in online journalistic texts</i>
2012–2016	National ANR project ASFALDA	Task co-leader <i>Development of a French FrameNet (lexicon, annotated corpus) and associated tools. PI: Marie Candito (ALPAGE)</i>
2011–2015	National FUI project PACTE	Local head for ALPAGE <i>Large-scale automatic correction of OCR outputs. Company leading the project: Numen Digital</i>
2010–2013	National ANR project EDyLex	Project leader (PI) <i>Dynamic extension of lexical resources</i>
2010–2012	Bilateral PROTEUS project (France-Slovenia)	Co-leader <i>Development and exploitation of parallel and comparable French-Slovene corpora for wordnet development. Other co-leader: Mojca Schalmberger Brezar (University of Ljubljana)</i>
2009–2011	National ANR project SEQUOIA	Local head for ALPAGE <i>Probabilistic parsing for French. PI: Alexis Nasr (Université de la Méditerranée)</i>
 Participation in research projects		
2017–	National ANR grant BASNAGE	<i>Digitalisation and automatic exploitation of Basnage's 1701 Dictionnaire Universel. PI: Laurent Romary (Inria)</i>
2016–	National ANR grant PARSITI	<i>Parsing and machine translation of noisy textual data as found on social media and more generally on the web. PI: Djamé Seddah (Université Paris-Sorbonne)</i>
2016–	Bilateral ANR-NSF project MCM-NL (France/USA)	<i>Investigation of correlations between neuroimaging data (fMRI, EEG) and automatic parsers, based on data from the Petit Prince read in both English and French. PI: John Hale (Cornell University)</i>
2015–	National ANR project SoSweet	<i>Parsing noisy data from Twitter for the joint sociolinguistic and graph-based analysis of the network of user relations. PI: Jean-Philippe Magué (ENS Lyon)</i>
2012–2018	Corpus Écrits then CORLI consortiums within the Huma-Num infrastructure	Member of the Board, participation in several research operations. In particular, participation in the CoMÉRÉ initiative led by Thierry Chanier (Université Blaise-Pascal) on computer-mediated content Consortiums led by the Institut de Linguistique Française, under the supervision of Franck Neveu (Université Paris-Sorbonne)
2012–2014	National IARPA “BABEL Program” (USA)	External expert <i>Fast and language-independent development of information retrieval systems for speech data. Expertise for the unsupervised morphological analysis task (project Lorelei) led by Owen Rambow and Nizar Habash (Columbia University)</i>
2009–2011	Bilateral ANR-DFG project PerGram (France/Germany)	<i>Development of linguistic descriptions and resources (lexicon, grammar) for Persian. PIs: Pollet Samvelian (Université Paris 3) and Stefan Müller (Freie Universität Berlin)</i>
2007–2009	National ANR project PASSAGE	<i>Automatic development and exploitation of large-scale treebanks. PI: Éric de La Clergerie (ALPAGE)</i>

2005–2007	National ILF project LexSynt	<i>Syntactic lexicons for French. PI: Sylvain Kahane (Université Paris X)</i>
2004, 2007	National Technolanguage project EASy	<i>National French parser evaluation campaign. Participant with the SxLFG parser, and indirect participant with the Lefff and the SxPipe shallow processing chain, use by both the SxLFG and the FRMG parsers)</i>
Research visits		
2013	Columbia University, USA	<i>One-week visit as part of a collaboration with Owen Rambow and Nizar Habash Formal and computational morphology, within the IARPA project “BABEL” (see above)</i>
2008–2012	University of Ljubljana, Slovenia	<i>Several one-week visits as part of a collaboration with Darja Fišer Development of wordnets for French and Slovene, partly within the above-mentioned bilateral PROTEUS project</i>
2007–2011	University of Vigo, Spain, and University of Nice, France	<i>Three short stays Development of lexical resources for French, Spanish and Galician, in relation with the Galician “Victoria” project</i>
2009	University of Padova, Italy	<i>One-week stay as part of a collaboration with Giorgio Satta Formal properties of mildly context-sensitive languages</i>
2007	Polish Academy of Sciences, Warsaw, Poland	<i>Two-month stay Computational morphology of Polish, development of morphological lexicons, improvement of the Polish national corpus</i>

