HABILITATION THESIS

SIMILARITY AND DECISION PROBLEMS IN COMPUTATIONAL LINGUISTICS
-SUMMARY-

LIVIU P. DINU

Chapter 1

Summary

Most of my research activity is conducted in the area of mathematical and computational linguistics and natural language processing. My interest in this research area stems from the the period of my Ph.D. studies when, under the guidance of my supervisor Solomon Marcus, I dedicated myself to the research activity. Solomon Marcus is one of the initiators of this research field in Romania and, worldwide, one of the first researchers in this domain. Therefore, I believe that our results and the present work can serve as a continuation of the tradition of the Romanian school of computational linguistics and can attract new researchers, especially from among the students. There is no longer a surprise that the results from a research field can find their echoes in different areas, often far remote from the area which inspired them. Given the need to identify a relationship of similarity between the Romance languages, we developed a metric that has found applications in the field of bio-informatics (particularly in the similarity of the DNA chains) and in the more general areas of string processing and aggregation and multi-criteria classification problems. These research issues constitute the subject of the present habilitation thesis, whose results and findings we will try to summarize below.

The results presented here are based on the 86 published papers (out of which 75 were published after defending the Ph.D. thesis), almost all of them published in international peer-reviewed journals or conference proceedings, on 2 books and 4 chapters published in books. The papers published after defending the Ph.D. thesis can be divided, according to the current standards, in 18 papers A category, 23

papers B category and 18 papers C category (plus 16 other papers), and, of these, 20 papers are indexed by ISI Web of Science, and 9 more papers are to be indexed by ISI (they are published in the proceedings of ISI rated conferences).

Starting from the necessity of identifying an appropriate metric to be used in the analysis of the natural language similarity, in [16] we proposed a metric called rank distance [17]. In chapter 3, we present the main mathematical and computational properties of this metric (most of them without proofs included) obtained until today. We address problems of maximum and average on rankings and strings (with equal or different length, and with the same or different composition) [51] [18], its properties of collinearity [52], the density of a segment [48], and we propose two linear algorithms (without indexing) for computing the metric [49].

Strongly related to chapter 3, in chapter 4 the transition to two important problems in theoretical computer science and their applications, namely the analysis of the median string and the closest string, is presented. Both problems were initiated from the Hamming distance, and, having demonstrated their polynomial intractability (for both the Hamming and the edit distance)[54], the existence of a metric that leads to an efficient (polynomial) approach to determining at least the median string has become an open problem. In chapter 4 we show that there is a polynomial approach for the median string via the rank distance [35], while the closest string remains intractable [42]. We propose a heuristic for determining the closest string [28]. We transform both problems in a problem of aggregation [16] and multi-criteria categorization [44], we make the junction with Arrow's impossibility theorem, we study the rationality criteria proposed by Arrow and expanded by Păun [16] [44], we propose and investigate a new method of clustering based on rank distance and the two aggregation problems [27] [31] [34], and we present two specific applications of the proposed technique of categorization: the identification of handwritten digits [44] [32] and the automatic categorization of texts [47]. We also study the consensus problem, by attempting to simultaneously optimize the two problems [33]. We also present here the main results obtained in bioinformatics, and specifically in the analysis of the similarity of the DNA chains. We analyzed whether the good computational and mathematical properties of the rank distance make it suitable for investigating the similarity of the DNA chains [49] [23] [50]. We studied whether an ordinal distance that counts rather distances between

the off-sets of two elements can perform well in this highly competitive research area. The good results we obtained were highlighted by applying the rank distance on the mithocondrial DNA of mammals, the final phylogenetic tree we obtained being comparable with the standards reported in this field [49] [34]. We investigated the same problem by applying the central string via rank distance [29] [31] and by employing an alignment algorithm based on LRD [30], a metric derived from the rank distance.

Starting with the 5th chapter we present the main results obtained in the area of computational linguistics, grouped into topics. thus, chapter 5 is devoted to a linguistic unit which was less studied than others, namely the syllable. Continuing the concerns from the period of the Ph.D. studies [15], we proposed a new formal approach to syllabification and implicitly to the syllable, based on insertion grammars [11] [19]. We present here a new formalization, namely a parallel model for syllabification, together with the main results obtained regarding the classification of the newly proposed class of languages [20]. We also introduced a new class of contextual grammars (called syllabic grammars) [25] and made the junction with the go-through automata [24], introduced by Gramatovici [53]. The quantitative aspects were investigated with respect to the entire dictionary of Romanian [12], and we presented a series of quantitative laws and their application on Romanian [13].

