The *eLexicon Mediae et Infimae Latinitatis Polonorum.* The Electronic Dictionary of Polish Medieval Latin

Krzysztof Nowak
Institute of the Polish Language, Polish Academy of Sciences
krzysztofn@ijp-pan.krakow.pl

Abstract

The paper presents goals, methods, and results of the project of the Electronic Dictionary of Polish Medieval Latin. First, a brief history of the paper dictionary, as well as an account of its main features are presented. Second, the main problems of the metalexicographic analysis and the subsequent XML encoding of the lexicographic content are discussed. The main purpose of both being a fine-grained description of linguistic resource, it was necessary to make explicit a fair amount of data which are coded only by means of convention. Third, the web interface of the dictionary is treated in more detail. Its most important of its design principles include separation of the expert and novice user perspective, system of aids and suggestions, integration with external sources.
Keywords: electronic lexicography; Medieval Latin; dictionary interface; TEI XML encoding; implicit information; lexicographic convention

1 Introduction

The *eLexicon Mediae et Infimae Latinitatis Polonorum* (henceforth referred as the *eLexicon*) is an electronic dictionary based on the first 7 volumes[1] of the paper *Lexicon Mediae et Infimae Polonorum* (henceforth the *Lexicon*) which has been published since 1953 under the auspices and with the financial support of the Polish Academy of Sciences (Plezia, Weyssenhoff-Brożkowa, Rzepiela 1953). The *Lexicon* was conceived by its first editor, the Polish eminent philologist Prof. Marian Plezia, as a work which would fully document the use of the Latin language on the Polish territory between the Xth and the mid-XVIth century (Plezia 1958). As such, it was meant to form a part of the European network of the national dictionaries of Medieval Latin which started to emerge at the same time in response to an appeal of the *Union Académique Internationale* (Bautier 1981: 433–436). Users to which the print *Lexicon* has been addressed are in particular members of a research community, which is the reason why so much emphasis has been put, among others, on the completeness of the source material included. The print dictionary provides, then, in-depth etymological, morphosyntactic and semantic description of each

1 They include entries from A to Q, which is ca. 6000 pages printed in two columns.

word attested in the Polish Latin during the Middle Ages. Sense definitions are formulated both in Polish and – with foreign readers in mind – in Latin, and are illustrated with appropriate source quotations, if the meaning was not known in the Antiquity. The audience of the *Lexicon* being scholarly community, it becomes partially clear why the *Lexicon* does not make much concessions as far as user friendliness is concerned, with its heavy use of the typographic conventions, tightly printed columns etc. The paper dictionary suffers also from the drawbacks symptomatic for every long-term academic publishing enterprise, and in particular from several inconsistencies of the editorial policy, which affect especially usage labelling system, semantic change description or sense nesting practice, to name only few.

The project of the electronic dictionary which would be based on the *Lexicon* was conceived by the author of the paper and has been carried out between mid-2010 and mid-2014 by the team of the Department of the Medieval Latin of the Institute of the Polish Language (Polish Academy of Sciences) in Kraków.[2] Regardless of its roots, from the beginning the *eLexicon* was expected to become a research tool on its own and not merely a digitized version of the paper work. Firstly, its content was to differ to various extent from what can be found in the print volumes. One source of substantial modifications was the incorporation into the main text of the *addenda et corrigenda*, 'supplements and corrections', printed at the end of each of the 7 volumes of paper dictionary. Another one was both manual and automatic update of the lexicographic content. The members of the team (and, in the same time, current authors of the paper dictionary) had to eliminate most obvious errors and, where only it was necessary, to adjust the text to the modern editorial rules.

Secondly, the *eLexicon* was expected to provide research community with capabilities that the print dictionary could not offer. Apart from the simple search and browse features, the on-line dictionary was meant to offer access to the wealth of information encapsulated either explicitly or implicitly in the dictionary entries. At the same time, the *eLexicon* was conceived as a constituent of a larger text analysis framework. It had not only to be integrated with the digital library of the scanned images of paper slips, but also to be actively linked to the bibliography list of the medieval sources and to constitute a *sui generis* wrapper around the Medieval Latin corpus.[3] Moreover, the on-line dictionary was to incorporate a fair amount of external resources, whether it be through locally triggered queries or by means of outward linking.

Finally, the *eLexicon* has been planned as an open-access and open-source project. From the beginning the access to the web service was meant to be free and unlimited, as was also the case of the XML annotated dictionary files, which are to be distributed under liberal licenses. Although there were many reasons to do so, the main of them was assuring the longevity of the project, a major challenge in academic projects with time-limited funding. The other factor expected to contribute to project's longev-

2 Its funding was provided by a grant of the Polish National Science Centre awarded to the chief-editor of the paper dictionary, Prof. Michał Rzepiela.

