Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI'2016, July 25-29, 2016, Chisinau, Republic of Moldova

Diachronic Analysis Using a Statistical Model

Daniela Gîfu

Abstract: This paper describes a statistical methodology for a diachronic study on a large corpus (a collection of publications, written from the second decade of the 19th century in two countries, Romania and Republic of Moldova, known as Bessarabia). The aim of this work is to analyse the lexical evolution of words in all four regions using a machine learning approach to identify the patterns that govern language changes. Basically, it was developed a mechanism for automatic correlation of different forms of the same words in order to train a statistical model on a list of known word-to-word correlations between lexicons.

Keywords: statistical model, diachronic study, corpora, lexical evolution, writing press.

1 Introduction

This study is based on diachronic exploration of Romanian and Bessarabian texts in order to implement a methodology for detecting automatically the language variation from the second decade of the 19th century to nowadays using a probability distribution estimation model, called MaxEnt (*Maximum Entropy*). Actually, this work is a continuation of a previous one (Gîfu & Simionescu, 2016), here a priori division of the temporal axis was excluded.

The present research is based on the question *how Romanian language has evolved at a particular period in different historical places?*

The language variation is often narrowed to consideration of change in one aspect of language: lexis, morphology, phonology, syntax, and semantics. A language variation in fact occurred also at the levels of discourse and pragmatics (Gass et al., 1989). The diachronically contrastive studies of the Romance languages (e.g. Romanian, Spanish,

French, Italian, and Portuguese) expose the presence of many similarities. (Densuianu, 1902). The Romanian language with approximately 24 million speakers has an important particularity. It is still spoken in Eastern Europe, with official status in Romania, Moldova, and parts of Serbia and Greece. Moreover, Romanian is recognized in Hungary (historical reasons) as a minority language and spoken in Ukraine, Albania, and Macedonia. (Miller-Broomfield, 2015).

For instance: the noun *prieten* → (EN. *friend*) in Romanian to five Romance languages has the same origin, Latin (*amicus*): Romanian – *amic* is synonym with prieten[1], Spanish – *amigo*, Catalan – *amic* is a dialect of Spanish, French – *ami*, Italian – *amico*, *and* Portuguese – *amigo*.

The paper is structured as follows: Section 2 presents briefly relevant literature that reveals a large interest for diachronic studies. Section 3 describes a methodology for research of language using a statistical model, on a list of known word-to-word correlations between lexicons, Section 4 presents the statistics results using machine learning model. Finally, the survey conclusions and future work are given in Section 5.

2 Related Work

Until now, the Romanian diachronic phenomenon was analysed using various methods. One of them relies on reconstructing a diachronic morphology for Romanian (Cristea et al., 2012), based on the digital version of the Romanian Language Thesaurus Dictionary (eDTLR) (Cristea et al, 2007). The authors detected the old form words occurring in the citations. For the Bessarabia, a group at the Institute of Mathematics and Computer Science, Academy of Sciences of Moldova proposed a technology based on transliteration and parallel texts alignment for creation of linguistic lexicon for Bessarabian corpus in Cyrillic script between 1967–1989 starting from actual Romanian lexicon. (Boian et al., 2014).

[1] In this study we used Romanian WordNet (http://dcl.bas.bg/bulnet/) the largest lexical ontology available today with a large collection of synsets. A synset is the wordnet's basic unit, being a set of synonyms which defines a specific meaning, common to the members of the synset. (Tufiș & Cristea, 2002; Tufiș et al., 2004, Ștefănescu, 2015).

Also, a study of language as an evolutionary phenomenon is included in (Mihalcea and Năstase, 2012). Their task was word epoch disambiguation, using text classification according to a specific epoch, knowing that the language is a dynamic phenomenon over time, being dependent on context. Actually, we found useful to statistical tests presented for epoch detection in (Popescu & Strapparava, 2013/2014), also, called temporal dynamics in (Wang & McCallum, 2006; Wang et al., 2008; Gerrish and Blei, 2010). Moreover, the diachronic text evaluation requires the automatic system in order to identify the epoch when the newspaper article was written (Gîfu, 2016; Popescu and Strapparava, 2015).

In order to evaluate the writing styles more researchers considered various indices: text features (Dascălu & Gîfu, 2015), textual formality (Eggins and Martin, 1997), and textual styles (Biber, 1987).