Chapter 6 is devoted to the results obtained on several important issues regarding the Romanian morphology. Romanian is a language with a rich morphology, and the current research results show that such languages have received an increasing attention lately. An important research problem is detecting the Romanian verb alternations and its conjugation in the present tense. We addressed this issue in a series of papers [26] [37], and we obtained 95% accuracy for predicting verb alternations. We also proposed a formal mechanism for resolving alternations, starting from the concept of variable letters introduced by Moisil [26] in 1965. Another problem presented here is the analysis of the plural form of the nouns and automatic gender detection [36] [39]. The main question here is the automatic detection of the neuter. We presented a technique for learning syllabification for Romanian words [40], using RoMorphoDict [1] as a database. Another research problem we addressed in this chapter is predicting the stress placement. Most linguists agree that the stress in Romanian is unpredictable (most recently, Pana-Dindelegan). We

proposed a system for learning stress placement [3] [2], and we were able to predict the position of the primary stress for Romanian words with 98% accuracy.

In chapter 7, we investigate and analyze problems of similarity and related words with respect to Romanian. Based on the problem of the syllabic similarity of the Romance languages [10], presented in the Ph.D. thesis and continued thereafter [17], we conducted a quantitative evaluation of the previous results [22], [21] and an extension of the investigation from the perspective of understanding texts written in related languages, i.e., their intelligibility [8]. We considered that a good starting point was detecting pairs of cognates between Romanian and other languages, and investigating their behavior [4] [6]. This problem is currently extensively studied, having a multitude of applications in machine translation, second language acquisition, language evolution, and so on. We proposed an algorithm for automatically extracting pairs of cognates between Romanian and any other language for which electronic resources are available [4], we determined the pairs of cognates between Romanian and the Romance languages and Turkish, we investigated the ability of deciding if two words are cognates or not using alignment algorithms [9] and suitable metrics [6], we investigated if we are able to distinguish between cognate pairs and word-etymon pairs, and we identified the orthographic changes undergone by words when entering a new language [5]. We proposed a new approach to quantifying the similarity between Romanian and other languages, and we presented the main results, which bring a new perspective on the similarity between Romanian and other Indo-European languages.

Chapter 8 is devoted to the authorship problems. We propose a new approach by applying rank distance on rankings of functional words corresponding to the investigated works [58]. We were among the first to have used, with good results, an approach based on kernels in authorship problems [57]. To evaluate our methods, we used standard texts, such as the Federalist Papers [43], texts on which previous research results were reported, and also controversial Romanian texts [59] [46]. We also investigated the stylistic unity of the Epistles of the Apostle Paul [14] and the results were always positive. We propose the use of others metrics as well, and we generally show the superiority of ordinal metrics [45]. We study translationese problems (especially the discrimination between works written in a language and those translated into that language) [55] [41] and problems of deception detection

and pastiche detection (authors who attempt to impersonate a given author) [38]. All our presentations are accompanied by applications.

The last chapter discusses the plan for academic and scientific development, by presenting the research topics in our interest and ways to achieve these objectives.

Chapter 2

Rezumat

Cea mai mare parte a activităţii mele de cercetare se desfăşoară în aria lingvisticii matematice şi computaţionale şi a procesării limbajului natural. Interesul pentru aceasta zonă de cercetare începe din anii doctoratului, sub îndrumarea coordonatorului meu, Acad. Solomon Marcus. Solomon Marcus este unul dintre iniţiatorii acestui domeniu in România, şi, pe plan mondial, se află printre primii cercetători care s-au dedicat acestui domeniu. În acest fel, considerăm că rezultatele noastre şi lucrarea de faţă se pot constitui ca o continuare a tradiţiei şcolii româneşti de lingvistică computaţională şi pot atrage noi cercetători în special din rândul studenţilor. Nu mai este o surpriză că rezultatele dintr-un domeniu îşi pot găsi ecoul în domenii diferite, adeseori aflate la mare distanţă de zona de la care au fost inspirate. Pornind de la necesitatea identificării unei relaţii de similaritate între limbile romanice, am dezvoltat o distanţă care şi-a găsit aplicaţii în zona bioinformaticii (în special a similarităţii lanţurilor ADN) şi in zona mai generală a procesării stringurilor şi a problemelor de agregare şi clasificare multicriterială. Toate acestea fac obiectul prezentei teze de abilitare, ale căror rezultate vom încerca sa le sintetizăm mai jos.