3 A 5 million words, balanced and representative corpus of the Polish Medieval Latin is now being developed by the same team and is due to be delivered by the end of the 2016.

ity was compliance with standards. Developed firstly as a set of the TEI-conformant files (TEI Consortium 2013), the electronic dictionary allows platform-independent implementations, the fact which implies two major consequences. On the one hand, one can benefit from the open-source technologies and existing text or data retrieval frameworks. On the other, one may hope that the available lexicographic data will be incorporated in other research contexts, integrated with NLP infrastructures, and, consequently, they will be steadily ameliorated and refined, even when the project itself comes to an end.

2 Methods

No phase of the e-dictionary creating was outsourced. After the volumes of the print dictionary had been scanned, the image pre-processing and OCR process began as a result of which machine-readable text was, firstly, obtained and, then, carefully proofread. After that metalexicographic analysis followed, its aim being twofold. First of all, it was expected to reveal the features of the print dictionary macro- and micro-structure to be retained in the *eLexicon*, but also to select lexicographic information worth retrieving by means of the on-line search interface. Contrary to what one might believe, first part of the analysis was far from trivial, since it was often equivalent to questioning the very foundation of the paper dictionary methodology and, at the same time, to designing principles of the future on-line dictionary. The main issues addressed included internal reference system, approach to the entries with deeply nested structure, status of idioms and multi-word expressions as lexical units etc. In what concerns lexicographic data, the guiding principle was to retrieve and make explicit as much linguistic and non-linguistic information as possible, since from the very beginning it was clear that the on-line dictionary should serve researchers of various expertise in medieval studies, from the historians working on the Medieval Latin sources, to the Latin and Polish linguists, to the historians of literature, art, philosophy and science. What is more, one of the goals of the *eLexicon* was also expanding the audience of its paper predecessor beyond the scholarly world to embrace students and teachers of Latin.[4] In order, then, to satisfy needs of the academic users[5], on the one hand, and to effectively distinguish between expert and lay users on the level of the web interface, on the other hand, a highly structured resource had to be created.

Secondly, the lexicographic analysis served also two other purposes, the first of them being to estimate the feasibility of the data annotation within project's time limits, that is without resorting to advanced NLP methods, the second – to conceptualize the dictionary macro- and microstructure by means of the TEI XML tagset. Although encoding standards in linguistic annotation constitute nowa-

4 Not only Medieval, but also Classical Latin, since there does not exist as yet any on-line Polish dictionary of Classical Latin, at least academic one.

5 Or what was believed to be their needs, since to my knowledge there do not exist any empirical studies of the needs of the users of (academic) Latin dictionaries.

days a topic on their own (Garside, Leech, McEnery 1997; Pustejovsky, Stubbs 2013), I will limit myself to indicating three main reasons why the TEI XML has been chosen as an output format of the dictionary files. First of them has been already mentioned: storing lexicographic data in text files (contrary to binary ones) makes them at least partially immune to platform or software-related issues. XML encoded resources are human-readable, so they may be easily subject to modifications, adaptations and further refinement even by less technical-oriented users. Secondly, the TEI XML encoding serves well the purpose of storing highly structured, paper-born documents. Since the print dictionary being a starting point of the *eLexicon* is a result of 60 years' work, not only is it far from unified, but it also makes heavy use of sense nesting, *ad hoc* usage hints etc., all of which makes putting it into database format a non-trivial task. Thirdly, the use of widely supported formats and standards becomes essential, if one wishes to benefit from the already existing software solutions, on the one hand, and, on the other, to make one's data useful in yet unpredicted research environments. As to the former, there was no intention to create from scratch a proprietary interface to serve the dictionary content. In fact, the *eLexicon*, rather than being a closed interface solution, attempts to initiate discussion about what tool do the medievalists need and to dynamically change as the community will express its expectations. In that it differs significantly from now outdated in their design, closed-source and paid resources, such as the *Database of Latin Dictionaries* published by Brepols[6]. The availability of the XML annotated files, in turn, should encourage dictionary content reuse, whereas applying the TEI recommendations should facilitate data exchange, as, despite their drawbacks, they were generally adapted in other open Medieval Latin dictionary projects, such as precursory digitisations of Lewis and Short's *A Latin-English Lexicon* (1879)[7] by the Perseus Project team (Crane, Seales, Terras 2009; Bamman, Crane 2009), DuCange's *Glossarium* by Ecole Nationale des Chartes (Glorieux, Thuillier 2010),[8] *Novum Glossarium* in frame of the project Omnia (Bon 2009; Bon 2010; Bon 2011).[9]

Once the analysis had come to an end, the annotation guide was created and the annotation itself started. After the XML files had been generated through PERL and XSLT processing of the OCR output, they were next distributed among project's team members who diligently proofread them, modified when necessary the dictionary content and adjusted automatic encoding. Verified for their well-formedness, the files were next validated with a previously generated *Document Type Definition* (DTD). The web interface of the *eLexicon*, which will be treated in more detail below, was built around the eXist-db, a free and open-source, XML native, no-SQL database running in the back-end. Programmed as a set of XQuery scripts, it produces on the front-end a light-weight HTML5+CSS web ap-

6 http://www.brepolis.net/.

