




Towards Interoperability in the European Poetry Community: The Standardization of Philological Concepts

Helena Bermúdez-Sabel¹(✉) , Mariana Curado Malta^{1,2} ,
and Elena Gonzalez-Blanco¹ 

¹ Laboratorio de Innovación en Humanidades Digitales,
Universidad de Educación a Distancia (UNED), Madrid, Spain
{[helena.bermudez](mailto:helena.bermudez@linhd.uned.es),[mariana.malta](mailto:mariana.malta@linhd.uned.es),[egonzalezblanco](mailto:egonzalezblanco@linhd.uned.es)}@linhd.uned.es

² Polytechnic of Oporto, CEOS.PP, Oporto, Portugal

mariana@iscap.ipp.pt
<http://linhd.uned.es>
<http://iscap.ipp.pt>

Abstract. This paper stems from the Poetry Standardization and Linked Open Data project (POSTDATA). As its name reveals, one of the main aims of POSTDATA is to provide a means to publish European poetry (EP) data as Linked Open Data (LOD). Thus, developing a metadata application profile (MAP) as a common semantic model to be used by the EP community is a crucial step of this project. This MAP will enhance interoperability among the community members in particular, and among the EP community and other contexts in general (e.g. bibliographic records). This paper presents the methodology followed in the process of defining the concepts of the domain model of this MAP, as well as some issues that arise when labeling philological terms.

Keywords: Digital humanities · Literary data · Metadata application profile · Terms standardization · Vocabulary encoding scheme · Linked open data · Semantic web

1 Introduction

The need for information exchange has made it necessary to create international standards in most fields. The humanities have evolved independently from other fields [1], as there are important factors such as history, creativity or self-identity that influence each particular tradition with different results. The case of poetry is especially significant, as every country, group and literary genre has followed an independent and idiosyncratic path [2]. As a result of this, online access to poetry collections is highly fragmented [3].

Our challenge as researchers is to reduce the digital gap between the humanities and technology, aiming for interoperable solutions and for an interdisciplinary approach with innovative results that transcend the current state-of-the-art.

Individual literary works have a set of metadata (such as author, title, date of composition or language) that is shared by all literary works. There is also a more limited set of properties that are specific to poetry (such as rhyme, metrical scheme or the number of stanzas). These tags are easy to recognize and would probably fit into any possible classification of any poetic corpus. The problem, however, is that every literary tradition has evolved in a particular way, coining different names and creating different conceptual systems to describe similar phenomena.¹ These differences apply to the way of naming lines, stanzas, poems, rhyme schemes and rhythmical patterns [5].

The first attempts of classification, shaped as metrical repertoires, were published in the late nineteenth century with the aim of gathering the lyrical materials that had been circulating around Europe. During the twentieth century, the number of repertoires and poetic catalogues grew, as also did the criteria and techniques to build them. In the nineteen nineties, some of these books were transformed into digital databases. A good example of this transformation is *Die nicht-lyrischen Strophenformen des Altfranzösischen* by Gotthold Naetebus, a poetic repertoire of French Medieval narrative poetry first published in 1891 [6]. This repertoire became the *Nouveau Naetebus* [7]: a digital resource that recovered the original book and incorporated useful information complemented by relevant studies. A very similar and interesting case is *Répertoire de la poésie hongroise ancienne* [9], a repertoire of Hungarian poetry by Horváth [8] already conceived as a digital collection. Further examples of this include the Galician-Portuguese database *MedDB, Base de Datos da Lírica Galego-Portuguesa*, directed by Mercedes Brea [10], which gathers both the secular lyric corpus and the metric and rhyming schemes of the metrical repertoire by Tavani [11], the metrical repertoire by Pillet and Carstens for Occitan lyrics [12], transformed into *BEDT - Bibliografia Elettronica dei Trovatori*, directed by Stefano Asperti [13], or *REMETCA: Repertorio métrico de la poesía medieval castellana* [14]. These, however, are just a few instances of scholarly attempts at metric systematization. Meanwhile, the traditional publications, printed in paper, continued with the repertoires of Antonelli [15], Pagnotta [16], Solimena [17] or Gorni [18] for the Italian lyrics, including, for instance, the Sicilian tradition or the so-called *Dolce Stil Novo*, Betti's repertoire for the *Cantigas de Santa Maria* [19], or a few Spanish examples such as [20] for medieval Catalan metrics or [21] for fifteenth-century Castilian metrics, which has been recently transformed into an online database [22].

