

Connecting the Content of Books to the Web and the Real World

Dan Cristea^{1,2}, Ionuț Pistol¹, Daniela Gîfu^{1,2}

¹Romanian Academy – Iași branch, Institute for Theoretical Computer Science
²„Alexandru Ioan Cuza” University, Faculty of Computer Science,
{dcristea, ionut.pistol, daniela.gifu}@info.uaic.ro

Abstract. In the last decade, the interest in developing a market of e-books has greatly increased, being preferred by many readers for their condensed volume, easy acquisition, some facilities that big providers started to introduce, such as the possibility to make personal notes, to search words in dictionaries and inside the book itself. However, this new format can have tremendously many more facets than proved by the printed book and has an extraordinary power to resonate differently with its particular readers. In this paper, we describe a technology that combines natural language processing with entity linking, web cartography, web mapping and augmented reality in order to bring to the reader of books new levels of knowledge, connection and leisure. The MappedBook is an electronic artifact that complements the textual content of a book with hyperlinks in a virtual space and augmented reality functionalities. Criteria to form social networks based on the MappingBooks technology are suggested. This study has a strong view on tourism business and educational systems.

Keywords: reading, name entity recognition, entity linking, geographical maps, augmented reality, evaluation of user satisfaction

1 Introduction

The last years witnessed a vividly developing market of e-books, which are preferred by many readers for their condensed volume, easy acquisition, some facilities that big providers started to introduce, such as the possibility to make personal notes, to search words in dictionaries and inside the book itself (e.g., the Amazon Kindle offer minimal, but very useful, interaction in the form of dictionary or Wikipedia definitions, lists of characters and their occurrences in the book) and, maybe not the least – ecological reasons, to diminish the abusive cutting down of trees. However, this new format can have tremendously many more facets than proved by the objects already existent in shops and implemented on reading devices.

The rich experience that a book creates for a reader is only partly due to the written text. Complementary knowledge adds to the genuine exposure of facts and descriptions. Moreover, the text has the extraordinary power to resonate differently in its particular readers, and these inner evoked representations and feelings may or may not match what the author intended (which is, clearly, the same for all readers). Depending on the background of the reader, their preferences and hobbies, some

mentions of the book may be ignored, some may be only shallowly deepened, and some may cause an interruption in the reading process to search for and retrieve additional information.

In this paper we describe a project and a technology intended to augment the text of a book with additional information acquired from the virtual and real world, which could be specific to the reader, as for instance their instantaneous position, the moment of reading and preferences.

The paper is organized as follows: section 2 gives a general overview of the MB technology, section 3 describes the collection of data used to train the chain of automatic linguistic annotations, section 4 presents the methodology that engines the automatic annotations, section 5 shortly depicts the main features of the implementation, section 6 discusses the evaluation of the technology, section 7 prefigures other functionalities that regard social networking, section 8 investigates connections of our work with other research, and section 9 offers a free discussion on the benefits of this realization, with a glance into the future.

2 Architecture and Functionality

MappingBooks (MB) is a recently finished project (June 2014 – Sept. 2017), which developed a new type of electronic product with a high impact in education, and potentially on tourism and leisure. The specifications of the project state the main envisioned users of the developed MB application to be school pupils and students. The technology mixes methods from natural language processing, web cartography and mapping, mixed reality techniques and ambient intelligence/ubiquitous computing to link mentions of geographical entities existing in school manuals onto data existing on the web, to localize these entities on 2D hypermaps and to put them in correlation with the reader's location and related data.

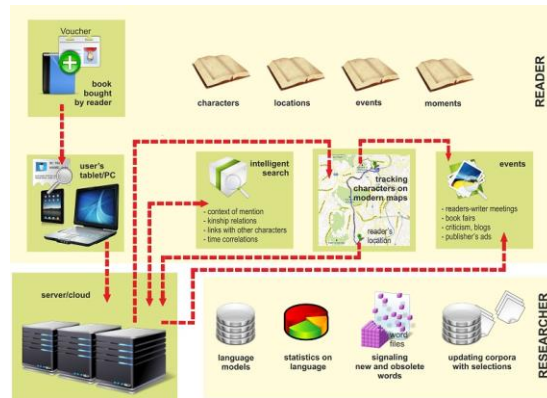


Figure 1: A bird's eye view of the MB functionality

The main result of the MB project was a functional proof of concept a *MappedBook*, i.e. an electronic artifact that complements the textual content of a book with hyperlinks in a virtual space and augmented reality functionalities.