7 http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0059, now consultable also within Perseus under Philologic Project (http://perseus.uchicago.edu/Reference/lewisandshort.html) or, in a more convenient way, within Logeion project (http://logeion.uchicago.edu/).

8 http://ducange.enc.sorbonne.fr/.

9 The TEI was also employed in other historical lexicography projects, such as the Anglo-Norman Hub (http://www.anglo-norman.net/,). For details, see (Trotter 2011).

plication. In order to introduce some elements of interactivity, as well as to provide users with instant headword suggestions and similar features, a moderate amount of jQuery scripting has been added.

3 Results

3.1 The XML Annotation

The principal rule applied in the process of the XML encoding was, as was already mentioned, to make as much lexicographic information explicit as possible without caring so much about the typographic peculiarities of the print dictionary. Formatting information was generally "translated" into appropriate "semantic" annotation and retained only if it could add to the on-line dictionary features. No special effort, then, was made to save indentation or font, since they should be otherwise easily deducible from the semantic encoding. Page and line numbers have been preserved in order to ensure correct resolution of the cross-references and, thus, successful intra-linking[10].

Each of the output files corresponds to one print volume and is preceded by a metadata header in which basic bibliographic information was recorded. There is, however, nothing that could prevent prospective users from decomposing original files according to their specific needs. As far as the annotation design is concerned, dictionary entries have been first separated from each other[11]. For the purpose of subsequent processing, unambiguous identification of the entries had to be assured by means of the automatically generated identifiers, however, the lemma-based identification has been also retained[12]. The entries were next classified[13], so as to distinguish between standard and reference entries of the type:

- "LETANIA *cf.* LITANIA", where one of the orthographic variants points to the canonical word form;
- "LETARG ... *cf.* LETHARG ...", where a word fragment (most frequently word prefix) points to the position in the dictionary text rather than to a precise headword.

Since the entry access in the *eLexicon* was meant to be subject to major redefinition, further refinements needed to be applied to the selected "secondary headwords" (Atkins, Rundell 2008: 235-236), with the most significant example being derived forms. Although such forms as n. *laureus* 'laurel' which is to be found in the paper dictionary as a sub-entry of the adj. *laureus* 'laurel', remain embedded in their respective superordinate entries, they will also function as separate lexical units during

10 In the *Lexicon* cross-references come generally in two forms and may point either 1) to a precise entry or one of senses (e.g. *„Cf.* LATIO *II*" under LEGISLATIO), or 2) to a volume, page and line(s) of the dictionary (e.g. *„cf. supra I 1076,49 sqq."* under LEX which may be rendered as 'cp. above, [volume] I, [page] 1076, [line] 49 and foll[owing]').

11 For this purpose <entryFree> tag („unstructured entry") was used which allows for a more liberal encoding of the paper-born and, thus, text-oriented dictionaries.

12 Here, the attributes @xml:id ("identifier") and @n ("number") were used.

13 By means of the @type attribute, with e.g. homonyms labelled as @type="hom".

alphabetical browsing or when listed in results lists.[14] The same can be said about other instances of secondary entries, such as multi-word or idiomatic expressions.

Orthographic, etymological and morphosyntactic information was subject to diligent, fine-grained encoding. In spite of the privileged position that a headword occupies in traditional lexicography, careful annotation of the variant orthographic forms is essential for a Medieval Latin dictionary to be a serious tool of research, and that from many reasons. With orthography changing often within one manuscript from scribe to scribe, a lexicographer can never know which word form dictionary user may be looking after, which makes selecting unique, "canonical" form somewhat anachronistic, if one takes into account the medieval sense of language correctness.[15] What is more, in the print *Lexicon* one can find headwords which serve only a purpose of identifying entries and they have been never attested in medieval texts. This is the case of such entries as "[LAVANDA] *s.* LAVENDA" 'lavender' , where the square bracket is employed to indicate that, unlike *lavenda*, the form *lavanda* does not occur in sources, but, in turn, was used by the lexicographer only as a conventional representation of the entry as an ideal, in his or her opinion, reconstruction of the Italian *lavanda*.[16] In a case like the one just mentioned, separation of the genuine linguistic material from what is only a pure convention becomes crucial.

Without going into unnecessary detail, let it suffice to say that the other elements of the grammatical description of the headwords were subject to likely minute encoding, which, apart from its direct and obvious goal, namely that of description of linguistic resource, had two secondary objectives. First of them was making implicitly coded lexicographic data fully explicit. The variety of the information which is tacitly conveyed in the *Lexicon* is striking, however, what may be only challenging for a human reader, if she is not accustomed enough to dictionary convention, makes a good deal of data inaccessible for automatic processing. Among those pieces of information which could be lost, if they were not scrupulously deduced from sometimes cryptic metalanguage and, then, redundantly added to the original files, one can mention part-of-speech labelling, which is explicit (that is, expressed with appropriate labels) for adverbs or sub-headwords,[17] but for verbs, nouns and adjectives is to be inferred from the inflectional information.[18] The same is true about the language from which the headword originated, since appropriate labels are in the print dictionary employed uniquely for languages other than Latin, so, for instance, while the entry LEXICON 'a dictionary' includes a self-explanatory etymology "*Gr. λεξικόν*" (where *Gr.* stands for "Greek"), in the entry LICENTIO 'to give a license' one

14 Here <re>, ie. „related entry", tag was used.

15 The problem of the abundance of word forms is even more striking for Medieval Latin as was used in France or Spain, where it experienced substantial assimilation to a vernacular language.