Although all of these repertoires and databases focus on the “poem” as their object of study, the way in which they conceptualize the information is very different, and the resulting databases cannot communicate. Thus, interoperability is very complex for two main reasons. First, for technological reasons, as each database is modeled in a different way and may use a different technology, and second, due to the philological tradition, as each literary tradition has followed an independent path to encode its metrical and poetic information. The result is a variety of terminological and classification systems that are very difficult to communicate, especially when it comes to finding equivalences or looking for common models in different traditions.

¹ See [4] for a more detailed definition of “literary tradition”.

The present article reports a work-in-progress that deals with the standardization of philological concepts and terms. The research for this paper was carried out in the context of POSTDATA, a European Research Council Starting Grant project that aims to reduce the digital gap between the humanities and technology by looking for interoperability solutions. The paper is organized in five sections. The following section briefly explains the technological context of the project. Section 3 presents the methodology followed during the analysis of the databases of the repertoires, and Sect. 4 expounds the issues that arose during the process of concept identification. Section 5 presents our conclusions and briefly explores future work.

2 Linked Open Data and Interoperability

“The classic definition is literal, based on the etymology of the word itself—metadata is ‘data about data’” [23, p. 1]; it is data about concepts that represent tangible and non-tangible objects that exist in the real world. This data is published in the Web of Data, a “giant global graph” [24] of data. The Web of Data is not only about publishing data on the Web, it is also about linking this data [25] to enable the access and analysis of data from different sources to final users. This access takes place through the use of software applications that can connect data and and make inferences about it. The data that is linked and open in the Web of Data is called Linked Open Data (LOD).

Publishing data as LOD in the Web of Data is a process that must start with a good data modeling. Linked data must endorse a semantic model before being published. Since this data comes from different sources that incorporate multiple contexts within various cultures and languages, this process of modeling becomes very complex. According to [26], metadata must be modeled as a metadata application profile (MAP) in order to become interoperable. [27] define a MAP as “a generic construct for designing metadata records.”

Semantic interoperability is a very important issue in regard to LOD. Interoperability makes it possible for multiple systems with different programs, hardware, and data structures and interfaces to exchange data without previous communication, losing a minimum of content and functionality [23].

As mentioned above (Sect. 1), this work concerns the European Poetry community. It stems from the POSTDATA project, which has as one of its main goals to provide the means for this community of practice to publish data as LOD (for more details about the project see [28]). In the next section, we explain the process whereby POSTDATA is developing a MAP for EP in order to enhance interoperability in the EP community.

3 Methodology

In order to develop a MAP for the European Poetry, the authors are following a systematic set of activities defined by Me4MAP [29]. The issues addressed in

this paper are part of the S1 and S2 activities “Defining the Functional Requirements” and “Defining the Domain Model”, respectively.

The definition of the domain model, a common conceptual model that should represent the informational needs of the EP community of practice, integrates the data requirements that result from S1, together with the results of the following sub-activities:

- analysis of the data model of a representative sample of EP databases, and
- analysis of a survey addressed to the final users of the repertoires in order to understand the data needs of the users of poetry databases.²

The process of developing the domain model is highly iterative: it is made of micro-steps of analysis of data models in which every micro-step might feed a previous analysis, depending on the conclusion of the analysis at hand. The technique used to analyze each data model is described in [30] and is not further explained here, as it is not the object of this paper.

During the process of analysis of a data model (1) every concept of the database, as well as the properties that characterize that concept, are identified and (2) the relationships between concepts are also identified. In what follows, these conceptual elements will be referred to as “Concepts” regardless of whether they are concepts, properties or relationships between concepts.

As it has been mentioned, the analysis is very iterative, which means that similar Concepts that have been already identified in previous analyses are compared with the current analysis, and that those Concepts that are equivalent are given the same description and named in the same way.

