Benoît SAGOT

Personal information

E-mail benoit.sagot@inria.fr Website http://atoll.inria.fr/∼sagot/

Professional experience

Research

2017– **Chargé de Recherches Inria (1st class)**, ***Head of the ALMAnaCH research team*** *(Automatic Language Modelling and Analysis & Computational Humanities), joint team involving Inria and the École Pratique des Hautes Études*, Paris, France.

2007–2016 **Chargé de Recherches Inria (2nd, then 1st class)**, *ALPAGE (Analyse Linguistique Profonde à Grande Échelle), joint research team involving Inria and Université Paris-Diderot*, Paris and Rocquencourt, France.
Head of ALPAGE from 2014 to 2016

2006–2007 **Chargé de Recherches Inria (2nd class)**, *Signes (Signes linguistiques, grammaire et sens: algorithmique logique de la langue), Inria research team within the UMR LabRI*, Bordeaux, France.

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 natural language processing: development of a natural language understanding application, under the supervison 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 natural language processing in the Mathematics, Vision and Machine Learning curriculum (24hrs/yr)

2009–2015 **Graduate (5th year)**, *Université Paris-Diderot*, Paris, France.
Course in natural language parsing 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 natural language processing.

2000 **Engineering degree from École polytechnique**, *École polytechnique*, Palaiseau, France, Specialisation in theoretical physics.

Research activities

Responsibilities within research projects

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 ALMAnaCH *Automatic identification of information concealment, especially in online journalistic texts*
2014–	**"French Reference Corpus" Initiative**	Co-leader, together with Franck Neveu (Université de Paris-Sorbonne), of the development of a preliminary version of the future French Reference Corpus, in charge of scientific and technical aspects *Project supervised by the Institut de Linguistique Française*

2012–2016	**National ANR project ASFALDA**	Task co-leader *Development of a French FrameNet (lexicon, annotated corpus) and associated tools. PI: Marie Candito (ALPAGE)*
2011–2015	**National FUI project PACTE**	Local head for ALPAGE *Large-scale automatic correction of OCR outputs. Company leading the project: Numen Digital*
2010–2013	**National ANR project EDyLex**	**Project leader (PI)** *Dynamic extension of lexical resources*
2010–2012	**Bilateral PROTEUS project (France-Slovenia)**	**Co-leader** *Development and exploitation of parallel and comparable French-Slovene corpora for wordnet development. Other co-leader: Mojca Schalmberger Brezar (University of Ljubljana)*
2009–2011	**National ANR project SEQUOIA**	Local head for ALPAGE *Probabilistic parsing for French. PI: Alexis Nasr (Université de la Méditerrannée)*

Participation in research projects

2016–	**National ANR grant PARSITI**	*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)*
2016–	**Bilateral ANR-NSF project MCM-NL (France/USA)**	*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)*
2015–	**National ANR project SoSweet**	*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)*
2012–	**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 *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)*
2009–2011	**Bilateral ANR-DFG project PerGram (France/Germany)**	*Development of linguistic descriptions and resources (lexicon, grammar) for Persian. PIs: Pollet Samvelian (Université Paris 3) and Stefan Müller (Freie Universität Berlin)*
2007–2009	**National ANR project PASSAGE**	*Automatic development and exploitation of large-scale treebanks. PI: Éric de La Clergerie (ALPAGE)*
2005–2007	**National ILF project LexSynt**	*Syntactic lexicons for French. PI: Sylvain Kahane (Université Paris X)*

2004, 2007	**National Technolangue project EASy**	*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)*

Research visits

2013	**Columbia University, USA**	*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)
2008–2012	**University of Ljubljana, Slovenia**	*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
2007–2011	**University of Vigo, Spain, and University of Nice, France**	*Three short stays* Development of lexical resources for French, Spanish and Galician, in relation with the Galician "Victoria" project
2009	**University of Padova, Italy**	*One-week stay as part of a collaboration with Giorgio Satta* Formal properties of mildly context-sensitive languages
2007	**Polish Academy of Sciences, Warsaw, Poland**	*Two-month stay* Computational morphology of Polish, development of morphological lexicons, improvement of the Polish national corpus