The development and use of software for natural language processing (NLP) highlight the defining aspects of two journalistic languages (Romania and Bessarabia) that have many similarities on the time axis that we have chosen. (Gîfu, 2014/2015). Furthermore, the diachronic study continues with exploring the patterns that govern the lexical differences between two lexicons, based on machine learning approach (Gîfu & Simionescu, 2016). This paper considers the study of the evolution of Romanian language focused on the lexical similarity based on statistical model.

3 Work Methodology

This section describes a language study based on a historical corpus used for investigating the evolution of words over time. This work is based on the Maximum Entropy (MaxEnt) text classifier being commonly used in Natural Language Processing (NLP) tasks, introduced first by Berger [Berger, et al, 1996] and Della Pietra [Della Pietra et al., 1997] in statistical estimation and pattern recognition. Noteworthy is that MaxEnt classifier has great results when the training corpus is limited, as in this case. Actually, the differences between the „source" lexicons (Moldavian, Transylvanian, Wallachian, and Bessarabian) and the „destination"

lexicon (DEX-online[2]) from the perspective of transformation patterns, were analyzed using this model. All the substring replacement operations (referred as "REP") are classified and extracted from the known correlations list, based on the character-level context in which they are applied in the source word. Based on these REPs, a set of fictive/candidate words are generated, each having a trust score attached. If a candidate word is found in the destination lexicon, the two words are marked as a corresponding pair.

Basically, all unknown words are extracted in order to find them the current correspondent. By applying these REPs operations on the first word of the pair (the old word), the present word[3] is obtained.

For instance the vowel [u] at the end of words that only had a phonetic significance.

Example: `tot`**`u`** `= totul` (Transylvania, 1881)

The consonant [s] (deaf) intervocalic is vocalized; thus it became the consonant [z].

For instance: `mu`**`s`**`ician = musician` (Wallachia, 1919)

The vowel [i] becomes in some situations [î].

An example: **`i`**`n =` **`î`**`n` (Moldavia, 1869)

The inflexion of words is often different: [ei] is transformed in [ii].

An example: `reclădir`**`ei`** `= reclădir`**`ii`** (Bessarabia, 1918)

Of course, these are the simplest situations, when we talk just one REP operation. But, in the present corpus, we have complex cases, when several REPs operations have intervened. To increase the accuracy of statistical data in identifying automatically the correlations, we decided to focus just three operations REPs for each words pair (old - new).

[2] www.dexonline.ro

[3] There was used the morphologic dictionary there (e.g. DEX on-line – www.dexonline.ro)

To illustrate this option, below is one example for each geographical area that includes 3 REPs operations:

Transylvania: a noun `serbâtoria` (En: *celebration*) in the direct case:

`serbâtoria = **sărbătoarea:** î->ă ria->area e->ă`

where `serbâtoria` is the "source" word, and `sărbătoarea` is the "destination" word.

Wallachia: a noun `esposiţiunea` (En: *exhibition*) in the direct case:

`esposiţiunea – **expoziţia:** s->z unea->a s->x`

Moldavia: a predicative verb `măngăemu` (En: *cosset*) in the indicative moode:

`măngăemu = **mîngîiem:** ă->îi ă->î u->`

Bessarabia: a noun `iantămplari` (En: *events*) in the direct case:

`iantămplari = **întîmplări:** ia->î ă->î a->ă`

As it was mentioned, this model is trained on a list of known word to word correlations between two lexicons (*source* and *destination*). For this study the size of the training data was not too big (40% from the current corpus), but we tried to cover a wide variety of lexical evolution phenomena. Basically, the trained model is used for predicting REPs which can be applied on a previously unknown word from the source lexicon.