Figure 1 shows the main ideas behind the project: when buying a book (in paper or eBook version) that is known to have a MappingBooks counterfact, the reader receives a voucher that allows them to install the application on their own mobile devices. The user's device will thus be coupled with a server which hosts the linguistically and geographically processed book and will allow two main types of interactions between user and app: either at the user's initiative, allowing them to search the electronic text and connect to sources of information related to the text on the web, or at the app's initiative, being prompted when the user happens to be physically located in the proximity of geographical entities mentioned in the book. Moreover, the dense annotations automatically produced over the texts and their links in the internet could be accumulated on the server and offered freely for computational linguistics research purposes.

For the linguistic processing, the project implements an NLP chain that includes: name entity recognition (NER-phase 1), sentence spitting, tokenization, POS-tagging, noun phrase (NP) chunking, NER-phase 2 and semantic relations detection. In MB, like in TTL (Tokenizing, Tagging and Lemmatizing) [Ion, 2007], some NER functions are applied before the sentence splitter, in order to correctly recognize dots to abbreviation as opposed to those signaling limits of sentences. More evolved NER functions are applied after the NPs are identified. Our NER implementations mix brute force methods (as provided by the use of a large collection of proper nouns, including abbreviations) with symbolic methods (a collection of regular expressions used to disambiguate types of proper names as indicated by contexts). Once name entities of type location are identified, semantic relations are searched for, by applying a collection of lexical-syntactic patterns anchored in triggers (keywords specific to different types of semantic relations),

Further on, mentions of geographic locations in the original text of the book are linked into the virtual space (in this implementation – Wikipedia). The application is sensitive to the location where the user (presumably the book reader) happens to be while using the application, as seized through their mobile phone or tablet, and if the mobile device is a tablet, elements of augmented reality can be experienced, as for instance arrows superposed with the captured image that indicate locations mentioned in the book and which are in a reachable vicinity of the reader. The project demonstrated these ideas by elaborating a functional prototype and transposing into the MB technology a first book, a Geography Schoolbook for the 8th grade¹.

3 Data Set

Texts are made of mentions of entities, events, thoughts, sensations, placed on one or more temporal axis. Representing these elements belonging to a more or less concrete world and recovering in our minds connections between them is indispensable for understanding the text. Without them, we would not be able to express any continuous thought rooted in the text, perhaps more than remembering a succession of places, characters, sensations and ideas, but each isolated from the others.

In MB, references to geographic entities are one or more contiguous words (usually NPs) that identify real-world locations (e.g. *Romania*, *Black Sea*, *mount*

¹ Neguț, S., Apostol, G., Ielenicz, M.: “Geografie”, Humanitas Educațional, București (2008).

Ceahlău). The notion of a geographical entity is a customization of the more general notion of named entity, which identifies textual terms whose meanings are known outside the text in which they are mentioned. In the case of geographical entities, external significance has the particular property that it can be generally associated with an area of a map and it possesses some specific properties (height, surface, population, etc.).

To help the machine learning process, we have manually developed an annotated file showing in XML entities and relations between them at both the surface language level and the knowledge representation level [Sălăvăstru and Gîfu, 2015; Gîfu et al., 2015]. Although, in general, entities are of a very diverse nature (persons, animals, places, organizations, moments of time, etc.) in this study we have been concerned only on persons and locations. Name entities (ENTITY elements) include at least one WORD and contain minimum the TYPE and SUBTYPE attributes. We exemplify below the main types and subtypes.