Rezultatele prezentate aici se bazează pe cele 86 de articole publicate (dintre care 75 dupa susţinerea tezei de doctorat), aproape toate apărute in jurnale şi/sau volume peer-reviewed ale unor conferinţe internaţionale, pe 2 cărti publicate şi pe 4 articole publicate în cărti.

Articolele publicate dupa susţinerea tezei pot fi împărţite, conform normelor actuale, în 18 lucrări categoria A, 23 categoria B si 18 lucrări categoria C (plus

16 alte lucraări), iar, dintre acestea, 20 lucrări sunt indexate ISI, iar alte 9 lucrări urmează să fie indexate ISI (apărute în volumele unor conferinţe cotate ISI).

Pornind de la necesitatea identificării unei metrici adecvate care să fie utilizată în analiza similarităţii limbilor naturale, în [16] am propus o metrică numită rank distance [17]. În capitolul 3 sunt prezentate principalele proprietăţi matematice şi computaţionale ale acesteia (marea majoritate fără demonstraţii incluse). Sunt studiate problemele de maxim şi medie pe clasamente şi cuvinte (cu lungime egală sau nu, cu aceeaşi compoziţie sau nu) [51] [18], proprietăţile de colinearitate ale acesteia [52], densitatea unui segment [48], sunt propuşi 2 algoritmi liniari ce permit calcului acesteia [49].

În strânsă legătură cu capitolul 3, în capitolul 4 facem trecerea spre două probleme importante ale informaticii teoretice si ale aplicaţiilor acestora. Este vorba de analiza şirului median si a şirului central. Ambele probleme au fost iniţiate pornind de la distanţa Hamming, şi, după ce s-a demonstrat intractabilitatea polinomială a lor (atât pentru Hamming cât şi pentru edit distance) [54], a devenit o problemă deschisă existenţa unei metrici care să conducă la o rezolvare eficientă (polinomială) măcar a şirului median. În capitolul 4 arătăm că şirul median via rank distance are o tratare polinomială [35], în schimb şirul central rămâne în continuare intractabil [42]. Transformăm ambele probleme într-o problemă de agregare [16] şi categorizare multicriterială [44], facem joncţiunea cu celebra teoremă de imposibilittae a agregării a lui Arrow, studiem criteriile de rationalitate propuse de Arrow şi extinse de Paun [16] [44], propunem o metodă nouă de clustering bazată pe rank distance şi cele două probleme de agregare [27] [31] [34], şi prezentăm două aplicaţii concrete ale tehnicii de categorizare propuse: identificarea cifrelor scrise de mână [44] [32] şi categorizarea automată a textelor [47]. Tot aici prezentăm principalele rezultate obţinute în bioinformatică, şi mai precis in analiza similarităţii lanţurilor ADN. Am căutat să vedem dacă proprietăc tile computaţionale şi matematice bune ale rank distance o califică pe aceasta în analiza similarităţii lanţurilor ADN [49] [23] [50]. Am căutat să vedem dacă o distanţa ordinală, care contorizează mai degrabă distanţa dintre off-seturile a doŭa elemente, poate funcţiona în această extrem de competitivă zonă. Rezultatele bune obţinute de noi au fost reliefate de aplicarea rank distance pe ADN-ul mitocondrial al mamiferelor, arborele filogenetic final obţinut fiind comparabil cu cel standard raportat [49] [34]. Am investigat acelaşi lucru prin aplicarea

şirului central via rank distance, şi prin aplicarea unui algoritm de aliniere bazat pe LRD (local rank distance)[30], o metrică derivată din rank distance.

