Dr. Ilia Markov

Curriculum Vitae

Postdoctoral researcher
Centre for Computational Linguistics and Psycholinguistics (CLiPS),
University of Antwerp,
Antwerp, Belgium

Google Scholar • ResearchGate • LinkedIn

Personal information

- Phone: (+32) 456-12-35-70
- E-mail: ilia.markov@uantwerpen.be
- Nationality: Russian
- Date of birth: 05/08/1983

Work experience

- **Postdoctoral researcher**, June 2019 – present
  Centre for Computational Linguistics and Psycholinguistics (CLiPS), University of Antwerp, Antwerp, Belgium
  Project: “The Linguistic Landscape of Hate Speech in Social Media”
  Advisor: Dr. Walter Daelemans
- **Postdoctoral researcher**, June 2018 – May 2019
  Automatic Language Modelling and Analysis & Computational Humanities (ALMAnaCH) Team, French Institute for Research in Computer Science and Automation (INRIA), Paris, France
  Project: “Automatic Analysis of Citizens’ Debates”
  Advisor: Dr. Eric Villemonte De la Clergerie

Education

- **PhD** (with honors), **Computer Science**, 2014 – 2018, **GPA 10/10**
  Natural Language Processing Laboratory, Center for Computing Research (CIC), Instituto Politécnico Nacional (IPN), Mexico City, Mexico
  PhD thesis: “Automatic Native Language Identification”
  Advisors: Dr. Grigori Sidorov, Dr. Obdulia Pichardo Lagunas
- **MSc, Language Sciences**, 2012 – 2014, **GPA 17/20**
  University of Algarve, Faro, Portugal
  MSc thesis: “Automatic Identification of Whole-Part Relations in Portuguese”
  Advisors: Dr. Jorge Baptista, Dr. Nuno Mamede
- **BSc** (with honors), **Computer Engineering**, 2001 – 2006, **GPA 5/5**
  Kaliningrad State Technical University, Russia
  BSc thesis: “Automatic Estimation of Stocking Norms”
  Advisor: Dr. Olga Toporkova

Research experience

- Native language identification, native language interference, discriminating between similar languages, author profiling, authorship attribution
- Sentiment analysis
- Computational semantics, semantic relations extraction, deep syntactic parsing
- Information retrieval, information extraction, text mining, argument mining, text similarity
- Corpus linguistics, lexical resources: dictionaries, ontologies
- Human-computer interaction

Teaching experience

- Natural language processing (Instituto Politécnico Nacional)
- Information retrieval (Instituto Politécnico Nacional)

Research internships

- March – July 2017: Fondazione Bruno Kessler, Trento, Italy, under Dr. Carlo Strapparava
- May – June 2015: University of the Aegean, Karlovassi, Samos, Greece, under Dr. Efstathios Stamatatos
- September 2013 – July 2014: Spoken Language Systems Laboratory (L2F), INESC-ID Lisboa, Lisbon, Portugal, under Dr. Nuno Mamede
- August 2005 – January 2006: Satakunta University of Applied Sciences, Pori, Finland

Scientific societies

- 2017, 2019: Association for Computational Linguistics (ACL)
- 2018 – present: Special Interest Group for NLP in Education (SIG EDU)
- 2015 – present: Mexican Network for Language Technologies
- 2014 – present: Mexican Society of Artificial Intelligence
- 2014 – present: Mexican Association of Natural Language Processing

Participation in research projects

- 2015–2018: **Automatic evaluation of semantic similarity of texts using syntactic n-grams and integrated syntactic graphs,** grant CONACYT 240844, Mexican Government. My responsibilities included building syntactic n-grams of various types.

- 2017: **Applications of convolutional neural networks for the analysis of social networks,** grant SIP 20172008, Mexican Government. My responsibilities included implementation of the author profiling methods and identification and interpretation of author's social group.