- a) Location (TYPE = "PLACE"), with subtypes: "VILLAGE", "TOWN", "CITY", "DISTRICT" and "REGION".
- b) Relief form (TYPE = "LANDFORM") with subtypes: "FIELD", "PLATEAU", "HILL", "PEAK", "BOTTOM-LAND", "DEPRESSION", "DEFILE", etc.
- c) Water (TYPE = "WATER") with subtypes: "LAKE", "BROOK", "DELTA", etc.
- d) Size (TYPE = "DIMENSION"), naming lengths, elevations of mountains, distances between localities, depth of the waters, etc. Subtypes here are: "HEIGHT", "DEPTH", "LENGTH", etc. with the unit of measurement expressed by the MEASUREMENT attribute ("M" for meter).

Among the extremely large class of semantic relations that a text could express, we deciphered only:

- a) "REFERENTIAL", with subtype "COREF", to express coreferential links, also complemented with the attributes FROM (indicating the ID of the right-most-in-the-text entity) and TO (its antecedent). For instance:

1: [*Romania*] ... 2: [*our country*] ... \Rightarrow <RELATION TYPE = "REFERENTIAL" SUBTYPE = "COREF" FROM = "2" TO = "1" />;

- b) Spatial relations: TYPE = "SPATIAL". Examples of subtypes are:

- "NEAR" and "FAR" – to express the proximity and farness between two locations,
- "POSITION", with the cardinal point indicated by the CARDINAL argument, for instance:

1: [*Parâng Mountains*] ... *is located at the East of* 2: [*Retezat Mountains*] \Rightarrow <RELATION TYPE = "SPATIAL" SUBTYPE = "POSITION" CARDINAL = "EST" FROM = "1" TO = "2"

- "PART-OF", for instance:

1: [*Romania*] *concentrates on its territory two thirds of the chain of* 2: [*Carpathian Mountains*] \Rightarrow <RELATION TYPE = "STRUCTURAL" SUBTYPE = "PART-OF" FROM = "2" TO = "1"

- "SOURCE", indicating the source of a running water, as here:

1: [*Târnava Mică*] *with springs in* 2: [*Gurghiu Mountains*] \Rightarrow <RELATION TYPE = "STRUCTURAL" SUBTYPE = "SOURCE" FROM = "1" TO = "2"

- "CONFLUENT-OF", indicating the confluent of a river, as here:

The basin of 1: [Mureş] ... 2: [Târnava], its most important confluent ⇒ <RELATION TYPE = "STRUCTURAL" SUBTYPE = "CONFLUENT-OF" FROM = "2" TO = "1"

In the present implementation we were not concerned with: negative relations (*Romania ... is not part of the Balkan Peninsula*), spatial relations of a very complex nature (*Romania is at a latitude halfway between the Equator and the North Pole*), or past names of toponyms (*Histria, Tomis, Callatis, Apulum, Ampelum, Napoca, Potaissa, Sucidava*, etc.).

4 Methodology of Generation of Inner and External Links

The main module of our NER is a gazetteer (GAZ), which uses a list of toponyms and other geographical names grouped by categories in order to identify potential candidates. Given the nature of the texts actually used to develop the MappingBooks project (all in the domain of Romanian geography schoolbooks), the gazetteer was developed to include as many geographic names as relevant to the geography of Romania. We identified 15 general types, including 103 subtypes, out of which 67 are of a geographic nature: entities of these types may be associated with locations on the map. In order to populate the gazetteer based on the defined types, we used Geonames as the main resource, frequently used by the community of researchers and developers in the field. Geonames has been developed from free government and educational resources, later supplemented with user-contributed data and validated by specialists. For Romania, Geonames includes 25,951 unique entities, with over 45,000 alternative names.

To complete the Geonames geotagger, it was necessary to map the types / subtypes defined in MappingBooks to those in Geonames. Due to the much larger list of subtypes in Geonames (652 vs. 67 in MappingBooks), we generally paired multiple Geonames subtypes to one MB subtype. The lower number of subtypes in our classification reduces the ambiguity of the GAZ module, the level of detail being considered sufficient for a potential user of the application [Cristea *et al.*, 2016a].