16 *S.* is here abbreviation for the Lat. *siue* 'or'. Such notation can be found on a regular basis when hypothetical Classical Latin form of the word is reconstructed, see, for example, „[RHINOCERON] *s.* RINOCERON" 'rhinoceros' .

17 See, for example, LICENTIOSE *adv.* 'violently'.

18 See, for example, entries for a verb LICENTIO, -are, -avi, -atum 'to give a licence', a noun LICENTIA, -ae *f.* 'licence' or an adjective LICENTIOSUS, -a, -um 'licentious', for which PoS information should be determined from inflectional description.

finds notation "licentia", from which one should infer that the word was coined during the Middle Ages from a Classical or Medieval Latin term. To make things even less transparent, the entries like LEX have no etymology at all, which, in turn, means they were inherited from the Classical Latin.

The list of information types which are encoded only indirectly is, naturally, far from complete and should also include such important features of historical lexicography as time and place of word's attestation. While geographical information is never explicitly given in the paper *Lexicon*, chronological data are provided for the sake of precision, that is only if the source quotation comes from a work which includes multifarious, chronologically diverse material.[19] Otherwise, spatio-temporal characteristics of the quotation should be deduced from the alphabetical list of the dictionary sources. Yet, there are many reasons why information of this sort should be accessible in the on-line dictionary and, thus, why it should be also explicitly declared in the XML files. The reason that comes first to mind is, obviously, more efficient and straightforward retrieval of these data within the search interface of the on-line dictionary. The other reason why aforementioned, but also e.g. genological properties of source quotations should be explicitly encoded is that it could greatly facilitate its interactive representation in form of maps, timelines or charts, as the example of the Wiki Lexicographica (Bon, Nowak 2013) demonstrated.

The other of the secondary goals of dictionary encoding was standardisation of the lexicographic description. This included, for instance, eliminating domain or usage labels which are now obsolete or were coined *ad hoc* at some point of the dictionary writing process and shortly after fell from use.[20] On the contrary, some of the subtle or nowadays less useful distinctions were subsumed on the encoding level under general or more frequently used ones. This was the case of the labels indicating direction of the semantic change. Thanks to their unification, one will get an access to words which experienced metaphorical extension, although they were originally marked in the paper dictionary either with the standard label *metaph.* standing for *metaphorice* 'metaphorically', or with a more verbose label, *in imagine* 'in the image of'.

It should be also added that the XML encoding, apart from its obvious, data-oriented objectives, has many practical, user-oriented ramifications. In the print dictionary, for instance, sense definitions are given, as was already noticed, both in Polish and Latin. Clear separation of the definition strings not only allows their subsequent retrieval and reuse, but also, on more practical level, allows Polish users to consult the on-line dictionary in their mother tongue, while serving foreign researcher with a Latin version of the entry. Fine-grained linguistic data encoding, in turn, facilitates differentiating basic

19 For example, the only quotation which can be found under the headword LEXICON is labelled as „AKap p. 61 (a. 1540)", where „AKap" is a source identifier (pointing to a multifarious collection of the chapter tribunal), „p." stands for *pagina* 'page' and „a. 1540" is a chronological hint which should be resolved as „anno 1540", ie. 'in the year 1540'.

20 Naturally, XML encoding allowed for more obvious ameliorations as well. It was possible, for example, to introduce explicit distinction between domain (e.g. *astr.* for 'astrology', *eccl.* 'ecclesiastical term') and attitude (eg. *in malam partem* 'pejorative') labels on the one hand and the syntax markers (such as *intrans.* for 'intransitive' or *refl.* for 'reflexive') on the other.

and advanced user scenarios, and enables adapting lexicographic content perspective to the varying user needs.

3.2 The Web Interface

Apart from the obvious goal of overcoming the well-known drawbacks of paper dictionaries, the web interface of the *eLexicon* was created in order to facilitate advanced retrieval of the data obtained in the process described above. Thus, expected to constitute the main entry point to the electronic dictionary and other tools of textual studies, it was meant to provide professional users with a fully-fledged research platform. At the same time, however paradoxical it may appear, it had to satisfy the needs of less-advanced users, students and language teachers, by clearly separating basic and advanced perspective on lexicographic content. In order to serve well both groups, namely that of expert, as well as that of novice users, the guiding principle of the web interface creation became to help users better understand what they are looking for and to produce meaningful output, even if the phrase the user looked for, was not found in the dictionary.