3.1 Corpus

The corpus includes articles (over 3 million lexical tokens), chronologically ordered, from the most important Romanian and Bessarabian publications since 1917 until nowadays (Table 1). Moreover, the corpus was developed and structured in four independent collections of publications corresponding to Moldavia (*Albina românească*; *Convorbiri literare*; *Curierul*. *Foaia intereselor generale*; *Constitutionalul*; *Moldova Socialistă*; *Scânteia*; *Noutatea*; *Deşteptarea*; *Bună ziua, Iaşi*; *Ziarul de Vrancea*; *Monitorul de Vaslui*; *Evenimentul regional al Moldovei*; *Imparţial*), Wallachia (*Curier românesc*; *Buletin. Gazeta oficială*; *România*; *Curierul românesc*; *Pressa*, *România liberă*; *Românulu*; *Timpul*; *Literatorul*; *Albina*; *Deşteptarea*. *Foaie pentru*

popor; *Adeverul*; *Curierul artelor*; *Dimineața*; *Universul*; *Viitorul*; *Curentul*; *Universul literar*; *Adevărul*; *Adevărul literar și artistic*; *Scânteia*; *Romania literară*; *Dimineața copiilor*; *Evenimentul zilei*; *Gândul*; *Ziua*; *Ziua news*; *Ziua veche*), Transylvania (*Organulu Luminarei*; *Gazeta de Transilvania*; *Gazeta Transilvaniei*; *Telegrafulu Românu / Telegraful român*; *Foaia pentru Minte Anima și Literatură*; *Transilvania*; *Federațiunea*; *Gura Satului*; *Albina*; *Telegraful Românu*; *Familia*; *Aradu*; *Patria*; *Chemarea tinerimei române*; *Dreptatea*; *Aradul*; *Curierul creștin*; *Vatra românească*; *Echinox*; *Adevărul de Cluj*; *Făclia*; *Monitorul de Cluj*; *Bihoreanul*), and Bessarabia (*Basarabia reînoită*; *Curierul*; *Candela*; *Deșteptarea*; *Viața economică din Bălți*; *Solidaritatea*; *Ehos*; *Buletinul Arhiepiscopiei Chișinăului*; *Cuvânt moldovenesc*; *Basarabia*; *România nouă*; *Sfatul țării*; *Democratul Basarabiei*; *Glasul Basarabiei*; *Luminătorul*; *Dreptatea*; *Basarabia Chișinăului*; *Literatura și artă*; *Moldova Socialistă*; *Jurnal*; *Contrafort*; *Jurnal de Chișinău*; *Moldova suverană*; *Ziarul de gardă*) that was a part of old Moldavia until 1812, and then between 1918-1941, becoming an independent state since 1991.

3.2 Preprocessing chain

The automatic preprocessing chain applied on this corpus consists of the following sequences: segmentation, tokenization, part-of-speech tagging, lemmatization, using the Romanian POS-tagger (Simionescu, 2011). The final XML includes an extra markup attribute, `NotInDict`. Each NotInDict is a token which is not recognized by DEX-online.

Below is a segmentation annotation in XML standoff format from Albina (Transylvania), 1884:

Trăim în nisce...

where *nisce* is an old form (marked with NotInDict) of the indefinite article, *nișce.*

```
  <?xml          version="1.0"          encoding="UTF-8"
standalone="no"?>
<POS_Output>
<S id="4" offset="186">
```

```
    <W   EXTRA="tranzitiv"   LEMMA="trăi"   MSD="Vmip1p"
Mood="indicative"       Number="plural"       POS="VERB"
Person="first"    Tense="present"    Type="predicative"
id="4.1" offset="0">Trăim</W>
    <W  LEMMA="în"  MSD="Sp"  POS="ADPOSITION"  id="4.2"
offset="6">în</W>
    <W          Case="direct"          Definiteness="no"
EXTRA="NotInDict"    Gender="feminine"    LEMMA="nisce"
MSD="Ncfsrn"         Number="singular"         POS="NOUN"
Type="common" id="4.3" offset="9">nisce</W>
…
</S>
```

The global situation is related in Table 1 and represented graphically in Figure 1.