- 2016: **Multi-Labeled Corpus of News in Spanish**, grants of the Mexican government 260178 and 271622 for the collaboration between researchers and research students working in natural language processing. My responsibilities included building a corpus of news in Spanish annotated with the varieties of the Spanish language, author, gender of the author, and topic; download.

- 2016: **Automatic question answering based on semantic and syntactic similarity,** grant SIP 20161947, Mexican Government. My responsibilities included developing of methods for syntactic similarity of texts.

- 2015: **Social Media Lexicon**, grants of the Mexican government 260178 and 271622 for the collaboration between researchers and research students working in natural language processing. My responsibilities included building a lexical resource for social media: slang words, contractions, abbreviations, and emoticons commonly used in social media for English, Spanish, Dutch, and Italian; download.

- 2015: **Syntactic and semantic analysis of texts applied to education, law, and social networks,** grant SIP 20152100, Mexican Government. My responsibilities included semiautomatic compilation of ontologies and disambiguation with the application of linguistic traits for machine learning.

- 2015: **Development of a corpus of program code and its annotation for automatic classification,** grant SIP 20151406, Mexican Government. I was responsible for data collection and annotation of the corpus.

- 2014: **Automatic identification of author using continuous and non-continuous syntactic n-grams,** grant SIP 20144274, Mexican Government. My responsibilities included building continuous and non-continuous syntactic n-grams of words and of tags of syntactic relations.

- 2013 – 2014: STRING **– A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese.** My responsibilities included improving the extraction of semantic relations between textual elements by targeting meronymy relations. I built a rule-based meronymy extraction module and integrated it in the grammar of the STRING system.

Participation in shared tasks on application of machine learning to NLP

- **Indian Native Language Identification (INLI) shared task at FIRE 2018.** Indian Native Language Identification using a machine-learning approach (ranked 3rd out of 14).
- **PAN RUSProfiling shared task at FIRE 2017.** Cross-genre gender identification in Russian texts using machine-learning and statistical approaches **(ranked 1st out of 22 systems).**
- **Native Language Identification (NLI) shared task at EMNLP 2017.** Native Language Identification using a combination of word- and character-level features **(ranked 1st out of 17).**
- **5th PAN Author Profiling competition at CLEF 2017.** Gender and language variety identification in English, Spanish, Portuguese, and Arabic adjusting feature selection and classifier parameters to each language and subtask (ranked 6th out of 22).
- **4th VarDial Discriminating between Similar Languages competition at EACL 2017.** Discriminating between similar languages within 6 language groups using typed and untyped character n-gram features and lexical features (ranked 6th out of 11).
- **4th PAN Author Profiling competition at CLEF 2016.** Cross-genre age and gender identification in English, Spanish, and Dutch using so-called transition point technique with typed character n-grams, lexical features, and non-lexical features (ranked 5th out of 22).
- **3rd PAN Author Profiling competition at CLEF 2015.** Age, gender, and personality traits identification in English, Spanish, Dutch, and Italian using syntactic dependency based n-grams of various types (ranked 17th – 19th out of 22, depending on language).
- **PAN Authorship Identification competition at CLEF 2015.** Author identification with textual patterns based on features obtained from shortest path walks over the Integrated Syntactic Graphs (novel structure introduced by our team) (ranked 11th out of 18).

Program committee member

- The Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019), the lexical semantics and word representations track
- 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT 2019), the text mining track
- 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA13) at NAACL-HLT 2018

Reviewer

Journals

- Cognitive Computation. JCR impact factor: 3.44

Conferences and workshops

- 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)
- 4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)
- 17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2016)
- 15th Mexican International Conference on Artificial Intelligence (MICAI 2016)

Advisor

BSc thesis: "Syllables as Features for the Authorship Attribution Task", ESIME, Instituto Politécnico Nacional (IPN).