When visiting the *eLexicon* page for the first time, users are proposed a quick tour of the search and browse features the dictionary offers. The main page is not meant, however, to overwhelm a visitor with a plethora of options (Figure 1). Rather the contrary is true, since apart from the simple menu, which gives direct access to the search and browse interface, it does not display anything but a simple search form which is, though, an actual entry point to the dictionary content.

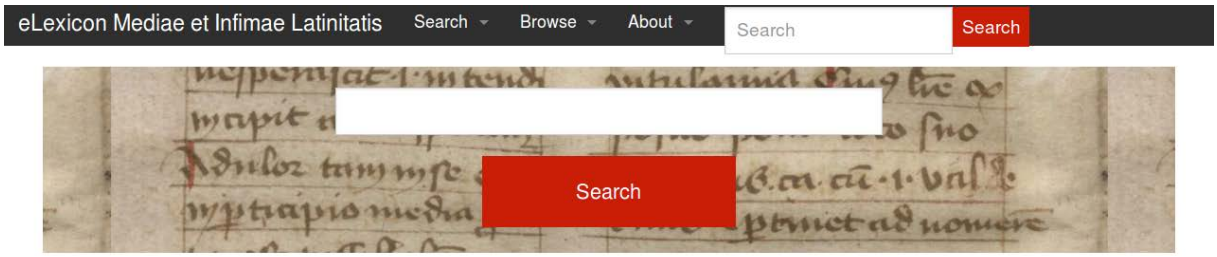

Figure 1: The on-line dictionary: the main page.

Its underlying logic is to support two expected use scenarios:

- a user is looking for a lemmatised word form, for which there exists a corresponding main or secondary headword;
- a user is looking either a) for an inflected form, a Polish or other non-Latin term, a Latin word which is not attested in Polish sources, or b) for an incorrect word form.

The first scenario, i.e. successful lookup of a headword included in the *eLexicon*, is promoted by means of the Ajax-based suggestion list which appears once the user types in three first letters of the phrase

she is looking for.[21] The list of the words suggested consists of all the orthographic variants of the headwords included in the dictionary, as well as of the multi-word expressions and idioms, which are, however, still distinguishable from the former thanks to their different formatting. The suggestion list not only should significantly speed up the lookup process, but also may handle potential typing errors and, since Latin is an inflected language, point user to a correct lemma of the word she is querying. Once suggested option is selected, the user is redirected to the appropriate entry.

Here, two perspectives on the dictionary content are provided as separate tabs which, once clicked, reveal, respectively, a basic and a full content view. First of them (Figure 2), under clearly separated headings presents selected morpho-syntactical properties of the word, a brief overview of its meaning, as well as various summaries of its use.[22] As such, it should aid the novice, as well as expert users to get the general idea of the word they are looking for, without necessarily overwhelming them with the full apparatus of the academic lexicography.

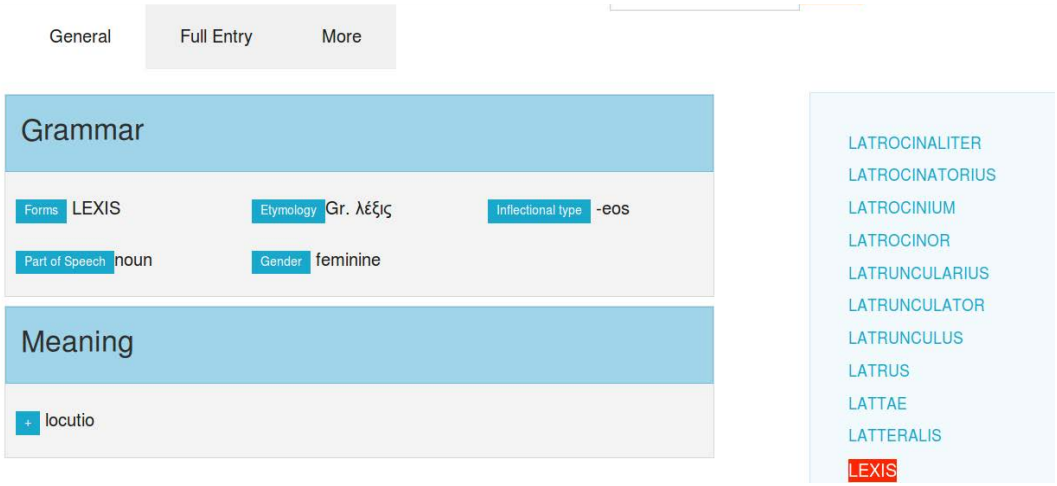

Figure 2: The on-line dictionary: a single entry ("Basic View" tab).

For both user groups the basic view may also be a convenient access point to the full version of the entry. The latter, in turn, makes heavy use of the CSS and JavaScript styling in order to improve the readability of the paper-born entry and facilitate information retrieval.