Începând cu capitolul 5 sunt prezentate principalele rezultate obţinute în aria lingvisticii computaţionale, grupate pe teme. Capitolul 5 este dedicat unei unităţi lingvistice mult mai puţin studiată decât altele, şi anume silaba. Continuând preocupările din timpul doctoratului [15], am propus o abordare formală nouă a silabificării şi implicit a silabei, bazându-ne pe gramaticile de inserţie [11] [19]. Prezentăm aici o formalizare nouă, şi anume un model paralel de silabificare, împreună cu principalele rezultate obţinute privind clasificarea noii clase de limbaje pe care am propus-o [20]. Am introdus de asemena o nouă clasă de gramatici contextuale (numită *syllabic grammars*) [25] şi am făcut joncţiunea cu automatele *go through* [24], introduse de Gramatovici [53]. Aspectele cantitative au fost investigate prin raportarea la întregul dicţionar al limbii române [12], şi am prezentat o serie de legi cantitative şi plierea acestora pe limba română [13].

Capitolul 6 este dedicat rezultatelor obţinute în analiza morfologică a limbii române. Româna este o limbă cu o bogată morfologie, şi ultima perioadă a arătat un avânt puternic al lucrărilor dedicate unor limbi aflate în această situaţie. În acest capitol prezentăm rezultatele noastre dedicate unor probleme importante ale morfologiei limbii române. O problemă de certă importanţă este detectarea alternanţelor verbului românesc şi conjugarea acestuia la indicativ prezent. Înr-o serie de articole [26] [37] am tratat această problemă, şi am reuşit să prezicem alternanţele verbului cu o precizie de 95%. Am propus de asemena şi un mecanism formal de rezolvare a alternanţelor [26], pornind de la conceptul de literă variabilă introdus de Moisil în 1965. O altă problemă prezentatăici este cea a analizei pluralului substantivelor şi detectarea genului acestora [36] [39]. Problema principală ridicată este detectarea automată a neutrului. Folosind tehnici de învă ţare automată, investigăm cu bune rezultate problema silabificării ortografice a cuvintelor din limba română[40]. O alta problemă a limbii române este predicţia accentului [3] [2]. Marea majoritatea a linvistilor cad de acord că accentul in română este nepredictibil (cel mai recent [56]). Propunem un mecanism de învăţare a accentului, cu o acurateţe de 98%.

În capitolul 7 ne dedicăm investigării problemelor de similaritate şi de influenţă şi analiză a cuvintelor înrudite referitorare la limba română. Pornind de la problema similarităţii silabice a limbii române (problemă prezentată în teza de doctorat

[10] şi continuată apoi în [17] [22], [21]), ne-am propus să extindem investigaţia în direcţia inteligibilităţii şi similarităţii limbii române tratate din perspectiva posibilităţii de înţelegere a unui text [8] [7]. Am considerat că un punct bun de pornire este detectarea perechilor de cuvinte cognates dintre română şi alte limbi, şi investigarea comportamentului acestora [4] [6]. Această problemă este o problemă de certă actualitate şi este intens studiată, având o multitudine de aplicaţii în traducerea automată, achiziţia unei limbi secundare, achiziţia limbajului, evoluţia limbii, etc. Am propus un algoritm automat de extragere a perechilor cognates dintre română şi orice altă limbă cu resurse electronice disponibile [4], am determinat perechile cognates (intr-o manieră exhaustiva) dintre română şi limbile romanice + turcă, am investigat capacitatea de a decide dacă două cuvinte sunt sau nu cognates prin utilizarea unor algoritmi de aliniere [9] şi folosirea unor metrici adecvate [6], am căutat să vedem dacă putem distinge intre cognates şi ethimons, am identificat evoluţia şi urmele lăsate de cuvintele intrate ţn limbă şi evoluţia acestora [5]. Am propus o nouă metodă de cuantificare a similarităţii ortografice a limbii romane cu celelalte limbi, şi am prezentat principalele rezultate obţinute, care aduc o nouă perspectivă asupra similarităţii limbii române vis-a-vis de limbile indo-europene studiate [7].

Capitolul 8 este dedicat problemelor de identificare de autor. Aici propunem o abordare nouă, prin aplicarea rank distance pe clasamentele cuvintelor funcţionale corespunzătoare operelor investigate [58]. Testăm atât pe romane englezeşti cât şi pe romane scrise în limba română. Am folosit printre primii metodele kernel [57] în problemele legate de identificarea autorului unor texte cu autor controversat. Penru evaluare, am folosit atât texte standard (ca de exemplu *Federalist Papers*, [43]), cât şi texte controversate din literatura română [59] [46]. Am investigat de asemnea unitatea stilistică a Episolelor Apostolului Pavel [14]. Am investigat în general distantele ordinale şi am motivat pentru alegerea aacestora [45]. Am investigat de asemnea problemele de traducere apărute (în special cele legate de discriminarea dintre operele scrise de un autor într-o limbă anume şi cele traduse în acea limbă) [55] [41] şi problemele legate de detectarea autorior care vor să se substiutie unui autor real (deception detection)[38]. Toate aceste probleme sunt însoţite de aplicaţii concrete.