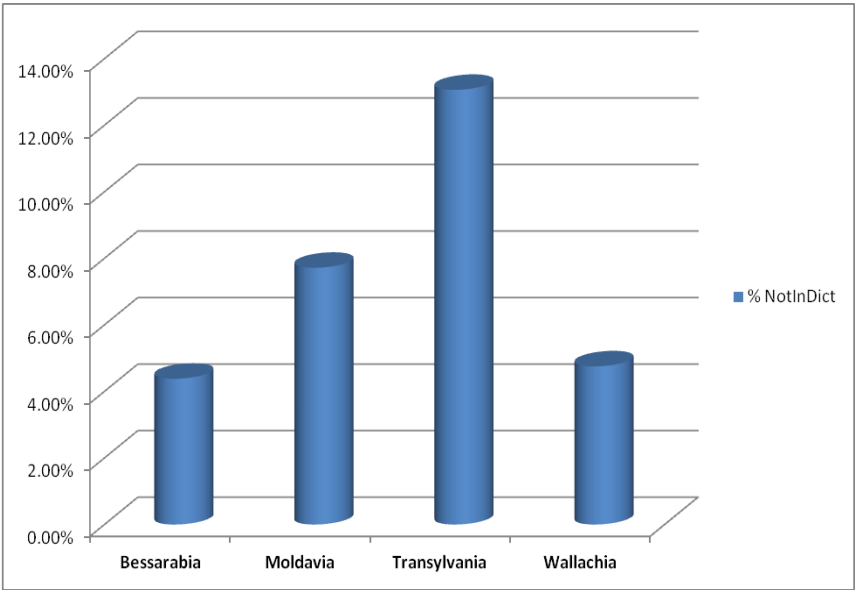

Figure 1. Percentage of NotInDict Words - 1817-2015

Table 1. General corpus statistics

Region	Time	Total considered words[4]	Total unknown Words	Total unique unknown words	%(total unknown words/total words)
Moldavia	1829-2015	65901	5085	2979	7.72
Wallachia	1829-2015	137261	6525	4105	4.75
Transylvania	1837-2015	160923	21023	8518	13.06
Bessarabia	1817-2015	107324	4703	2891	4.38

Although the four corpora are slightly disproportionate as number, the Transylvania case is different from the other three. The language in Transylvania is marked by historical waves: 1849-1860, the official language is German, including in administration; 1860-1866, the Romanian language returned as the official language, following the Romanian claims; 1867-1914, the Dual Monarchy is installed, which grants visible linguistic concessions by nationalities law.

[4] From the total tokens the punctuation, the numbers and the words with less than two characters were removed.

4 Statistics and interpretation

In this section the statistical results for the 4 collections of journalistic texts that correspond to Moldova, Transylvania, Wallachia and Bessarabia will be highlighted. There was used a mechanism of automatic correlation of unknown words with the new ones, presented above.

In Table 2, we consider the most common REPs (12) for the present corpus for a common period (1840-2015).

Table 2. The percentage of REPs

REP	1840 - 2015 Wallachia	1840 - 2015 Transylvania	1840 - 2015 Moldavia	1840 - 2015 Bessarabia
u ->	2.61%	**18.26%**	**11.55%**	5.79%
s -> z	**12.89%**	5.19%	9.06%	4.06%
e -> ă	5.91%	3.83%	6.22%	1.66%
ei -> ii	6.27%	2.56%	3.58%	**7.86%**
e -> i	2.61%	2.10%	3.09%	2.07%
i -> î	0.75%	1.99%	5.82%	1.82%
i -> e	3.22%	1.23%	3.04%	3.23%
ĭ -> i	2.04%	1.77%	1.14%	1.99%
a -> ă	1.22%	1.97%	1.19%	1.16%
s -> x	3.58%	1.48%	0.70%	0.08%
a ->	1.97%	1.30%	2.39%	2.73%
e -> î	1.72%	1.38%	2.34%	0.91%

It can be seen that in Transylvania and Moldavia similarities appear in writing rules, if we look at the hierarchy of these REPs (first 5). There was a special situation, Bessarabia between 1945-1989, a period when nothing was published anymore in Latin script (except the war years). This period has not been considered in this study. The vales from the Table 2 are represented in Figure 2.

All these REPs will become an important rules-based system very useful to develop a diachronic POS tagger for Romanian, another future work direction. We believe this MaxEnt model could be used to add

enhanced support for unknown words in order to develop the POS-tagger[5] for contemporary Romanian used in this paper (a free online service). This collection of publications can be considered a start for developing a Gold Corpus required for training such a diachronic POS tagging model.