Supervision

PhD supervision

2019–	Clémentine Fourier	<i>Neural models of language evolution Inria-funded PhD</i>
2019–	Robin Algayres	<i>Unsupervised segmentation and lexicon extraction from raw speech data State-funded PhD PhD director: Emmanuel Dupoux (CoML).</i>
2018–	Benjamin Muller	<i>Multi-task neural approaches for the automatic analysis of noisy texts from social media PhD thesis funded by the SoSweet and ParSiTi national (ANR) projects Other supervisor: Djamé Seddah (ALMAnaCH)</i>
2018–	Pedro Javier Ortiz Suárez	<i>Neural approaches for the automatic analysis of and information extraction from legacy dictionaries PhD thesis funded by the BASNUM national (ANR) project PhD director: Laurent Romary (ALMAnaCH)</i>
2018–	Louis Martin	<i>Automatic text simplification PhD thesis co-funded by the Facebook Artificial Intelligence Research lab Other supervisors: Éric Villemonte de La Clergerie (ALMAnaCH), Antoine Bordes (FAIR Paris, Facebook)</i>

2012–2015	Marion Baranes	<i>Normalisation of noisy text for opinion mining</i> PhD thesis co-funded by the viavoo company <i>PhD director: Laurence Danlos (ALPAGE).</i> Defended on 23rd October 2015, with honours
2011–2015	Valérie Hanoka	<i>Semi-automatic construction and extension of multilingual lexical networks</i> PhD thesis co-funded by the Verbatim Analysis company <i>PhD director: Laurence Danlos (ALPAGE).</i> Defended on 6th July 2015, with honours
2010–2013	Pierre Magistry	<i>Unsupervised Word Segmentation and Wordhood Assessment. The case for Mandarin Chinese</i> State-funded PhD <i>PhD director: Sylvain Kahane (Université Paris X); other co-supervisor: Marie-Claude Paris (Université Paris-Diderot).</i> Defended on 19th December 2013, with the highest honours
2009–2013	Rosa Stern	<i>Automatic identification of named entities to enrich textual data</i> PhD thesis co-funded by the Agence France-Presse <i>PhD director: Laurence Danlos (ALPAGE).</i> Defended on 28th June 2013, with honours

Post-doc supervision

2019–	Gaël Guibon	<i>Modelling the Old French lexicon and its diachronic evolution</i> National (ANR) project Profiterole
2016–2017	Hector Martínez Alonso	<i>Neural part-of-speech tagging, Automatic identification of omissions in journalistic texts</i> FUI project VerDI
2013–2015	Kata Gábor	<i>Identification and correction of domain-specific named entities in OCR outputs</i> FUI project PACTE
2012–2013	Damien Nouvel	<i>Dynamic extension of lexical resources using dynamic corpora (news wires)</i> National (ANR) project EDyLex
2012–2013	Yves Scherrer	<i>Lexicon development and part-of-speech tagging for non-resourced languages based on closely related resources languages</i> LabEx (excellency cluster) EFL
2010–2011	Yayoi Nakamura–Delloye	<i>Automatic extraction of relations between named entities</i> National (ANR) project EDyLex
2010–2011	Marianna Apidianaki	<i>Automatic extension of semantic lexicons (wordnets)</i> National (ANR) project EDyLex
2009–2011	Sattisvar Tandabany	<i>Algorithmics of non-contextual probabilistic parsing</i> National (ANR) project Sequoia

Master 2 dissertation supervision

2019	Charlotte Rochereau	<i>Cognitive Science master at École Normale Supérieure (CogMaster).</i> <i>Co-supervised with Emmanuel Dupoux (CoML).</i> Comparing human judgments and neural language models' assessment of argument order in German.
2014	Sacha Beniamine	<i>Computational linguistics Masters at Université Paris-Diderot.</i> Towards a linguistically motivated treatment of French multi-token units in constituency parsing

Member of PhD and Habilitation committees

2020	Jacobo Abitbol	Levy	<i>Reviewer (“rapporteur”)</i> <i>PhD in computer science, Université de Lyon, France</i> <i>PhD supervisors: Éric Fleury (Inria) and Márton Karsai (Central European University & Inria)</i>
------	-----------------------	-------------	---

