

**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](mailto:ilia.markov@uantwerpen.be)
- 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 (ALMAAnaCH) 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 (Natural Language Processing, AI)

- Hate speech and abusive language detection
- 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

- 2020–2021: Natural Language Processing, UA\_2007FLWDTA (University of Antwerp), with Dr. Walter Daelemans: morphology and syntax, computational semantics, deep learning
- 2019–2020: Computational Linguistics, UA\_2010FLWTAA (University of Antwerp), with Dr. Walter Daelemans: syntactic analysis and parsing, computational semantics
- 2017–2018: Natural language processing (Instituto Politécnico Nacional)
- 2016–2017: Information retrieval (Instituto Politécnico Nacional)

## Co-advisor

- 2020: MSc thesis “Cross-Lingual Age Detection on English and German Tweets: Evaluating the Impact of Individual Parts of Speech”, CLiPS, University of Antwerp
- 2018: 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, TensorFlow, PyTorch, Keras, NLTK, pandas, spaCy, scikit-learn, FreeLing, Gensim, Stanford Core NLP
- **Operating systems:** MacOS, Linux, Windows
- **Languages:** English (fluent), Spanish (fluent), Portuguese (intermediate), Dutch (beginner), Russian (native)

## Publications

### Journals

1. I. Markov, V. Nastase, C. Strapparava. Exploiting Native Language Interference for Native Language Identification. *Natural Language Engineering*, pp. 1–31, 2020. ([DOI](#))  
JCR impact factor: 1.465
2. 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
3. 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
4. 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
5. 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
6. 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

7. 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
8. 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
9. 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 chapters in books

1. I. Markov, N. Ljubešić, D. Fišer, W. Daelemans. Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection. In: *Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2021)*, Online, ACL, pp. 149–159, April 19, 2021. ([PDF](#))
2. J. Bevendorff, B. Chulvi, G. L. De La Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle. Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection. In: *43<sup>rd</sup> European Conference on Information Retrieval (ECIR 2021)*, Online, LNCS, Springer, vol. 12657, pp. 567–573, March 28 – April 1, 2021. ([DOI](#))
3. N. Ljubešić, I. Markov, D. Fišer, W. Daelemans. The LiLaH Emotion Lexicon of Croatian, Dutch and Slovene. In: *Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media (PEOPLES 2020)*, Barcelona, Spain (Online), ACL, pp. 153–157, December 13, 2020. ([PDF](#))
4. E. Lotfi, I. Markov, W. Daelemans. A Deep Generative Approach to Native Language Identification. In: *28<sup>th</sup> International Conference on Computational Linguistics (COLING 2020)*, Barcelona, Spain (Online), International Committee on Computational Linguistics, pp. 1778–1783, December 8–13, 2020. ([DOI](#))
5. J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, E. Zangerle. Overview of PAN 2020: Authorship Verification, Celebrity Profiling, Profiling Fake News Spreaders on Twitter, and Style Change Detection. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2020)*, Thessaloniki, Greece. LNCS, Springer, vol. 12260, pp. 372–383, September 22–25, 2020. ([DOI](#))

6. M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, B. Stein. Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In: *Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece. CEUR, vol. 2696, September 22–25, 2020. ([PDF](#))
7. J. Lemmens, B. Burtenshaw, E. Lotfi, I. Markov, W. Daelemans. Sarcasm Detection Using an Ensemble Approach. In: *Second Workshop on Figurative Language Processing (FigLang 2020)*, Online, ACL, pp. 264–269, July 9, 2020. ([DOI](#))
8. I. Markov, V. Nastase, C. Strapparava. Anglicized Words and Misspelled Cognates in Native Language Identification. In: *14<sup>th</sup> Workshop on Innovative Use of NLP for Building Educational Applications (BEA14 2019)*, Florence, Italy. ACL, pp. 275–284, August 2, 2019. ([PDF](#))
9. I. Markov, E. De la Clergerie. INRIA at SemEval-2019 Task 9: Suggestion Mining Using SVM with Handcrafted Features. In: *13<sup>th</sup> International Workshop on Semantic Evaluation (SemEval-2019)*, Minneapolis, Minnesota, USA. ACL, pp. 1204–1207, June 6–7, 2019. ([PDF](#))
10. I. Markov, G. Sidorov. CIC-IPN@INLI2018: Indian Native Language Identification. In: *Working Notes of FIRE 2018 – 10<sup>th</sup> International Forum for Information Retrieval Evaluation*, Gandhinagar, India. CEUR-WS.org, vol. 2266, pp. 82–88, December 06–09, 2018. ([PDF](#))
11. I. Markov, V. Nastase, C. Strapparava, G. Sidorov. The Role of Emotions in Native Language Identification. In: *9<sup>th</sup> Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2018)*, Brussels, Belgium. ACL, pp. 123–129, October 31, 2018. ([DOI](#), [PDF](#))
12. 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](#))
13. I. Markov, V. Nastase, C. Strapparava. Punctuation as Native Language Interference. In: *27<sup>th</sup> 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](#))
14. 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 – 9<sup>th</sup> International Forum for Information Retrieval Evaluation*, Bangalore, India. CEUR-WS.org, vol. 2036, pp. 20–24, December 08–10, 2017. ([PDF](#))  
**Ranked 1<sup>st</sup> in the PAN RUSProfiling shared task 2017.**

15. I. Markov, L. Chen, C. Strapparava, G. Sidorov. CIC-FBK Approach to Native Language Identification. In: *12<sup>th</sup> Workshop on Innovative Use of NLP for Building Educational Applications (BEA12 2017)*, Copenhagen, Denmark. ACL, pp. 374–381, September 8, 2017. ([PDF](#))  
**Ranked 1<sup>st</sup> in the NLI shared task 2017**
16. 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](#))
17. 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](#))
18. I. Markov, E. Stamatatos, G. Sidorov. Improving Cross-Topic Authorship Attribution: The Role of Pre-Processing. In: *18<sup>th</sup> 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
19. 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: *4<sup>th</sup> Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*, Valencia, Spain. ACL, pp. 137–145, April 3, 2017. ([PDF](#))
20. I. Markov, H. Gómez, J.-P. Posadas, G. Sidorov, A. Gelbukh. Author Profiling with Doc2vec Neural Network-Based Document Embeddings. In: *15<sup>th</sup> 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**
21. 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](#))
22. J. Baptista, I. Markov. Morphosyntactic processes involving body-part nouns in Portuguese. In: *Perspectives Harrissiennes*. CRL - Cellule de Recherche en Linguistique, pp. 255–267, 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 5<sup>th</sup> 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: *13<sup>th</sup> Mexican International Conference on Artificial Intelligence (MICAI 2014)*, Tuxtla Gutiérrez, Mexico. Springer, vol. 8856, pp. 37–50, November 16–22, 2014. ([DOI](#), [PDF preprint](#))
27. I. Markov, N. Mamede, J. Baptista. Automatic Identification of Whole-Part Relations in Portuguese. In: *3<sup>rd</sup> 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: *11<sup>th</sup> 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: *11<sup>th</sup> 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: *2<sup>nd</sup> PARSEME General Meeting*, Athens, Greece, March 10–11, 2014.

**References:** [Dr. Walter Daelemans](#), [Dr. Alexander Gelbukh](#), [Dr. Grigori Sidorov](#)