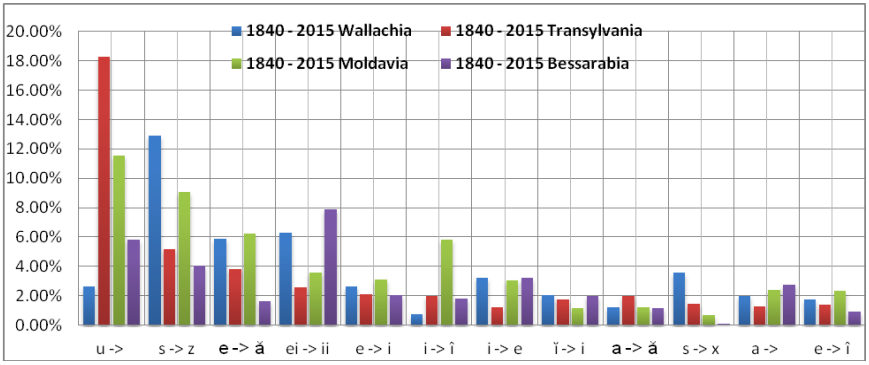

Figure 2. The most frequent REPs

The results from Table 3 are very promising.

Table 3. Known words vs. identifiable words comparison

Region/ statistical parameter	Transylvania	Wallachia	Moldavia	Bessarabia
Known words	85.96%	94.98%	91.82%	95.91%
Identifiable words	97.40%	98.11%	96.55%	97.80%
MaxEnt model - Precision	**81,39**	**82,92**	**77,05**	**86,81**

The "Identifiable words" represent the percent of words from the texts with known words or which can be correlated automatically with a known form. As we can see, for all geographical areas this indicator is over 95%. In this case, over 96% of the words can be recovered, but only

[5] http://nlptools.info.uaic.ro/WebPosRo/

76% of these are automatically correlated with a precision of 82%. Regarding the indicator "Known Words", only for Transylvania the result is smaller, because the corpus was bigger.

5 Conclusions and discussions

The methodology presented is language independent and it offers a basis for future large-scale studies, having a large impact on reducing the amount of human effort required by linguistic analysis of language variants.

This work presents a language variation over time in order to compare the journalistic language changes in four regions, Moldavia, Wallachia and Transylvania (Romania) and Bessarabia (a historical part of Romania). This survey investigates the problem of journalistic language similarity between cognate languages. The statistical results show the fact that there exists a high level of similarity between the lexicons of those four historical Romanian regions analyzed, at least in the newspapers.

In the future, an interesting experiment could be focused on the transliteration differences from Cyrillic to Latin both in Romania and Bessarabia until 1862, when in Romania the texts were published in both alphabets. Moreover, given that in the period 1944-1989 (excluding the war years) in Bessarabia the writing in Latin alphabet was prohibited, the process of collecting and transliterating publications of those times - with the support of the Academy of Sciences of Chisinau – should continue.

Acknowledgments. I would like to thank my colleagues, Radu Simionescu and Augusto Perez from the Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi for all support to finish this work.

References

[1] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. (1996). *A Maximum Entropy Approach to Natural Language Processing*. In: Computational Linguistics, 22(1), pp. 39-71.

[2] D. Biber. *A textual comparison of British and American Writing*. American Speech, (62) (1987), pp. 99–119.

[3] E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, L. Malahov. *Cultural and Historical Heritage Digitization, Recognition and Conservation*. In: Akademos: Revista de Ştiinţă, Inovare, Cultură şi Artă, nr. 1 (32) (2014), pp. 61-68.

[4] D. Cristea, M. Răschip, C. Forăscu, G. Haja, C. Florescu, B. Aldea, E. Dănilă, *The Digital Form of the Thesaurus Dictionary of the Romanian Language*. In: Proceedings of SpeD 2007 Speech Technology and Human - Computer Dialogue, Iași, May 10-12 (2007).

[5] D. Cristea, R. Simionescu, G. Haja. *Reconstructing the Diachronic Morphology of Romanian from Dictionary Citations*. In: Proceedings of LREC-2012, Istanbul, 21-25 May (2012).

[6] M. Dascălu, and D. Gîfu. *Evaluating the Complexity of Online Romanian Press*. In: Proceedings of the 11th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2015, 26-27 Nov. 2015, Iași, Romania, "Alexandru Ioan Cuza" University Publishing House, Iași, Romania (2015), pp. 149-162.

[7] S. A. Della Pietra, V. J. Della Pietra, and J. Lafferty, J. (1997). *Inducing features of random fields*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(4):380–393.