In parallel, a pattern-matching module (PAT) uses a set of templates, identified in that part of the text manually annotated, to find other candidates for geographic entities. The major difference between GAZ and PAT modules is that GAZ uses the exact textual form of potential candidates, while PAT also uses contextual information.

The templates used have been built using Graphical Grammar Studio (GGS) [Simionescu, 2011], a graphical environment that allows the development and application of contextual grammars on annotated and raw texts. GGS defines a constraint language that allows the description of composite elements, complemented with look-ahead and look-behind assumptions, and the association of scores on the arcs defining the parsing sequence. A network described in GGS (directed chart) consumes an input and produces an annotated output with additional information according to the paths traversed in the network. The nodes in these graphs define conditions that can be reached, while the directed arcs identify conditions for consumption of the input tokens. A GGS grammar is basically a finite state machine. The PAT module receives an XML annotated document, resulted from the sentence splitting, tokenization and POS-tagging pipeline, and a GGS network. If,

corresponding to a subsequence of the input sentence, one path is found that links the initial node to the final node in the network, then that input sequence is consumed and an XML sequence is generated. The result is a document containing annotations added over the names found in the resource used (type, subtype, coordinates, other geographic attributes). In the case of an ambiguity (identifying the same name in different categories), the annotation will add all possible variants and the decision is taken by the PAT module.

In [Cristea *et al.*, 2015], the issue of annotating relations linking entities in texts is addressed. In the mentioned paper we say that any mention of an entity (restricted only to persons, gods, group of persons and of gods) is a mapping from a text expression to a corresponding «container», which is associated with each entity at its first mention and is subsequently filled in with pieces of information that define details, as contributed by the text (name, sex, kinship connections, etc.). In all corpus-based approaches, mentions, not containers, are annotated, but if semantic reasoning is tried on these annotations, then the more complex containers are the keepers of the semantic representations. In MappingBooks we adopt the same simplified representation, associating inner links to mentions of entities and not to their containers. As such, all relations will have a local dimension, being perceived contextually, in the exact place of their occurrence in the text.

The last step of the linguistic processing chain is realized by the Relations Detection module (RD), which is responsible for putting in evidence semantic relations mentioned in the text. The set of patterns of the RD module are regular expressions [Colhon *et al.*, 2016] that link fixed textual sequences (for instance: *is/was redone by*, *is/was finished by*, *is/was the creation of*, *is/was realized by*, etc.) with two slots, XML ENTITY elements (in the example above, the first of type institution and the second of type person). This way, buildings build by persons are evidenced in the text.

Other relations put in evidence geographical relations that link two entities of type location. All patterns are extracted from examples, by observing common types of relevant text fragments. At the end of the described chain, a heavily annotated XML file is accumulated. Thus, the entity names, as identified by the NER component, are enriched with information of historical and geographical nature.

5 Interface

The user interface scales down to different users' devices running Android. Without entering into details, the dimension of each displayed component is expressed in percentages of the dimensions of the screen. Thus, on a larger device, a button will appear larger, and on a smaller device, it will appear smaller. In other words, the same aspect ratio will be maintained regardless of the device. This solution involves also the use of different geographical resources of different pixel densities.

The following geographical resources are running on the server, continuously fetching data towards users' devices: GeoNetwork² – an open source web platform allowing the creation of spatial data catalogues, the editing of their metadata and

² <https://sourceforge.net/projects/geonetwork/>

spatial search by key words; OpenLayers 3.19.1³ – a Javascript library that allows the display of web maps in any browser; Nominatim – a search engine for geo-data operated by a web interface; GeoServer – a Java-based application that implements web services to create and deliver spatial data as maps.