If the user does not decide to follow the suggestions and her query does not correspond directly to one of the entries, she is taken to the disambiguation page, where the second of the aforementioned scenarios is handled. User's input is being processed and searched for in the sections of the *eLexicon* different than headwords.

21 The threshold was selected as a compromise between acceptable server load and usefulness. It is certain that it will be adjusted, once the user search logs are collected and analysed.

22 It is strongly inspired by the basic view previously implemented in the WikiLexicographica (Bon, Nowak 2013).

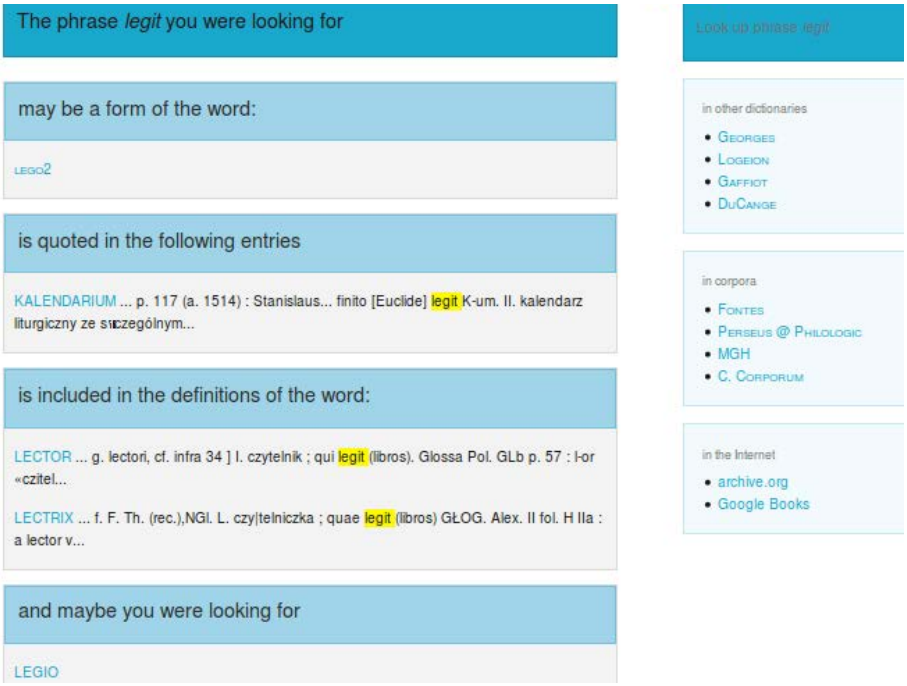

Figure 3: The on-line dictionary: disambiguation page.

The user is next presented with a result list (Figure 3) which, depending on the case, includes all or only some of the following parts:

- lemmatised form, in case if the user's input was an inflected form of a word;
- results of the word lookup in source quotations and/or definitions, if the user was using the dictionary as a source of attestations or synonyms, if she was after translation of a Polish term or if she was mimicking onomasiological search;[23]
- suggestions of similar words, if the input does not yield any meaningful result, so instead a correct word form should be suggested.[24]

The *eLexicon* content, however, may be accessed not only from the simplified main page search form, but also from the browse and expert search pages. The former functions as an equivalent of turning pages of the paper dictionary. Entries may be, then, selected by specifying the respective volume, page and line of the print edition. What is, however, more important, the browse interface offers dynamic (changing as the user is typing), synchronous lookup of the user's input at the beginning, in the middle and at the end of the dictionary headwords. This seemingly trivial feature was included to serve especially the Medieval Latin paleographers and manuscript readers in general who used to consult dictionaries looking for a reading suggestion of a hardly legible characters rather than for precise sense explanation.

23 By looking, for example, for all words of which the Latin definitions employ word *color* 'colour'.

24 Since the eXist-db makes use of the Lucene engine for text search, this feature is implemented as a fuzzy search of the user's phrase in the dictionary headwords. As the default Levenshtein distance value (Jurafsky, Martin 2009: 152) looks to be too liberal to produce helpful output, it will be certainly adjusted, once the data about the actual queries are collected.

The third access point is constituted by the advanced search facility, which takes full advantage of the scrupulous XML encoding of the lexicographic information (Figure 4)