[8] A. Delmestri, and N. Cristianini, N. *String Similarity Measures and PAM-like Matrices for Cognate Identification*. Bucharest Working Papers in Linguistics, 12(2) (2010), pp. 71-82.

[9] O. Densusianu. *Filologia Romanică în universitatea noastră*, Bucuresci, J. V. Socecu Editeur (1902), p. 23.

[10] Ș. D. Dumitrescu. *Rowordnetlib – the First API for the Romanian WordNet*. In: Proceedings of the Romanian Academy, Series A, vol. 16: 1 (2015), pp. 87-94.

[11] S. Eggins, J. R. Martin. *Genres and Register of Discourse*. In: Dijk, T.A.v. (ed.) Discourse as Structure and Process (Discourse Studies – A Multidisciplinary Introduction), Vol. 1, Sage Publications, London, UK (1997), pp. 231–232.

[12] S. M. Gass, C. Madden, D. Preston, and L. Selinker (eds.) *Variation in second language acquisition*, vol. 1: Discourse and pragmatics, Clevedon, England (1989).

[13] S. M. Gerrish and D. M. Blei. *A language-based approach to measuring scholarly impact*. In Proceedings of International Conference of Machine Learning (2010).

[14] D. Gîfu. Diachronic Evaluation of Newspapers Language between Different Idioms at the first *Workshop on Natural Language Meets Journalism*, NLPMJ-2016, held at the *International Joint Conference on Artificial Intelligence*, IJCAI-16, 9-15 July 2016, New York, USA.

[15] D. Gîfu. *The Emotional Orientation*. In: Proceedings of the third Conference of Mathematical Society of the Republic of Moldova, "IMCS-50", 9-23 August 2014, Chișinău, Republic of Moldova (2014), pp. 511-516.

[16] D. Gîfu. *Contrastive Diachronic Study on Romanian Language*. In: Proceedings FOI-2015, S. Cojocaru, C. Gaindric (eds.), Institute of Mathematics and Computer Science, Academy of Sciences of Moldova (2015), pp. 296-310.

[17] D. Gîfu and R. Simionescu. *Tracing Language Variation for Romanian* at the 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2016, Konya, Turkey (2016).

[18] R. Mihalcea and V. Năstase, V. *Word epoch disambiguation: Finding how words change over time*. In: Proceedings of ACL 2012 (2012).

[19] C. Miller-Broomfield. *Romanian: The Forgotten Romance Language*. In: Unravel: The Accesible Linguistics Magazine (2015), http://unravellingmag.com/articles/romanian-the-forgotten-romance-language/.

[20] O. Popescu and C. Strapparava. *Behind the Times: Detecting Epoch Changes using Large Corpora*. In: International Joint Conference on Natural Language Processing, Nagoya, Japan, 14-18 October 2013 (2013), pp. 347-355.

[21] O. Popescu and C. Strapparava. *Time corpora: Epochs, opinions and changes*. Knowledge-Based Systems (2014).

[22] O. Popescu and C. Strapparava. Semeval-2015 task 7: *Diachronic text evaluation*. In: Proceedings of SemEval (2015).

[23] R. Simionescu. *UAIC Romanian Part of Speech Tagger*, resource on nlptools.info.uaic.ro, "Alexandru Ioan Cuza" University of Iași (2011).

[24] D. Tufiș, E. Barbu, V. Barbu Mititelu, R. Ion, L. Bozianu. *The Romanian WordNet*, Romanian Journal of Information Science and Technology, 7, 1–2 (2004), pp. 107–124.

[25] D. Tufiş, D. Cristea. *Ro-BALKANET – ontologie lexicalizată în context multilingv pentru limba română*. In: Limba Română în Societatea Informațională – Societatea Cunoașterii, Ed. Expert, Tufiş, D., Filip, F. Gh. (coord.) (2002), pp. 139-166.

[26] X. Wang and A. McCallum. *Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends*. In: KDD 2006, USA (2006).

[27] X. Wang, M. S. Gerber, and D. E. Brown. *Automatic Crime Prediction using Events Extracted from Twitter Posts*. SBP, LNCS 7227:231-238 (2012).

Daniela Gîfu

Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
e-mail: daniela.gifu@info.uaic.ro