Skills

- **Linguistics**: computational linguistics, corpus linguistics, discourse analysis, semantics, syntax, rule-based approaches
- **Programming languages and packages:** Python, NLTK, spaCy, scikit-learn, FreeLing, Gensim, Stanford Core NLP, WEKA
- **Operating systems:** MacOS, Linux, Windows
- **Languages:** English (fluent), Spanish (fluent), Portuguese (intermediate), French (Beginner), Russian (native)

Certificates

Universitat Politècnica de Catalunya course of Approaches to Machine Translation: Rule-based, Statistical and Hybrid (grade 118.7 out of 120, online)

Awards and scholarships

- **Best academic performance of PhD students**, Instituto Politécnico Nacional (IPN) 2017. There are ca. 1,500 PhD students at IPN. The diploma is awarded for best scores and publications.
- **1st rank (**out of 22 systems) in the PAN RUSProfiling shared task at FIRE 2017.
- **1st rank (**out of 17) in the Native Language Identification shared task at EMNLP 2017.
- **Best paper award** (1st place) at the 15th Mexican International Conference on Artificial Intelligence (MICAI 2016), for the paper *Author Profiling with doc2vec Neural Network-Based Document Embeddings*.

- **Mexican Government Scholarship** for obtaining a PhD degree, 2014 – 2018.
- **Research Grant BEIFI** of the Instituto Politécnico Nacional (IPN), 2014 – 2018.
- **Erasmus Mundus Action 2 2011-2574 Triple I - Integration, Interaction and Institution** for obtaining MSc at the University of Algarve, Faro, Portugal, 2012 – 2014.
- **Governor of the Region Scholarship** for showing the best academic results at Kaliningrad State Technical University, Kaliningrad, Russia, 2002 – 2006.

Publications (31)

Journals (8)

1. H. Gómez, R. Fuentes, I. Markov, G. Sidorov, A. Gelbukh. A Convolutional Neural Network Approach for Gender and Language Variety Identification. *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4845–4855, 2019. (DOI)

   JCR **impact factor**: 1.426

2. G. Sidorov, I. Markov, O. Kolesnikova, L. Chanona. Human Interaction with Shopping Assistant Robot in Natural Language. *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4889–4899, 2019. (DOI)

   JCR **impact factor**: 1.426

3. I. Markov, J. Baptista, O. Pichardo. Authorship Attribution in Portuguese Using Character N-grams. *Acta Polytechnica Hungarica*, vol. 14, no. 3, pp. 59–78, 2017. (DOI, PDF)

   JCR **impact factor**: 0.745

4. G. Sidorov, M. Ibarra, I. Markov, R. Guzman, L. Chanona, F. Velásquez. Measuring Similarity Between Karel Programs Using Character and Word N-grams. *Programming and Computer Software*, vol. 43, no. 1, pp. 47–50, 2017. (DOI, PDF preprint)

   JCR **impact factor**: 0.230

5. H. Gómez, I. Markov, G. Sidorov, J.-P. Posadas, M. Sanchez, L. Chanona. Improving Feature Representation Based on a Neural Network for Author Profiling in Social Media Texts. *Computational Intelligence and Neuroscience*, vol. 2016, 13 pages, 2016. (DOI)

   JCR **impact factor**: 1.215

6. G. Sidorov, M. Ibarra, I. Markov, R. Guzman, L. Chanona, F. Velásquez. Automatic Detection of Similarity of Programs in Karel Programming Language based on Natural Language Processing Techniques. *Computación y Sistemas*, vol. 20, no. 2, pp. 279–288, 2016. (DOI, PDF)

   Scopus Q2

7. H. Gómez, I. Markov, G. Sidorov, J.-P. Posadas, C. Fócil. Compiling a Lexicon of Social Media for the Author Profiling Task. *Research in Computing Science*, vol. 115, pp. 19–27, 2016. (PDF)

   DBLP

8. I. Markov, N. Mamede, J. Baptista. A Rule-Based Meronymy Extraction Module for Portuguese. *Computación y Sistemas*, vol. 19, no. 4, pp. 661–683, 2015. (DOI, PDF)

   Scopus Q2

Conferences and workshops (22)