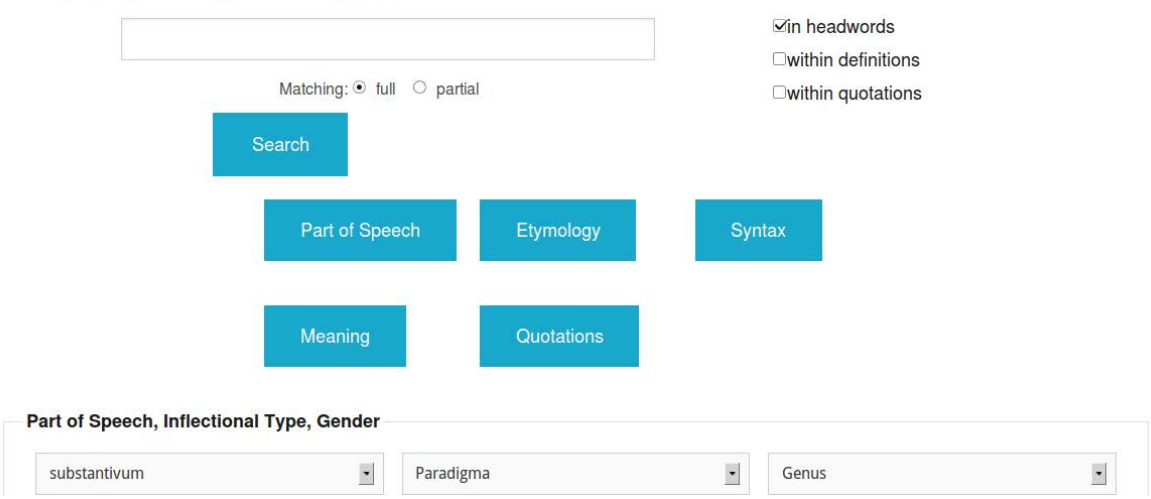

Figure 4: The on-line dictionary: an advanced search page.

The user is here provided with a search form consisting of two main parts:

- the text input field, in which query string should be typed. The user may further specify scope of her search,[25] as well as matching strategy of her choice.[26]
- the list of additional restrictions to apply to search results. The user is free to refine her query and limit results by means of the morphosyntactic, etymological, semantic and chronological criteria.[27] In case the query string is not specified, selected criteria are applied to all dictionary entries and the interface functions as a tool of the exploratory analysis of the Medieval Latin lexicon. The results page allows for further refinement, since it contains a filtering list of linguistic properties of the previously queried headwords, which behaves in a manner similar to faceted browsing widgets.

As was already mentioned above, the *eLexicon* is expected to become a centre of a fully-fledged research platform. For that purpose, it is meant to be closely bound with the corpus of the Polish Medieval Latin. Although there still remains much work to be done, even now, whenever possible, use is made of the already existing external resources that are expected to be of help for the *eLexicon's* expert users. Since, as for now, resources in question are stored externally, only appropriate links to the freely available corpora, dictionaries or on-line encylopedias may be provided to the users. In not so distant future, however, external resources are planned to be exploited locally and the content of at least some

25 That is, specific section of the dictionary entry within which the phrase should be looked for (currently options are limited to headwords, quotations and definitions).
26 That is, whether exact or approximate matching search should be applied.
27 Therefore, each query may be restricted, for instance, only to entries belonging to a specified inflectional class, originating from certain language, labelled as technical terms of a given domain or attested only in certain period.

of them will be directly embedded in the *eLexicon* search results.[28] In the current state of the interface, external resources are displayed:

- as a sidebar on the disambiguation page, as a way to suggest to the user other localisations in which she may find word absent from the *eLexicon* (see Figure 3 above);
- as a separate tab under a single entry view, so as to extend lexicographic perspective with corpus and knowledge base extracted data (Figure 5).[29]

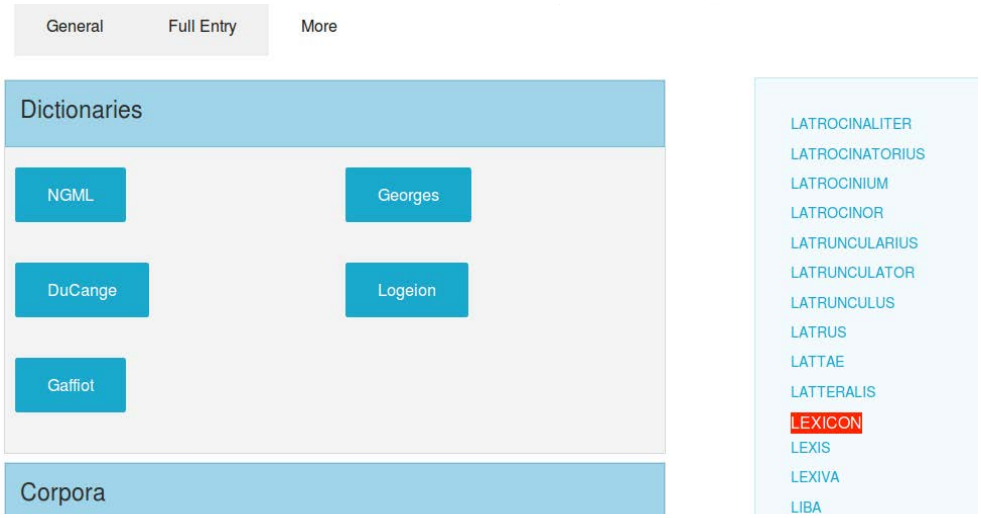

Figure 5: The on-line dictionary: a single entry ("More" tab).