Occasionally, the level of abstraction increases and names are changed retroactively; that is, previous analyses are re-evaluated. In the beginning of the process abstraction is low, but it increases with the number of analyses made. As a result, at the end of the procedure the level of abstraction is higher than it was at the start, which decreases the level of granularity of the final model. As in any process of semantic modeling, there is always some tension between interoperability and semantics. The level of semantics is related to the possibilities of data sharing, which means that the researchers look for the highest level of meaning in the definition of the Concepts without compromising interoperability. The same Concepts from different databases will contain data that can be shared, and different Concepts will contain data that cannot be shared. However, if a specific Concept is different but similar to other Concept that has been already identified while analyzing other databases, semantics may be lost in favor of interoperability gain: a new broader Concept is created.

Every Concept analysis integrates the actions presented in Fig. 1. The process begins with the identification of a Concept in a data model analysis. Then a study of similar Concepts in the previous data model analyses is carried out in order to understand if the Concept at hand is new or if it has already been identified. Two possibilities arise:

² Survey available at <http://postdata.linhd.es/limesurvey/index.php/113575>.

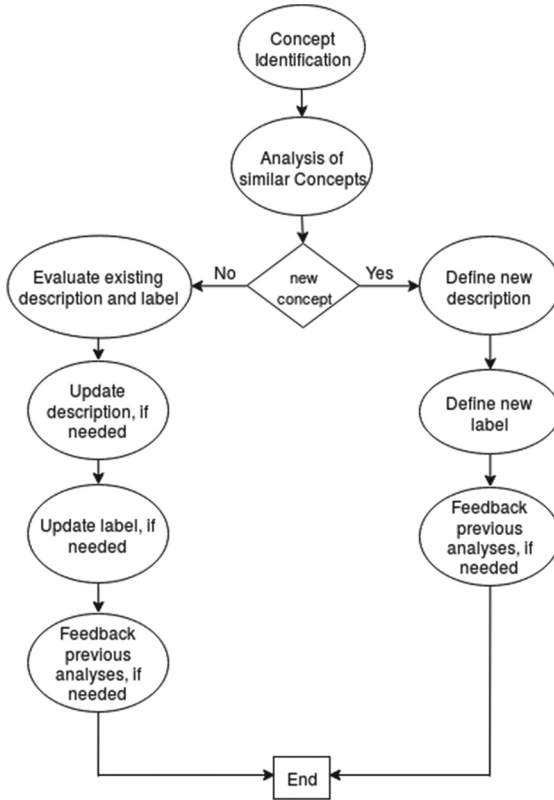


Fig. 1. Diagram with the sequence of actions of a Concept analysis process

- If there is a new Concept: a new description and a new label are defined. If the working group realizes that there are changes to be implemented in previous descriptions or labels during the definition of the description for the new Concept, an update is done in order to re-define preceding data model analyses. The process ends here and the analysis follows the study of a new Concept.
- If there is no new Concept: an evaluation of the description for the same Concept is undertaken, and from that evaluation, there might be an update of the description and label. If this update is required, changes are introduced to the previous analyses, otherwise, the process ends here. The analysis follows the study of a new Concept.

The process of analysis ends when there are no more new Concepts to analyze. The working group moves then to the study of the next data model. The procedure ends when all the data models are analyzed and there are no remaining feedback updates. As it has been explained, the methodology forces the process of analysis to continually reiterate over previous results, so the feedback process

is constant. This method avoids any over-representation or bias that could arise due to the order in which the databases are studied.

It is very important to integrate application domain experts since the early stages of development to guarantee the soundness of the foundations on which we keep building the project [29]. Every database has a delegate that verifies the accuracy of the interpretation of the concepts that the working group make. In addition, the aforementioned survey brings a broader sense of the data needs of the community of practice.

4 Concepts Standardization

Some of the fields of study included within the *jurisdiction* of philology have already paid great attention to the construction of standards. For instance, libraries have led the standardization projects in the humanistic field, and as such, the efforts to regularize the metadata of bibliographic records have been remarkably successful [31, 32]. The same can be said about other consolidated projects, such as the Text Encoding Initiative, which succeeded at presenting to the research community highly-adopted recommendations for multi-purpose encoding of digital texts [33]. Since these initiatives already convey the means to describe most of the physical artifacts studied by philologists, we make use of the same terminology as much as possible.

On a similar note, much work has been devoted in the last decades³ to the standardization of linguistic resources (for an updated state-of-the-art see [35]). However, in our study of poetry databases, we could corroborate that the implementation of linguistic standards in projects whose main object of analysis is not purely linguistic seems to be inadequate: none of the analyzed projects that include linguistic information used any existing standard or recommendation, but employed their own model. As a result, we decided that the conceptualization of morphosyntactic and lexical features will be left out during these first stages of the project. Nevertheless, we do model phonetic information when it is recorded due to metrical implications.