9. I. Markov, V. Nastase, C. Strapparava. Anglicized Words and Misspelled Cognates in Native Language Identification. In: *14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA14 2019),* Florence, Italy. ACL, pp. 275–284, August 2, 2019. (PDF)

10. I. Markov, E. De la Clergerie. INRIA at SemEval-2019 Task 9: Suggestion Mining Using SVM with Handcrafted Features. In: *13th International Workshop on Semantic Evaluation (SemEval-2019)*, Minneapolis, Minnesota, USA. ACL, pp. 1204–1207, June 6–7, 2019. (PDF)

11. I. Markov, G. Sidorov. CIC-IPN@INLI2018: Indian Native Language Identification. In: *Working Notes of FIRE 2018 – 10th International Forum for Information Retrieval Evaluation*, Gandhinagar, India. CEUR-WS.org, vol. 2266, pp. 82–88, December 06–09, 2018. (PDF)

12. I. Markov, V. Nastase, C. Strapparava, G. Sidorov. The Role of Emotions in Native Language Identification. In: *9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2018),* Brussels, Belgium. ACL, pp. 123–129, October 31, 2018. (DOI, PDF)

13. I. Markov, H. Gómez, M. Jasso-Rosales, G. Sidorov. CIC-GIL Approach to Author Profiling in Spanish Tweets: Location and Occupation. In: *Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, Seville, Spain. CEUR-WS.org, vol. 2150, pp. 97–101, September 18, 2018. (PDF)

14. I. Markov, V. Nastase, C. Strapparava. Punctuation as Native Language Interference. In: *27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA. The COLING 2018 Organizing Committee, pp. 3456–3466, August 20–26, 2018. (PDF)

15. I. Markov, H. Gómez, G. Sidorov, A. Gelbukh. The Winning Approach to Cross-Genre Gender Identification in Russian at RUSProfiling 2017. In: *Working Notes of FIRE 2017 – 9th International Forum for Information Retrieval Evaluation*, Bangalore, India. CEUR-WS.org, vol. 2036, pp. 20–24, December 08-10, 2017. (PDF)

    Ranked 1st in the PAN RUSProfiling shared task 2017.

16. I. Markov, L. Chen, C. Strapparava, G. Sidorov. CIC-FBK Approach to Native Language Identification. In: *12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA12 2017),* Copenhagen, Denmark. ACL, pp. 374–381, September 8, 2017. (PDF)

    Ranked 1st in the NLI shared task 2017

17. M. Sanchez, I. Markov, H. Gómez, G. Sidorov. Comparison of Character n-grams and Lexical Features on Author, Gender, and Language Variety Identification on the Same Spanish News Corpus. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2017)*, Dublin, Ireland. LNCS, Springer, vol. 10456, pp. 145–151, September 11–14, 2017. (DOI)

18. I. Markov, H. Gómez, G. Sidorov. Language- and Subtask-Dependent Feature Selection and Classifier Parameter Tuning for Author Profiling. In: *Working Notes of CLEF 2017 – Conference and Labs of the Evaluation Forum*, Dublin, Ireland. CEUR, vol. 1866, September 11–14, 2017. (PDF)

19. I. Markov, E. Stamatatos, G. Sidorov. Improving Cross-Topic Authorship Attribution: The Role of Pre-Processing. In: *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017),* Budapest, Hungary. Springer, vol. 10762, pp. 289–302, April 17–23, 2017. (DOI, PDF preprint)

    Best poster award, third place

20. H. Gómez, I. Markov, J. Baptista, G. Sidorov, D. Pinto. Discriminating between Similar Languages Using a Combination of Typed and Untyped Character N-grams and Words. In: *4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*, Valencia, Spain. ACL, pp. 137–145, April 3, 2017. (PDF)

21. I. Markov, H. Gómez, J.-P. Posadas, G. Sidorov, A. Gelbukh. Author Profiling with Doc2vec Neural Network-Based Document Embeddings. In: *15th Mexican International Conference on Artificial Intelligence (MICAI 2016)*, Cancún, Mexico. Part II, LNAI, Springer, vol. 10062, pp. 117–131, October 23–29, 2016. (DOI, PDF preprint)