Finally, an augmented reality application [Cristea and Pentiu, 2016] couples the geographical position of the device with the image captured by the built-in camera, on which it projects Points Of Interest (POIs) that appear mentioned in the book and are momentary in the close vicinity of the user. Different pruning strategies are used to cut down the number of POIs when an area larger than 5km is selected by the user, or otherwise the image would become too dense in information and hard to read. The purger ranks the localities in the selected area by the population number (information stored on the geo-server) and retains the most important. Other strategies have been used to stabilize the image (for instance, a frequency actualization threshold had to be established for purging of the data coming from the built-in compass, originally too sensible at very small rotation angles). When the user touches a POI displayed on the image, the interface will display its mentions in the text as well as supplementary information got from external sources (Wikipedia).

6 Results

A development objective of the mobile application created within the MB project has always been the users' satisfaction. The target group identified for the purpose of demonstrating the application was formed of school pupils studying "Geography".

After several stages of application testing in partners' project laboratories, we went to evaluate the app in the context of the identified target groups. The investigative approach trying to understand how potential users relate to MappingBooks was based on the use of a social survey using the questionnaire as a working tool. The investigation was based on twenty closed statements, with Likert type scales, and two open questions [Lewis and Sauro, 2009], and mainly pursued:

- measure the overall satisfaction (usability);
- measure the change of appetite (negative-positive) of pupils towards the studied notions when using the MB technology as compared to the classical use of the manual;
- identify the interestingness of the implemented features of the application;
- identify advantages and disadvantages;
- identify design suggestions (recommendations).

The sample used for the assessment was probabilistic, consisting of 156 respondents, VIIIth and IXth grade school pupils from two Romanian cities (Iași and Botoșani). The construction of the sample was that of the "snowball" [Sauro and Lewis, 2012]. The sample has the following characteristics: 50% male, 50% female; 21.2% were in the VIIIth grade, and 78.8% – in the IXth grade; of those in the IXth grade, 59.3% are in the mathematics-computer science profile, 15.4% in the philology profile, and 25.2% in the natural sciences profile. In relation to the general connection of the respondents with the technologies, 53.2% use the mobile device very frequently, 30.8% – frequently, 10.3% – rarely and 5.8% – very rarely or never.

³ <https://openlayers.org>

The possible answers were scaled from the lowest to the highest values of the item "agreement". 60.3% of respondents agreed with the statement "I like to use this app" and 30.8% were undecided. 90% of questioned people agreed that the app could be potentially further developed. Significant is also the score of over 80% for agreeing to the statement "The application helps me understand the notions of geography better than the schoolbook", which indicates the direct didactic benefits perceived by the target group in fixing the notions expressed in the text.

The lowest total agreement (58.3%) was received for the statement, "I like how I can interact with the text of the book in the application." The dissatisfaction proved to be related to the type of the device: the majority of pupils were having mobile phones and therefore small size screens, as compared to larger displaying texts on tablet screens. Actually, 66% of the respondents agreed with the statement "I would like the mobile device we tested the app on to have a larger screen."

A significant difference in satisfaction was between the VIIIth and IXth graders, the IXth graders proving more than 20% greater agreement on the statement "The application is easy to use", respectively 48.8% versus 27.3%. The probable cause of this difference is the greater experience that class IX students have with both the contents of the manual and the use of test devices.

Regarding the possible enhancements to the application (included in the last 7 statements of the questionnaire), the highest value of the "agreement" item was obtained for "being able to communicate with other users via the application" (23.7%) [Novak-Marcincin et al., 2014; Gifu & Teodorescu, 2014], a recommendation that we retain for possible future developments (see the next section for a short description).

7 Networking Rooted on Common Readings

The evaluation seems to reflect a strong desire of users to interact with other people through apps like MB. As proved in those areas in which NLP technologies have been particularly successful in enhancing a social network user experience, which are e-learning [Romero and Ventura, 2013] and consumer reviews [Netzer *et al.*, 2012], links between users can be established based on their consumed product, in both cases this being shown to significantly increase users satisfaction and retention.