The aforementioned context justifies the focus of this paper on literary terms since they are the weakest link of the philological field in terms of standardization. Despite the existence of glossaries and dictionaries that collect information on literary concepts and their denomination, the goal of these works is to be a reference on the subject, to *accumulate* knowledge, not to offer a standardization proposal (least of all, a proposal to be implemented in a digital environment for interoperability purposes).

The aim of this paper is not to expound a taxonomy of literary devices, but to delineate the decision-making process that the authors went through during the selection of concepts and labels. The core points of our criteria are the following:

³ For instance, the *EAGLES Guidelines*, a set of recommendations for *de facto* standards and for good practice in computational linguistics, was already a consolidated project in 1996. For more information visit its website at <http://www.ilc.cnr.it/EAGLES/home.html>.

Lingua franca. When establishing a norm, choosing the most widespread label seems like a straightforward criterion. However, our starting point is a multilingual corpus, so the selection of the most common term could not be acknowledged. More so when the same language may present various terms to refer to the same concept. Considering the status of English as a *lingua franca* in science and technology, the most prevalent terminology in the English-speaking scientific community was taken as reference [36,37].

For instance, the concept that defines “the grouping of lines forming the basic recurring metrical unit in a poem or song” receives different names depending on the language (*estrofa*, *strophe*, *strofă* . . .). Moreover, certain particularities of a poetic school might determine the existence of various terms for this same concept in one language. In Portuguese, for instance, the term *cobra* is used in relation to the lyric poetic movement, while the more generic term *estrofe* is used in other contexts. In the present standardization proposal, all recurring line groupings were conceptualized as “stanza”.

Neutral terms. The selection of “marked” terms is avoided. Although existing terms are hardly ever completely neutral, great care was taken to select labels that do not have special connotations depending on the theoretical approach. For example, this proposal considers the term “apparatus” instead of the more common “critical apparatus” or even “*apparatus criticus*” in order to wrap any recording of textual variants under this label. Hence, we separate the concept from the ecdotic model of critical editions, enabling its use by genetic or synoptic editions without inheriting any theoretical baggage.

Semantic efficiency. The authors’ judgments as philologists when facing the materials also play an important role in the decision making process. For instance, our proposal pays great attention to the discerning of textual materials that are previous to the work under analysis (source texts), from those that are contemporary and which may present intertextual relations, and those texts that might have been influenced by the work at hand (derived). Opposite to these concepts, any type of subsequent bibliographical source (including previous editions of the work) is categorized separately. The reason behind these categorial distinctions is that we have identified the potentiality of the application of LOD to map interrelationships between texts. If not explicitly stated in the primary source, these intertextual relations are established by the researchers, which means that they are limited to their knowledge of other literary traditions. Thanks to this method, a researcher in French poetry, with no previous knowledge of Hungarian, can find out which Hungarian poems share the same Latin influences as poems in the corpus s/he is studying. Furthermore, many research questions regarding the existence of intermediary texts in the interrelatedness of literary works could be explored.

Although the level of abstraction in relation to the number of analyzed databases increases (see Sect. 3), the need to further restrain a concept arises sporadically. For instance, most projects include the concept of “metrical scheme”. Usually, in the context of a certain tradition the term is unequivocal. However, we decided to establish different categories to define metrical patterns according to the type of meter, that is, syllabic, accentual, accentual-

syllabic, and quantitative. Through this course of action, we facilitate that, for instance, a syllabic-verse tradition may find more efficiently similar metrical patterns by taking out the schemes that are not analogous.

Validity check. We evaluate the input of the targeted community of users in order to make decisions regarding the relevance of terms and the quality of their denomination. Our work is always open to be reviewed.

5 Conclusion and Future Work

“There have been great societies that did not use the wheel, but there have been no societies that did not tell stories” [34, p. 22]. If we were to make a comprehensive analysis of how these stories have been built in the European context, we would get a very complex network of relations between the multiple literary traditions developed by the different linguistic communities. As stated by Even-Zohar [38], there are not any European literatures which have not leaned heavily on some other literature [38, p. 48]. As a result, even outside of Comparative Literature, any type of approach that engages with the cultural analysis of a community demands the study of cultural heritages other than the one under study.