- 2019 **Hazem Al Saied** *Committee member*
PhD in computer science, Université de Lorraine, France
PhD supervisors: Matthieu Constant (Université de Lorraine), Marie Candito (Université Paris-Diderot)
- 2019 **Tamara Álvarez López** *Reviewer ("rapporteur")*
PhD in computer science, University of Vigo, Spain
PhD supervisors: Enrique Costa Montenegro and Milagros Fernández Gavilanes (University of Vigo)
- 2019 **Nourredine Alliane** *Reviewer ("rapporteur")*
PhD in computer science, Université Paris 8, France
PhD supervisor: Gilles Bernard (Université Paris 8)
- 2018 **Sébastien Delecraz** *Reviewer ("rapporteur")*
PhD in computer science, Aix-Marseille Université, France
PhD supervisors: Frédéric Béchet, Alexis Nasr (Aix-Marseille Université)
- 2018 **Kim Gerdes** *Committee president*
Habilitation in linguistics, Université Paris X Nanterre, France
Habilitation mentor: Sylvain Kahane (Université Paris X Nanterre)
- 2015 **Wajdi Zaghouni** *Committee member*
PhD in computer science, Université Paris X Nanterre, France
PhD supervisor: Sylvain Kahane (Université Paris X Nanterre)
- 2014 **Édouard Grave** *Committee member*
PhD in computer science, Université Pierre et Marie Curie, France
PhD supervisors: Francis Bach (Inria) and Guillaume Obozinski (École Nationale des Ponts et Chaussées)
- 2011 **Rania Voskaki** *Committee member*
PhD in computational linguistics, Université Paris-Est Marne-la-Vallée, France
PhD supervisor: Tita Kyriakopoulou
- 2010 **Claire Mouton** *Committee member*
PhD in computer science, Université Paris Sud, France
PhD supervisor: Anne Vilnat (Université Paris Sud). Co-supervisor: Gaël de Chalendar (CEA)
- 2010 **Lionel Nicolas** *Committee member*
PhD in computer science, Université de Nice, France
PhD supervisor: Jacques Farré
- 2009 **Juan Otero Pombo** *Reviewer*
PhD in computer science, Universidade de Vigo, Spain
PhD supervisors: Manuel Vilares Ferro (Universidade de Vigo) and Jorge Graña Gil (Universidade de A Coruña)
- 2008 **Laurence Delort** *Committee member*
PhD in linguistics, Université Paris Diderot
PhD supervisor: Laurence Danlos

Community activities

- 2019– **Standardisation** *Project leader on the standardisation of inflectional morphology within the Working Group 4 (Lexical resources) of the Subcommittee 4 (Language resource management) of the Technical Committee 37 (Language and terminology)*
- 2018 **Invited co-editor** *Co-editor, with Olivier Bonami, of the special issue of the Morphology journal on computational approaches to morphology*
- 2012 **Invited co-editor** *Co-editor, with Núria Bel, of the special issue of the Traitement Automatique des Langues journal on language resources*
- 2016– **Responsibility within Inria** *Member of the Executive Board of the Inria Paris research centre*

2011–2019	Responsibility within Inria	<i>Member of the working group on international relations of the Inria Scientific and Technological Committee</i>
2015–	Scientific Board	<i>Member of the Scientific Council of the EquipEx Ortolang (French national infrastructure for language resources)</i>
2012–	Scientific Board	<i>Member of the Executive Committee of the Corpus Écrits then CORLI consortiums, within the TGIR (very large research infrastructure) Huma-Num</i>
2011–	Scientific Board	<i>Member of the Restricted Scientific Council (RSC) of the LabEx (excellency cluster) EFL, former head of the research strand 6 Language Ressources — 2011–2015, now deputy head, and alternate member of the RSC</i>
2016	Workshop co-chair	<i>Workshop on “Computational methods for descriptive and theoretical morphology” co-located with the 2016 edition of the International Morphology Meeting (Vienna, Austria). Co-organised with Olivier Bonami</i>
2011	Workshop chair	<i>WoLeR, International Workshop on Lexical Resources, co-located with the ESSLLI 2011 summer school (Ljubljana, Slovenia)</i>
2016	Expertise	<i>Expert for the French Research Evaluation Agency HCERES, member of the evaluation committee for the ATILF laboratory</i>
2011–	Expertise	<i>Expert for the French national funding agency ANR, in charge of the evaluation of research proposals in several fields (humanities and social sciences, information technology, generic)</i>
2005–	Programme and reviewing committees	<i>Member of programme or reviewing committees for a number of journals and conferences such as ACL, EAACL, CoLing, IJCNLP, Computational Linguistics, Language Resources and Evaluation, Natural Language Engineering, Journal of Language Modelling and Traitement Automatique des Langues</i>
2017–	Learned society	<i>Elected member of the Board and Deputy Treasurer of the Société de Linguistique de Paris</i>
2005–2016	Learned society	<i>Elected member of the Board (2007-2016) and former Secretary (2010-2013) of ATALA, the French learned society for NLP</i>
2014–2016	Learned society	<i>Member of the Permanent Committee of the TALN and RECITAL conferences, representing the Board of ATALA</i>

Languages

Native	French
Fluent	English
Intermediate	German
Conversational	Slovak

Publications

My full publication list can be accessed on the HAL platform (<http://hal.inria.fr>) and on my web page (<http://alpage.inria.fr/~sagot/index.html#allpubs>). A selection of my publications is given below.