When the processed text (the host document) will be freed from the constraints of approaching only the domain of teaching of geography, which we had to observe in the present deployment, the annotations automatically added to the text could be extremely rich. Items marked could be entities with very diverse real-world significance: locations, institutions, notorious people, historical events, etc. as well as links between these entities, both inside and outside the text. But semantic and locational links could be a good source of forming social communities of readers. Some examples of possible social networks are [Cristea *et al.*, 2016b]:

- co-readers of a certain book B could be formed out of those readers who have declared the "subscribed for B " field visible;
- co-readers of a certain book B AND actually being in the proximity of a location L would include readers who also declared visible their "instantaneous location";
- co-readers of B (a book) AND co-track of T (a trajectory rooted in that book) would include readers reading a book and having followed a more or less similar sequence of places mentioned in that book (as appears, for instance, in a traveling

guide).

It is easy to imagine other ways to form social communities rooted in readings. Imagine, for instance, that common readings would be intersected with attended places in the past and levels of friendship reported by social media, like Facebook or WhatsApp, or real-world events and entities mentioned in a book associated with real-world locations and particular moments of the year/day.

A good example of maps rooted on literature are those built within the Vilnius' literature mapping project⁴, developed by several Vilnius University literature scholars and cartographers. Their maps, with data manually acquired from Lithuanian texts, include mentions of places of the city of Vilnius. It is clear that the MB technology could be used to automatize such an enterprise, to build cultural maps and to connect people based on their readings. In enterprises of this kind, a great importance will have name entity coreferences in multiple documents.

8 Related Work

While performing this study, we had in mind three components, for which we searched appropriate approaches in the literature: natural language processing (NLP), linguistic linked open data (LLOD) and information extraction (IE).

As an NLP approach, entity identification and linking touches some issues related to word sense disambiguation. Wikipedia and Open Linked Data entities inventories combine numerous sources of knowledge, such as DBpedia, WordNet, Freebase, GeoNames, US Census, PubMed, etc. The task of knowledge base population with a subtask of entity linking was launched in 2009 [McNamee and Dang, 2009] focusing the research in this area and providing gold standard datasets and evaluation measures. This task is still very challenging; the most successful approaches rely on ranking the candidates for linking in a supervised manner, using a diversity of features like similarity metrics, entity profiling and so on. To exploit the coherence of a text as a "feature", some methods can target single entities at a time [Zheng et al., 2010] and other several groups [Milne and Witten, 2008; Wick et al., 2013; Cheng and Roth, 2013]. It is well known that the performance of the systems depends on the evaluation metrics in connection with how a task is formulated. State-of-the-art results in named entity recognition and linking reach an F-score of approximately 77% [Ji et al., 2014]. In MB, we restricted the task of entity identification and linking to only full and pronominal mentions, by applying a rudimentary anaphora resolution engine, ignoring other types of surface referencing. Zhou et al. [2014] report successful supervised entity discovery and linking, but our task was much simplified by the access to collections of names, such as Geonames.

Relations detection, as performed in MappingBooks, has many similarities with task of Information Extraction (IE) – for instance, detection of the date of an event. In MB the goal of applying IE techniques was to gather additional information that could be useful to the reader. Classical exercises campaigns in IE are the DARPA initiated Message Understanding Conference series, which covered various types of extracted information – from structured relations, to more free-form text fragments. Approaches range from unsupervised [Riedel et al., 2010] or Open Information Extraction [Wu

⁴ http://www.vilniusliterature.flf.vu.lt/en/?page_id=18

and Weld, 2010] to self-supervised [Downey *et al.*, 2010; Fader *et al.*, 2011] and supervised and targeted using seed relations from resources such as FreeBase [Mintz *et al.*, 2009] or Wikipedia infoboxes [Wu and Weld, 2010].

9 Conclusions

It is clear that not all types of books should stand the same MappingBooks processing in order to become a *MappedBook*. A large palette of editing tools can be put in the hands of the book editor that they could employ to augment the leisure and the benefits of the reading act.

Certain semantic processing techniques – such as named entity recognition and entity linking – are largely insensitive to genres and topics. Others, like event identification, temporal analysis and coreference resolution, may be more difficult in the presence of dialogue and shifting or alternating timelines. Fiction and non-fiction books may have different requirements with respect to information to be added to the identified entities, and within fiction books we may encounter a mixture of fictional and non-fictional characters, events and locations. The purpose of the final book object could also vary – being destined for education at various levels, as guides during travelling, or reading for pleasure.