Ultimul capitol este dedicat prezentării evoluţiilor academice şi ştiintifice ulterioare ale autorului, prin prezentarea unor teme de cercetare aflate în imediata atenţi

a noastră, precum si modalităţilor pe care în acest moment le identificăm în scopul atingerii acestor obiective.

Contents

1 Summary **3**

2 Rezumat **9**

Bibliography

[1] Ana-Maria Barbu. Romanian lexical databases: Inflected and syllabic forms dictionaries. In *Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

[2] Ioana Chitoran, Alina Maria Ciobanu, Liviu P. Dinu, and Vlad Niculae. Using a machine learning model to assess the complexity of stress systems. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 331–336, 2014.

[3] Alina Maria Ciobanu, Anca Dinu, and Liviu P. Dinu. Predicting romanian stress assignment. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 64–68, 2014.

[4] Alina Maria Ciobanu and Liviu P. Dinu. A dictionary-based approach for evaluating orthographic methods in cognates identification. In *Recent Advances in Natural Language Processing, RANLP 2013, 9-11 September, 2013, Hissar, Bulgaria*, pages 141–147, 2013.

[5] Alina Maria Ciobanu and Liviu P. Dinu. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of*

the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers, pages 99–105, 2014.

[6] Alina Maria Ciobanu and Liviu P. Dinu. Building a dataset of multilingual cognates for the romanian lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 1038–1043, 2014.

[7] Alina Maria Ciobanu and Liviu P. Dinu. An etymological approach to cross-language orthographic similarity.application on romanian. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2014), October 25–29, 2014, Doha, Qatar (accepted)*, 2014.

[8] Alina Maria Ciobanu and Liviu P. Dinu. On the romance languages mutual intelligibility. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 3313–3318, 2014.

[9] Antonella Delmestri and Liviu P Dinu. An assessment of string similarity methods for cognate identification. In *Methods and Applications of Quantitative Linguistics. Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO 2012), Belgrad, Serbia, april 26-29 2012*, pages 152–163, 2012.

[10] Anca Dinu and Liviu P. Dinu. On the syllabic similarities of romance languages. In *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005, Proceedings*, pages 785–788, 2005.

[11] Anca Dinu and Liviu P. Dinu. A parallel approach to syllabification. In *Computational Linguistics and Intelligent Text Processing, 6th International Con-*

ference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005, Proceedings, pages 83–87, 2005.

[12] Anca Dinu and Liviu P. Dinu. On the data base of romanian syllables and some of its quantitative and cryptographic aspects. In *Proceedings of The 5th Internationa Conference onl Language Resources and Evaluation (LREC), Genova, Italy, may 2006*, pages 1795–1798, 2006.

[13] Anca Dinu and Liviu P. Dinu. On the behavior of romanian syllables related to minimum effort laws. In *Proceedings of the Workshop on Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages, 14-16 September, 2009, Borovets, Bulgaria (co-located with RANLP)*, pages 9–13. Association for Computational Linguistics, 2009.

[14] Anca Dinu, Liviu P. Dinu, Alina Resceanu, and Ionut Resceanu. Some issues on the authorship identification in the apostles' epistles. In *Language Resources and Evaluation for Religious Texts, LREC 2012 workshop, Istambul, Turkey*, pages 18–25, 2012.

[15] Liviu P. Dinu. An approach to syllables via some extensions of marcus contextual grammars. *Grammars*, 6(1):1–12, 2003.

[16] Liviu P Dinu. On the classification and aggregation of hierarchies with ifferent constitutive elements. *Fundamenta Informaticae*, 55(1):39–50, 2003.

[17] Liviu P. Dinu. Rank distance with applications in similarity of natural languages. *Fundamenta Informaticae*, 64(1):135–149, 2005.

[18] Liviu P. Dinu. On the behavior of certain metrics on the permutation group. *Proceedings of the Romanian Academy, series A*, 7(2):105–110, 2006.

[19] Liviu P. Dinu. On the quantitative and formal aspects of the romanian syllables. *Revue Roumaine de Linguistique*, pages 3–4, 2006.