    Best paper award, first place

22. I. Markov, H. Gómez, G. Sidorov, A. Gelbukh. Adapting Cross-Genre Author Profiling to Language and Corpus. In: *Working Notes of CLEF 2016 – Conference and Labs of the Evaluation Forum*, Évora, Portugal. CEUR, vol. 1609, pp. 947–955, September 5–8, 2016. (PDF)

23. G. Sidorov, H. Gómez, I. Markov, D. Pinto, N. Loya. Computing Text Similarity using Tree Edit Distance. In: *Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, joint with 2015 5th World Conference on Soft Computing (WConSC), Redmond, WA, USA. IEEE, pp. 1–4, August 17–19, 2015. (DOI, PDF)

24. H. Gómez, G. Sidorov, D. Pinto, I. Markov. A Graph Based Authorship Identification Approach. In: *Working Notes of CLEF 2015 – Conference and Labs of the Evaluation Forum*, Toulouse, France. CEUR, vol. 1391, September 8–11, 2015. (PDF)

25. J.-P. Posadas, I. Markov, H. Gómez, G. Sidorov, I. Batyrshin, A. Gelbukh, O. Pichardo. Syntactic N-grams as Features for the Author Profiling Task. In: *Working Notes of CLEF 2015 – Conference and Labs of the Evaluation Forum*, Toulouse, France. CEUR, vol. 1391, September 8–11, 2015. (PDF)

26. I. Markov, N. Mamede, J. Baptista. Whole-Part Relations Rule-Based Automatic Identification: Issues from Fine-Grained Error Analysis. In: *13th Mexican International Conference on Artificial Intelligence (MICAI 2014)*, Tuxtla Gutiérrez, Mexico. Springer, vol. 8856, pp. 37–50, November 16–22, 2014. (DOI, PDF priprint)

27. I. Markov, N. Mamede, J. Baptista. Automatic Identification of Whole-Part Relations in Portuguese. In: *3rd Symposium on Languages, Applications and Technologies (SLATE 2014)*, Bragança, Portugal. Dagstuhl Publishing, vol. 38, pp. 225–232, June 19–20, 2014. (DOI, PDF)

28. J. Baptista, N. Mamede, I. Markov. Integrating Verbal Idioms into an NLP System. In: *11th International Conference on the Computational Processing of the Portuguese Language (PROPOR 2014)*, São Carlos, SP, Brazil. Springer, vol. 8775, pp. 250–255, October 6–9, 2014. (DOI, PDF preprint)

29. I. Markov, N. Mamede, J. Baptista. Body-Part Nouns and Whole-Part Relations. In: *11th International Conference on the Computational Processing of the Portuguese Language (PROPOR 2014)*, São Carlos, SP, Brazil. Springer, vol. 8775, pp. 125–136, October 6–9, 2014. (DOI, PDF preprint)

30. J. Baptista, N. Mamede, I. Markov. Integrating a Lexicon-Grammar of Verbal Idioms in a Portuguese NLP System. In: 2nd *PARSEME General Meeting,* Athens, Greece, March 10–11, 2014.

Chapters (1)

31. J. Baptista, I. Markov. Morphosyntactic processes involving body-part nouns in Portuguese. In: *Perspectives Harrissiennes*. CRL - Cellule de Recherche en Linguististique, pp. 255–267, 2016. (PDF)

Submitted (4)

1. I. Markov, G. Sidorov, A. Gelbukh. Artificial Intelligence Research in Mexico. *AI Magazine.*

2. I. Markov, V. Nastase, C. Strapparava. Exploiting Native Language Interference in the Native Language Identification Task.

3. I. Markov, V. Nastase, C. Strapparava. How Different Linguistics Characteristics Appear in Different L2s: The Case of English and Portuguese.

References: Alexander Gelbukh, Grigori Sidorov, Carlo Strapparava, Efstathios Stamatatos, Jorge Baptista.