Our vision is that the MappingBooks technology or one emerged from ideas developed in this project could put at the basis of a future printing house technology, in which the book editor would be given a palette of tools, any of them added to the book under processing, yielding the inclusion in the final electronic artifact of a theme that exploits the linkage of entities, internally or/and externally, including elements of augmented reality or not, synchronized or not to the specificities of the reader, able to trigger or not formation of social communities, etc.

All these being set, we are also aware that a technology like MB should be handled with care. One cannot avoid annoying questions, like the following: to what degree should the technology interfere with the act of literary consumption? Augmented reality techniques are at our touch, still which of the descriptions of perception mentioned in a book should be added to the text? For instance, perhaps we would not want to rebuild in the virtual space sensations, images, sounds or smells as described by the author of a fiction book. What we want is only to bring the reader to another level of knowing notorious entities mentioned there and maybe to another level of sensing the text. But how rich in electronic effects should this level be must be judged well or otherwise the reader can be suffocated and the text trivialized.

Acknowledgments: MappingBooks was financed by UEFISCDI (Executive Agency for Higher Education, Research, Development and Innovation Funding) of the Romanian Ministry of Education and Research, under the Partnerships Programme (PN II, competition PCCA 2013), project code: PN-II-PT-PCCA-2013-4-1878. Partners in the project have been: „Alexandru Ioan Cuza” University of Iași, SIVECO Romania, and „Ștefan Cel Mare” University of Suceava. Our colleagues Mihai Niculiță and Dorina Ticu have designed and implemented the geographical mapping and the evaluation with subjects, respectively. We thank Tommaso Caselli, Piek Vossen, Vivi Năstase, Danilo Giampiccolo and Carlo Strapparava for a past collaboration fostering some ideas mentioned in this paper.

References

1. Cheng, X. and Roth, D. Relational inference for wikification, Proceedings of EMNLP (2013).
2. Colhon, M., Cristea, D., Gifu, D. Discovering Semantic Relations within Nominals. In: D. Trandabăț and D. Gifu (eds.): *Proceedings of the Workshop on Social Media and the Web of Linked Data, RUMOUR-2015*, A satellite event of EUROLAN-2015, Sibiu, Romania, July 2015, Springer (2016).
3. Cristea, D., Gifu, D., Colhon, M., Diac, P., Bibiri, A.-D., Măranduc, C. and Scutelnicu L.-A. Quo Vadis: A Corpus of Entities and Relations. In N. Gala, R. Rapp and G.B. Enguix (eds.): *Language Production, Cognition, and the Lexicon*, Springer International Publishing Switzerland (2015).
4. Cristea, D., Gifu, D., Pistol, I.-C., Sfirnaciuc, D., Niculiță, M. A Mixed Approach in Recognising Geographical Entities in Texts. In: D. Trandabăț and D. Gifu (eds.): *Proceedings of the Workshop on Social Media and the Web of Linked Data, RUMOUR-2015, A satellite event of EUROLAN-2015*, Sibiu, Romania, July 2015, Springer International Publishing (2016a).
5. Cristea, D., Pentiu G.-Ș. A New eBook Concept and Technology Dedicated to Geographical Information, *Proceedings of the 18-th International Conference on Scientific Research & Education in the Air Force – AFASES*, Brașov, May, pp., 417-412, DOI: 10.19062/2247-3173.2016.18.1.57 (2016).
6. Cristea, D., I. Pistol, D. Gifu and D. Anechitei. Networking Readers: Using Semantic and Geographical Links to Enhance e-Books Reading Experience, in *Proceedings of the 2nd Workshop on Social Media and the Web of Linked Data, RUMOUR 2016*, together with the 8th International Conference on Computational Collective Intelligence Technologies and Applications, ICCCI 2016, September 28-30, 2016, Halkidiki, Greece (2016b).
7. Downey, D., O. Etzioni, S. Soderland. Analysis of a probabilistic model of redundancy in unsupervised information extraction, *Artificial Intelligence* 174(11): 726-748 (2010).
8. Fader, A., S. Soderland, O. Etzioni. Identifying relations for open information extraction, Proceedings of EMNLP (2011).
9. Gifu, D. and Teodorescu, M. Communication concepts vs. Sciences concepts. In: ILSHS, 18, 48-57 (2014).
10. Gifu, D., Pistol, I., Cristea, D. Corpus of Entities and Semantic Relations with Application in Geographical Domains. In: *Proceedings of the 11th ConsILR*, D. Gifu, D. Trandabăț, D. Cristea, D. Tufiş (eds.), "Al.I. Cuza" University Publishing House, Iași, 67-78 (2015).
11. Sălăvăstru, A. and Gifu, D. Annotating Geographical Entities. In: *Research in Computing Science*, Al. Gelbukh, Samhaa R. El-Beltagy (eds.), issue 90: *Advances in Computational Linguistics*, México, 45-56 (2015).
12. Ion, R. Methods of Automatic Semantic Disambiguation. Applications for English and Romanian (in Romanian). PhD. Thesis, Research Institute for Artificial Intelligence, Romanian Academy, Bucharest (2007).
13. Ji, H., J. Nothman, B. Hachey. Overview of the TAC-KBP 2014 Entity discovery and linking tasks. In: *Proc. Text Analysis Conference (TAC)* (2014).
14. Lewis, J.R. & Sauro, J. The factor structure of the system usability scale. HCII 2009, San Diego CA, USA (2009).