[20] Liviu P. Dinu. On insertion grammars with maximum parallel derivation. *Fundamenta Informaticae*, 93(4):357–369, 2009.

[21] Liviu P. Dinu and Denis Enachescu. Languages similarity: Measuring and testing. In *Recent Advances in Stochastic Modelling and Data Analysis*, pages 51–520.

[22] Liviu P. Dinu and Denis Enachescu. On clustering romance languages. In *Recent Advances in Stochastic Modelling and Data Analysis*, pages 521–529.

[23] Liviu P. Dinu and Florin Ghetu. Circular rank distance: A new approach for genomic applications. In *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on*, pages 397–401. IEEE, 2011.

[24] Liviu P. Dinu, Radu Gramatovici, and Florin Manea. On the syllabification of words via go-through automata. In *Proceedings of the 1st International Conference on Language and Automata Theory and Applications (LATA 2007)*, pages 223–236, 2007.

[25] Liviu P. Dinu, Radu Gramatovici, and Florin Manea. Syllabic languages and go-through automata. *Fundamenta Informaticae*, 131(3):409–424, 2014.

[26] Liviu P. Dinu, Emil Ionescu, Vlad Niculae, and Octavia-Maria Sulea. Can alternations be learned? A machine learning approach to romanian verb conjugation. In *Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria*, pages 539–544, 2011.

[27] Liviu P. Dinu and R-T Ionescu. Clustering methods based on closest string via rank distance. In *14th International Symposium on Symbolic and Nu-*

meric Algorithms for Scientific Computing, SYNASC 2012, Timisoara, Romania, September 26-29, 2012, pages 207–213. IEEE, 2012.

[28] Liviu P. Dinu and Radu Ionescu. A genetic approximation of closest string via rank distance. In *Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2011 13th International Symposium on*, pages 207–214. IEEE, 2011.

[29] Liviu P. Dinu and Radu Ionescu. An efficient rank based approach for closest string and closest substring. *PloS one*, 7(6):e37576, 2012.

[30] Liviu P Dinu, Radu Ionescu, and Alexandru Tomescu. A rank-based sequence aligner with applications in phylogenetic analysis. *PloS One*, 9(8):e104006, 2014.

[31] Liviu P. Dinu and Radu-Tudor Ionescu. Clustering based on rank distance with applications on DNA. In *Neural Information Processing - 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part V*, pages 722–729, 2012.

[32] Liviu P. Dinu and Radu-Tudor Ionescu. A rank-based approach of cosine similarity with applications in automatic classification. In *14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2012, Timisoara, Romania, September 26-29, 2012*, pages 260–264. IEEE, 2012.

[33] Liviu P. Dinu and Radu-Tudor Ionescu. An efficient algorithm for rank distance consensus. In *AI*IA 2013: Advances in Artificial Intelligence - XIIIth International Conference of the Italian Association for Artificial Intelligence, Turin, Italy, December 4-6, 2013. Proceedings*, pages 505–516, 2013.

[34] Liviu P. Dinu and Radu Tudor Ionescu. Clustering based on median and closest string via rank distance with applications on dna. *Neural Computing and Applications*, 24(1):77–84, 2014.

[35] Liviu P. Dinu and Florin Manea. An efficient approach for the rank aggregation problem. *Theoretical Computer Science*, 359(1):455–461, 2006.

[36] Liviu P. Dinu, Vlad Niculae, and Octavia-Maria Sulea. Dealing with the grey sheep of the romanian gender system, the neuter. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Demonstration Papers, 8-15 December 2012, Mumbai, India*, pages 119–124, 2012.

[37] Liviu P. Dinu, Vlad Niculae, and Octavia-Maria Sulea. Learning how to conjugate the romanian verb. rules for regular and partially irregular verbs. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, pages 524–528, 2012.

[38] Liviu P. Dinu, Vlad Niculae, and Octavia-Maria Şulea. Pastiche detection based on stopword rankings: exposing impersonators of a romanian writer. In *Proceedings of the Workshop on Computational Approaches to Deception Detection, Avignon, France, April 23-27, 2012 (co-located with EACL)*, pages 72–77. Association for Computational Linguistics, 2012.

[39] Liviu P. Dinu, Vlad Niculae, and Octavia-Maria Sulea. The romanian neuter examined through A two-gender n-gram classification system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 907–910, 2012.