My publication list includes:

- 2 book chapters,
- 12 articles in international journals, including three articles in *Language Resources and Evaluation*, two in *Linguisticae Investigationes*, one in the *Journal of Language Modelling*, one in *Indogermanische Forschungen*,

- and one (to appear) in the *Münchener Studien zur Sprachwissenschaft*,
- 7 articles in French language journals, including four in *Traitement Automatique des Langues*,
- 2 co-edited journal issues for the *Traitement Automatique des Langues* (2011) and *Morphology* (2018) journal,
- 110 articles in the proceedings of international peer-reviewed conferences, including four papers presented at ACL and two presented at CoLing.
- 40 articles in the proceedings of national peer-reviewed conferences, mostly in France but also in Spain and Slovenia

Selected publications

Apidianaki, M. and Sagot, B. (2014). Data-driven synset induction and disambiguation for wordnet development, *Language Resources and Evaluation* **48**(4): 655–677.

Beniamine, S., Bonami, O. and Sagot, B. (2018). Inferring inflection classes with description length, *Journal of Language Modelling* **5**(3): 465–525.

URL: <https://hal.inria.fr/hal-01718879>

Bonami, O. and Sagot, B. (2017). Computational methods for descriptive and theoretical morphology: a brief introduction, *Morphology* **27**(4). Special Issue: Computational Methods for Descriptive and Theoretical Morphology (ed.: O. Bonami and B. Sagot).

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

Fišer, D. and Sagot, B. (2015). Constructing a poor man's wordnet in a resource-rich world, *Language Resources and Evaluation* pp. 1–35.

Garnier, R. and Sagot, B. (2017). A shared substrate between Greek and Italic, *Indogermanische Forschungen* **122**(1): 29–60.

URL: <https://hal.inria.fr/hal-01621467>

Magistry, P. and Sagot, B. (2012). Unsupervised word segmentation: the case for Mandarin Chinese, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Jeju, Korea.

Nicolas, L., Sagot, B., Molinero, M. A., Farré, J. and Villemonte de La Clergerie, E. (2008). Computer aided correction and extension of a syntactic wide-coverage lexicon, *Proceedings of the 22nd conference of the International Committee on Computational Linguistics (CoLing 2008)*, Manchester, UK.

Sagot, B. and Boullier, P. (2008). SxPipe 2: architecture pour le traitement pré-syntaxique de corpus bruts, *Traitement Automatique des Langues* **49**(2): 155–188.

Sagot, B. and Martínez Alonso, H. (2017). Improving neural tagging with lexical information, *15th International Conference on Parsing Technologies*, Pisa, Italy, pp. 25–31.

URL: <https://hal.inria.fr/hal-01592055>

Sagot, B. and Satta, G. (2010). Optimal rank reduction for linear context-free rewriting systems with fan-out two, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden.

Sagot, B. and Villemonte de La Clergerie, E. (2006). Error mining in parsing results, *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (ACL-CoLing 2006)*, Sydney, Australia, pp. 329–336.

Sagot, B. and Villemonte de La Clergerie, E. (2008). Fouille d'erreurs sur des sorties d'analyseurs syntaxiques, *Traitement Automatique des Langues* **49**(1).

Seddah, D., Sagot, B., Candito, M., Moulleron, V. and Combet, V. (2012). The French Social Media Bank: a Treebank of Noisy User Generated Content, *Proceedings of the 24th conference of the International Committee on Computational Linguistics (CoLing 2012)*, Mumbai, India.

Villemonte De La Clergerie, É., Sagot, B. and Seddah, D. (2017). The ParisNLP entry at the ConLL UD Shared Task 2017: A Tale of a #ParsingTragedy, *Conference on Computational Natural Language Learning*, Proceedings of the CoNLL

2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, pp. 243–252.

URL: <https://hal.inria.fr/hal-01584168>

Walther, G. and Sagot, B. (2011). Modélisation et implémentation de phénomènes flexionnels non-canoniques, *Traitement Automatique des Langues* **52**(2): 91–122.