Although we can agree on the existing connections among all European literary traditions, and even though the digitalization of cultural heritages facilitates the exploration and retrieval of information, the lack of standards to define the mechanisms employed in the creation of a literary work is a hindrance for complex comparative studies; a hindrance that we aim to surmount, as it has been delineated in this paper. Our solution entails the construction of the required means for literary data to be published as LOD. Due to the relation of this proposal with the POSTDATA (Poetry Standardization and Linked Open Data) project, our object of analysis is data related to European poetics.

The foundation of the POSTDATA work is the development of a MAP, which depends on the definition of a domain model, that is, the common conceptual model that represents the informational needs of the EP community. In order to elicit these informational needs, the authors studied a representative sample of existing resources and consult the EP community as a whole through a survey. In addition, the work undertaken by POSTDATA enables the analysis of poetry within a broader and more complete context. The standardization process overcomes any linguistic barriers and it collocates the cultural products of minoritized communities with major traditions.

After defining the domain model, future work will entail the definition of the RDF vocabulary terms that best describe the concepts of the domain model and the development of vocabulary encoding schemes to constrain certain terms of the model in order to further enhance interoperability. The enrichment of our proposal with this last process will be a major step as regards the standardization of terms in the philological field. It will thus open new lines of inquiry on Literary and Cultural studies and it will enhance the existing ones in order to gain a better understanding of Western cultures.

Acknowledgments. The authors would like to thank the researchers in charge of the analyzed repertoires for their availability and willingness to share information and to discuss issues related to their projects with the POSTDATA team. Mariana Curado Malta thanks the Polytechnic of Oporto for the 3-year leave that granted her the possibility to work in POSTDATA, a wonderful and challenging professional experience in Madrid-UNED. Research for this paper has been achieved thanks to the Starting Grant research project *Poetry Standardization and Linked Open Data: POSTDATA* (ERC-2015-STG-679528), funded by European Research Council (ERC) under the research and innovation program Horizon2020 of the European Union (<http://postdata.linhd.es>).

References

1. Bod, R.: How a new field could help save the humanities. <http://www.chronicle.com/article/How-a-New-Field-Could-Help/239209/>
2. Gonzalez-Blanco, E.: Actualidad de las Humanidades Digitales y un ejemplo de ensamblaje poético en la red: ReMetCa. Cuadernos Hispanoamericanos **761**, 53–67 (2013)
3. González-Blanco, E., Selaf, L.: Megarep: A comprehensive research tool in medieval and renaissance poetic and metrical repertoires. In: Humanitats a la xarxa: món medieval/Humanities on the Web: The Medieval World, pp. 321–332. Peter Lang (2014)
4. Literary Articles: Literary tradition (2012). <https://literacle.com/literary-tradition>
5. Gonzalez-Blanco, E., del Rio Riande, G., Martínez Cantón, C.: Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires. In: Proceedings of the LREC (Tenth International Conference on Language Resources and Evaluation) 2016 Workshop “LDL 2016–5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources” (2016)
6. Naetebus, G.: Die nicht-lyrischen Strophenformen des Altfranzösischen: ein Verzeichnis. Druck von J.B. Hirschfeld, Leipzig (1891)
7. Seláf, L.: Le Nouveau Naetebus. Poèmes strophiques non-lyriques en français des origines jusqu’à 1400. <http://nouveaunaetebus.elte.hu/>
8. Horváth, I.: Répertoire de la poésie hongroise ancienne: Manuel de correction d’erreurs dans la base de données. Editions du Nouvel Objet (1992)
9. Horváth, I.: Répertoire de la poésie hongroise ancienne. <http://rpha.elte.hu/>
10. Centro Ramón Piñeiro para a Investigación en Humanidades: MedDB - Base de datos da Lírica Profana Galego-Portuguesa. www.cirp.es/meddb
11. Tavani, G.: Repertorio metrico della lirica galego-portoghese. Edizioni dell’Ateneo, Roma (1967)
12. Pillet, A., Carstens, H.: Bibliographie der Troubadours. M. Niemeyer, Halle (Saale) (1933)
13. Asperti, S., De Nigro, L.: Bibliografia Elettronica dei Trovatori. http://www.bedt.it/BEdT_04.25/index.aspx
14. Gonzalez-Blanco, E.: ReMetCa. <http://www.remetca.uned.es>
15. Antonelli, R.: Repertorio metrico della Scuola poetica siciliana. Centro di Studi Filologici e Linguistici Siciliani (1984)
16. Pagnotta, L.: Repertorio metrico della ballata italiana: secoli XIII e XIV. R. Ricciardi (1995)
17. Solimena, A.: Repertorio metrico dei poeti siculo-toscani. Centro di Studi Filologici e Linguistici Siciliani (2000)