15. McNamee, P. and Dang, H.T. Overview of the TAC 2009 Knowledge Base Population track. In: Proc. Text Analysis Conference (TAC) (2009).
16. Milne, D. and I. Witten. Learning to link with Wikipedia. In: 17th ACM Conference on Information and Knowledge Management, Napa Valley, CA, 2008, pp. 509-518 (2008).
17. Mintz, M., S. Bills, R. Snow, D. Jurafsky. Distant supervision for relation extraction without labeled data, Proceedings of ACL (2009).
18. Netzer, O., Feldman, R., Goldenberg, J., Fresko, M. Mine your own business: market structure surveillance through text mining. *Mark. Sci.* 31(3), 521-543 (2012).
19. Novak-Marcincin, J., Gifu, D., Nicolescu, A. The standard of axes in ontology of communication. In: *International Letters of Social and Humanistic Sciences* 30 (2), 176-183 (2014).
20. Riedel, S., L. Yao, A. McCallum Modelling relations and their mentions without labelled text. In: J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part III, LNAI 6323, pp. 148–163, Springer-Verlag Berlin Heidelberg (2010).
21. Romero, C. and Ventura, S. Data mining in education. *Wiley Interdisc. Rev. Data Min. Knowl. Disc.* 3(1), 12-27 (2013).
22. Sauro, J., & Lewis, J.R. Quantifying the user experience: Practical statistics for user research. Morgan Kaufmann, Waltham MA, USA (2012)
23. Simionescu, R. Graphical grammar studio as a constraint grammar solution for part of speech tagger". In *Proceedings of the International Conference Resources and Tools for Romanian Language – ConsILR-2011*, Bucharest, “Alexandru Ioan Cuza” University of Iași Publishing House (2011).
24. Wick, M., S. Singh, H. Pandya, A. McCallum. A joint model for discovering and linking entities, *Proceedings of the Third International Workshop on Automated Knowledge Base Construction (AKBC)* (2013).
25. Wu, F. and D. Weld. Open Information Extraction using Wikipedia, Proceedings of ACL (2010).
26. Zheng, Z., F. Li, M. Huang, X. Zhu. Learning to link entities with knowledge bases, Proceedings of NAACL-HLT (2010).
27. Zhou, S., C. Kruengkrai, K. Inui. Tohoku’s entity discovery and linking system at TAC 2014. In: Proc. Text Analysis Conference (TAC2014).