4 Conclusions

The electronic dictionary of Polish Medieval Latin may become an important tool of the medieval studies, and this for many reasons. It will be freely accessible both in form of diligently encoded XML files and as a research-driven web application. To provide user with better insight into the medieval lexicon, the internet dictionary employs external sources, either by means of dynamically linking or direct embedding. Advanced users not only should find its single entry more readable, but they also will benefit from a configurable expert search and browsing interface, which provides the *a tergo*-like lookup. In turn, clear separation of a basic and advanced perspective on lexicographic content, as well as the use of suggestion lists and disambiguation pages may contribute to its becoming an effective tool for the Latin language students and teachers.

At the same time, there, naturaly, still remains much room for improvement. As far as data presentation layer is concerned, maps, timelines, charts and other alternative displays need to be implement-

28 This is the case of the freely available volumes of the *Novum Glossarium Mediae Latinitatis* or the *Glossarium* of DuCange, but also of the texts collected in the Perseus Library. The similar approach has been already applied in such inspiring tools as *Logeion* (http://logeion.uchicago.edu) or *Le Dictionnaire vivant de la langue française* (http://dvlf.uchicago.edu/).

29 In its current form, it is clearly still very far from being fully implemented.

ed. There is also a serious NLP work which has to be done, since the *eLexicon* is expected to provide conceptual search interface and to better integrate with knowledge bases.

5 References

Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bamman, D., Crane, G. (2009). Computational Linguistics and Classical Lexicography. In *Digital Humanities Quarterly*, 3(1). Accessed at: http://www.digitalhumanities.org/dhq/vol/3/1/000033/000033.html [10/04/2014].

Bautier, A.-M. (1981). La lexicographie du latin médiéval. Bilan international des travaux. In La lexicographie du latin médiéval et ses rapports avec les recherches actuelles sur la civilisation du Moyen Age: Paris, 18-21 octobre 1978, Paris: CNRS, pp. 433–53.

Bon, B. (2009). OMNIA – Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins. In *BUCEMA. Bulletin du centre d'études médiévales d'Auxerre*, 13, pp. 291–92. Accessed at: http://cem.revues.org/11086 [10/04/2014].

Bon, B. (2010). OMNIA: outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (2). In *BUCEMA. Bulletin du centre d'études médiévales d'Auxerre*, 14, pp. 251–52. Accessed at: http://cem.revues.org/11566 [10/04/2014].

Bon, B. (2011). OMNIA : outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3). In *BUCEMA. Bulletin du centre d'études médiévales d'Auxerre*, 15. Accessed at: http://cem.revues.org/12015 [10/04/2014].

Bon, B., Nowak, K. (2013). Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langements, M. Tuulik (eds.) Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013. Tallinn - Ljubljana: Trojina, Institute for Applied Slovene Studies, Eesti Keele Instituut, pp. 407-420. Accessed at: http://eki.ee/elex2013/proceedings/eLex2013_28_Bon+Nowak.pdf [10/04/2014].

Crane, G., Seales, B., Terras M. (2009). Cyberinfrastructure for Classical Philology. In *Digital Humanities Quarterly*, 3 (1). Accessed at: http://www.digitalhumanities.org/dhq/vol/003/1/000023/000023.html [10/04/2014].

Garside, R., Leech, G. N., McEnery, T. eds. (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London-New York: Longman.

Glorieux, F., Thuillier, S. (2010). Grec ancien, latin médiéval, balisage comparé de deux dictionnaires, vers des ressources linguistiques. In *ALMA. Archivum Latinitatis Medii Aevi*, 68, pp. 161–81.

Jurafsky, D., Martin, J. H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River: Pearson Prentice Hall.

Lewis, Ch. T., Short, Ch. (1879). *A Latin Dictionary*. Oxford: Clarendon Press.

Plezia, M. (1958). Lexicon mediae et infimae latinitatis Polonorum. In *ALMA. Archivum Latinitatis Medii Aevi*, 28, pp. 271–84.

Plezia, M., Weyssenhoff-Brożkowa, K., Rzepiela, M. eds. (1953). *Słownik łaciny średniowiecznej w Polsce. Lexicon mediae et infimae Latinitatis Polonorum*. Vols. 1–8. Kraków: Wydawnictwo IJP PAN.

Pustejovsky, J., Stubbs, A. (2013). *Natural language annotation for machine learning*. Sebastopol, CA: O'Reilly Media.

TEI Consortium. (2013). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version: 2.5.0. Accessed at: http://www.tei-c.org/Guidelines/P5/ [10/08/2013].

Trotter, D. A. (2011). Bytes, Words, Texts: The Anglo-Norman Dictionary and Its Text-Base. In *Digital Medievalist*, 7. Accessed at: http://www.digitalmedievalist.org/journal/7/trotter/ [10/04/2014].

Acknowledgments

The *eLexicon Mediae et Infimae Latinitatis Polonorum* as well as the work on the present paper were supported by a research grant of the Polish National Science Centre (*eLexicon Mediae et Infimae Latinitatis Polonorum (A-Q)*, nr 3736/B/H03/2011/40).