18. Gorni, G.: *Repertorio metrico della canzone italiana dalle origini al Cinquecento (REMCI)*. Franco Cesati, Firenze (2008)
19. Betti, M.P.: *Diciassettesimo Repertorio metrico delle Cantigas de santa Maria di Alfonso X di Castiglia*. Pacini Editore, Ospedaletto (Pisa) (2005)
20. Parramon i Blasco, J.: *Repertori mètric de la poesia catalana medieval*. Publicacions de l'Abadia de Montserrat, Barcelona (1992)
21. Gómez-Bravo, A.M.: *Repertorio métrico de la poesía cancioneril castellana del siglo XV*. University of California, Berkeley (1991)
22. Escribano, J.J., Gonzalez-Blanco, E., del Rio Riande, G.: *PoeMetCa—Recursos digitales para el estudio de la Poesía Medieval Castellana* (2016). <http://poemetca.linhd.es/>
23. Riley, J.: *Understanding Metadata: What is Metadata, and What is it For?: A Primer*. NISO Press, Bethesda (2017)
24. Heath, T., Bizer, C.: *Linked data: evolving the web into a global data space*. *Synth. Lect. Semant. Web Theor. Technol.* **1**, 1–136 (2011)
25. Berners-Lee, T.: *Linked data - design issues*. <https://www.w3.org/DesignIssues/LinkedData.html>
26. Nilsson, M., Baker, T., Johnston, P.: *Interoperability levels for Dublin core metadata*. <http://dublincore.org/documents/interoperability-levels/>
27. Coyle, K., Baker, T.: *Guidelines for Dublin core application profiles*. <http://dublincore.org/documents/profile-guidelines/>
28. Curado Malta, M., González-Blanco, E., Martínez Cantón, C., del Rio Riande, G.: *Digital repertoires of poetry metrics: towards a linked open data ecosystem*. In: *Proceedings of the First Workshop on Digital Humanities and Digital Curation co-located with the 10th Conference on Metadata and Semantics Research, MTSR 2016, CEUR Workshop Proceedings*, pp. 1–11 (2016)
29. Curado Malta, M., Baptista, A.A.: *Me4MAP V1.0: a method for the development of metadata application profiles*. Submitted for publication (n.d.)
30. Curado Malta, M., Centenera, P., Gonzalez-Blanco, E.: *Using reverse engineering to define a domain model: the case of the development of a metadata application profile for European poetry*. In: *Developing Metadata Application Profiles*, pp. 146–180. IGI Global (2017)
31. DCMI Usage Board: *DCMI metadata terms*. <http://dublincore.org/documents/dcmi-terms/>
32. IFLA: *Functional requirements for bibliographic records* (2009). http://archive.ifla.org/VII/s13/frbr/frbr_current.toc.htm
33. TEI Consortium (eds.): *TEI P5: guidelines for electronic text encoding and interchange*. <http://www.tei-c.org/Guidelines/P5/>
34. Guin, U.K.L.: *The Language of the Night: Essays on Fantasy and Science Fiction*. HarperCollins, New York (1992)
35. Herzog, G., Heid, U., Trippel, T., Bański, P., Romary, L., Schmidt, T., Witt, A., Eckart, K.: *Recent initiatives towards new standards for language resources*. In: *Presented at the International Conference of the German Society for Computational Linguistics and Language Technology*, 30 September 2015
36. Baldick, C.: *The Oxford Dictionary of Literary Terms*. Oxford University Press, Oxford (2015)
37. Abrams, M.H.: *A Glossary of Literary Terms*. Harcourt Brace College Publishers, New York (1999)
38. Even-Zohar, I.: *Papers in Historical Poetics*. Porter Institute for Poetics and Semiotics, Tel Aviv (1978)