Supervision

PhD co-supervision

2018–	**Louis Martin**	*Automatic text simplification* PhD thesis co-funded by the Facebook Artificial Intelligence Research lab *PhD director: Laurent Romary (ALMAnaCH).*
2012–2015	**Marion Baranes**	*Normalisation of noisy text for opinion mining* PhD thesis co-funded by the viavoo company *PhD director: Laurence Danlos (ALPAGE).* Defended on 23rd October 2015, with honours
2011–2015	**Valérie Hanoka**	*Semi-automatic construction and extension of multilingual lexical networks* PhD thesis co-funded by the Verbatim Analysis company *PhD director: Laurence Danlos (ALPAGE).* Defended on 6th July 2015, with honours
2010-2013	**Pierre Magistry**	*Unsupervised Word Segmentation and Wordhood Assessment. The case for Mandarin Chinese* State-funded PhD *PhD director: Sylvain Kahane (Université Paris X); other co-supervisor: Marie-Claude Paris (Université Paris-Diderot).* Defended on 19th December 2013, with the highest honours
2009–2013	**Rosa Stern**	*Automatic identification of named entities to enrich textual data* PhD thesis co-funded by the Agence France-Presse *PhD director: Laurence Danlos (ALPAGE).* Defended on 28th June 2013, with honours

Master 2 dissertation supervision

2014 **Sarah Beniamine** — *Towards a linguistically motivated treatment of French multi-token units in constituency parsing*
Defended on 27th June 2014

Member of PhD committees

2015 **Wajdi Zaghouani** — *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

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– **Responsibility within Inria** — *Member of the working group on international relations of the Inria Scientific and Technological Committee*

2015– **Scientific Board** — *Member of the Scientific Council of the EquipEx Ortolang (French national infrastructure for language resources)*

2012– **Scientific Board** — *Member of the Executive Committee of the Corpus Écrits then CORLI consortiums, within the TGIR (very large research infrastructure) Huma-Num*

2011– **Scientific Board** — *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*

2016 **Workshop co-chair** — *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*

2011	**Workshop chair**	*WoLeR, International Workshop on Lexical Resources, co-located with the ESSLLI 2011 summer school (Ljubljana, Slovenia)*
2016	**Expertise**	*Expert for the French Research Evaluation Agency HCERES, member of the evaluation committee for the ATILF laboratory*
2011–	**Expertise**	*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)*
2005–	**Programme and reviewing committees**	*Member of programme or reviewing committees for a number of journals and conferences such as ACL, EACL, CoLing, IJCNLP,* Computational Linguistics, Language Resources and Evaluation, Natural Language Engineering, Journal of Language Modelling *and* Traitement Automatique des Langues
2017–	**Learned society**	*Elected member of the Board and Deputy Tresurer of the Société de Linguistique de Paris*
2005–2016	**Learned society**	*Elected member of the Board (2007-2016) and former Secretary (2010-2013) of ATALA, the French learned society for NLP*
2014–2016	**Learned society**	*Member of the Permanent Committee of the TALN and RECITAL conferences, representing the Board of ATALA*

Languages

Native	**French**
Fluent	**English**
Intermediate	**German**
Conversational	**Slovak**
Notions	**Polish, Italian**

Publications

My full publication list can be accessed on the HAL platform (http://hal.inria.fr or https://hal.archives-ouvertes.fr). A selection of my publications is given below.
My publication list includes:

- 2 book chapters,
- 14 articles in international journals, including three articles in *Language Resources and Evaluation*, two in *Linguisticæ Investigationes*, four in *Traitement Automatique des Langues*, one in the *Journal of Language Modelling*, one in *Indogermanische Forschungen*, and one (to appear) in the *Münchener Studien zur Sprachwissenschaft*,
- 2 co-edited journal issues for the *Traitement Automatique des Langues* (2011) and *Morphology* (2018) journal,
- 133 articles in the proceedings of international peer-reviewed conferences, including three papers presented at ACL and two presented at CoLing.

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). Unsupervized 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., Mouilleron, 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.