[40] Liviu P. Dinu, Vlad Niculae, and Octavia-Maria Sulea. Romanian syllabication using machine learning. In *Text, Speech, and Dialogue - 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings*, pages 450–456, 2013.

[41] Liviu P. Dinu and Sergiu Nisioi. Authorial studies using ranked lexical features. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Demonstration Papers, 8-15 December 2012, Mumbai, India*, pages 125–130, 2012.

[42] Liviu P. Dinu and Alexandru Popa. On the closest string via rank distance. In *Combinatorial Pattern Matching - 23rd Annual Symposium, CPM 2012, Helsinki, Finland, July 3-5, 2012. Proceedings*, pages 413–426, 2012.

[43] Liviu P. Dinu and Mariu. Popescu. Language independent kernel methods for classifying texts with disputed paternity. In *Proceedings of the XIII International Conference Applied Stochastic Models and Data Analysis (ASMDA 2009), june 30-july 3, Vilnius, Lithuania*, pages 66–70, 2009.

[44] Liviu P. Dinu and Marius Popescu. A multi-criteria decision method based on rank distance. *Fundamenta Informaticae*, 86(1):79–91, 2008.

[45] Liviu P. Dinu and Marius Popescu. Ordinal measures in authorship identification. *Proceedings 3rd PAN workshop. Uncovering plagiarism, authorship and social software misuse.(PAN'09), San Sebastian, september*, pages 62–66, 2009.

[46] Liviu P. Dinu, Marius Popescu, and Anca Dinu. Authorship identification of romanian texts with controversial paternity. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 2008.

[47] Liviu P. Dinu and Andrei A. Rusu. Rank distance aggregation as a fixed classifier combining rule for text categorization. In *Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010. Proceedings*, pages 638–647, 2010.

[48] Liviu P Dinu and Andrea Sgarro. Estimating similarities in dna strings using the efficacious rank distance approach. In *Systems and Computational Biology - Bioinformatics and Computational Modeling, Prof. Ning-Sun Yang (Ed.)*, pages 121–138.

[49] Liviu P. Dinu and Andrea Sgarro. A low-complexity distance for dna strings. *Fundamenta Informaticae*, 73(3):361–372, 2006.

[50] Liviu P. Dinu and Andrea Sgarro. Rank distance: a soft tool for comparison of dna strings. In *Proc. IPMU 2006 (11th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems), Les Cordeliers, Paris, France, July*, pages 2791–2799, 2006.

[51] Liviu P. Dinu and Andrea Sgarro. Maximal rank distance for binary sequences. *Mathematica Pannonica*, 19:125–131, 2008.

[52] Liviu P. Dinu and Ioan Tomescu. From rankings' collinearity to counting sdr's via chromatic list expression. *International Journal of Computer Mathematics*, 86(9):1483–1489, 2009.

[53] Radu Gramatovici and Florin Manea. Parsing local internal contextual languages with context-free choice. *Fundam. Inform.*, 64(1-4):171–183, 2005.

[54] F. Nicolas and E. Rivals. Complexities of the centre and median string problems. In *CPM*, pages 315–327, 2003.

[55] Sergiu Nisioi and Liviu P. Dinu. A clustering approach for translationese identification. In *Recent Advances in Natural Language Processing, RANLP 2013, 9-11 September, 2013, Hissar, Bulgaria*, pages 532–538, 2013.

[56] Gabriela Pană Dindelegan. *The Grammar of Romanian*. Oxford University Press, 2013.

[57] Marius Popescu and Liviu P. Dinu. Kernel methods and string kernels for authorship identification: The federalist papers case. In *Recent Advances in Natural Language Processing, RANLP september 27-29, 2007, Borovets, Bulgaria*, pages 484–487, 2007.

[58] Marius Popescu and Liviu P. Dinu. Rank distance as a stylistic similarity. In *COLING 2008, 22nd International Conference on Computational Linguistics, Posters Proceedings, 18-22 August 2008, Manchester, UK*, pages 91–94, 2008.

[59] Marius Popescu and Liviu P. Dinu. Comparing statistical similarity measures for stylistic multivariate analysis. In *Recent Advances in Natural Language Processing, RANLP 2009, 14-16 September, 2009, Borovets, Bulgaria*, pages 349–354, 2009.