Formal Representation and the Digital Humanities

Formal Representation and the Digital Humanities

Edited by

Paola Cotticelli-Kurras

and Federico Giusfredi

Cambridge
Scholars
Publishing

Formal Representation and the Digital Humanities

Edited by Paola Cotticelli-Kurras and Federico Giusfredi

This book first published 2018

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-0803-X
ISBN (13): 978-1-5275-0803-3

TABLE OF CONTENTS

INTRODUCTION

This book contains the papers that were presented at the workshop "Formal representation and digital humanities: text language and tools", organized by Paola Cotticelli-Kurras and Federico Giusfredi under the auspices of the Department of Cultures and Civilizations of the University of Verona, and with the funds of the Marie Skłodowska Curie Project "SLUW: A computer-aided study of the Luwian (morpho-)syntax", that received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 655954. The editors of this collection are grateful to Alfredo Rizza, who acted as the third member of the scientific committee of the workshop and helped selecting the papers that were accepted.

The University of Verona has traditionally shown a strong interest in the problem of the "new trends" in humanities, including the computational and digital approaches; in 2012 a volume dedicated to *Linguistica e filologia digitale: aspetti e progetti,* appeared, containing the proceedings of a conference held in 2011 and edited by P. Cotticelli-Kurras. Such conference was part of a larger set of annual events that go by the title of *Giornate di filologia digitale*, promoted and organized by Prof. Adele Cipolla. This book represents a new step in the constant process of renewal of the field of humanities in the digital era.

"Digital Humanities" is a large, umbrella-term. The word "humanities" itself is, of course, the label of a large and varied set of disciplines. All of them, from history to philology, from literature to linguistics, are facing new challenges both from the methodological point of view and from the point of view of their very scientific contents.

The new technologies are an important aspect of research. Machines are capable, in a few seconds, to compute and compare the syntactic patterns annotated in a corpus of texts, to transform bracketed strings in trees or tables, to count and list all of the occurrences of a word or regular expression, to compare images and assign metrics to similarity and distances. This certainly made the lives of the scholars easier, and made it possible to easily manage large amounts of data. Dealing with a corpus of

thousands of texts used to take years or even decades of work; now, dozens of queries can be ran in a few hours. Today, we give technology for granted, but this, of course, has not always been the case.

***

Tools are not generated by themselves, and, once they are successfully developed, it is necessary to learn how to use them. The strengths and weaknesses of the several applications of computation and technology to the study of humanities are manifold and complex. Furthermore, the "computational" and the "digital" approaches in human sciences can be very different from each other: indeed, the labels "computational" and "digital" have much in common, but they are not properly synonymic.

The computational side has much to do with *formalization*. It deals with using formal languages and tools based on formal languages in order to explore the structure of an object of study. Computational approaches, in their *formalizing* acceptation, may have little or nothing to do with technology. For instance, while computation was one of the bases, and perhaps even the one real base, of the revolution of the American post-structuralism that produced the generative study of the syntactic structures, the role of formal languages in the Chomskyan revolution turned out to have a relatively small impact on the development of modern "computational linguistics", that in the more recent years is mainly concentrated on Natural Language Processing (NLP) and concerned with problems of automatization and optimization. Still, computation acts as a criterion and tool for formalization in all generative grammars, that still belong, with full right, within the large and varied family of computational linguistics.

In other cases, computation can be the tool employed for investigating specific aspects of an object of study: scanning, measuring, annotating, browsing and querying data in order to discover something new. This is increasingly happening in linguistics, epigraphy, philology and even in fields that deal with iconography and iconology: textual and visual data can be formally represented and categorized in order to find out patterns on large scales of data, in time and space.

Finally, technologies can be employed to produce cutting-edge tools, that can be made available for further inquiry and investigation in specific fields. This is the case of "digital" collections, annotated corpora and

lexica, compiled by scholars, teams and research facilities and eventually made available to the wider scientific community: a more practical and technical approach, that, nevertheless, quite often ends up producing important scientific contributions. As one builds the corpus, it is not uncommon to discover patterns during the very process of annotation.

***

Our understanding is that all of these different nuances belong to a large and comprehensive set of issues that regard the current approach to formalization, formal investigation and formal representation of data in humanities. As a consequence, the Workshop "Formal Representation and Digital Humanities: text, language and tools" included papers ranging from the formal representation of linguistic data to the annotation of corpora, from the theory and practice of creating digital lexicographic databases to the current issues and trends in the field of NLP.

On the front of formal linguistics, Andrej Sideltsev explores the possible extension of the models of generative grammar to the Hittite language; Bernhard Koller and Hannes Fellner employ quantitative approaches to discuss the prosodic and phonological features of Tocharian; and Rosemarie Lühr presents the results of a quantitative modeling of topicality and salience in Indo-European.

As far as lexica and corpora are concerned, Ancient Anatolian linguistics and philology are strongly represented in this volume. Markus Frank and Zsolt Simon present the structure and some preliminary results of the German project eDiAna (*Digital Philological-Etymological Dictionary of the Minor Ancient Anatolian Corpus Languages*), funded by the Deutsche Forschungsgemeinschaft and hosted by the Universities of Munich and Marburg. Lorenzo Castellano, Lorenzo D'Alfonso and Annick Payne describe the preliminary results of their project for a digital study of the paleography of Anatolian Hieroglyphs, and Claudia Posani outlines a perspective digitalization of textual resources. Finally, Guglielmo Inglese's paper is somewhere in the middle between theory and practice: it deals with the technical and theoretical issue of annotating broken and fragmentary Hittite texts. Beside Anatolian, Ancient Greek corpora are also represented by Chiara Zanchi, Eleonora Sausa and Silvia Luraghi's project HoDeL (Homeric Dependency Lexicon), a dependency lexicon of Homeric Greek.

A last group of papers is dedicated to "computational linguistics" proper, with a contribution by Marco Passarotti who outlines and discusses the developments and merits of dependency grammars. Machine translation and automatic annotation of semantic information and meaning change are dealt with by the papers by Fahad Khan, Javier E. Díaz-Vera and Monica Monachini, while Eleonora M. Litta presents her tool for the morphological annotation of Latin (European Marie Skłodowska-Curie project WDL). Claudio Russo discusses the perspectives and issues of annotation of an Italian corpus available on the web, while the annotation of temporal data in a small historical corpus is the topic of the contribution by Manuela Speranza e Rachele Sprugnoli. Finally, Katja Zupan and Tomaž Erjavec explore the field of statistics-based machine translation.

***

The number of disciplines, fields and approaches that this collection of papers reunites is large. This reflects the complexity of the fields of formalization, computation and digitalization of data and resources of humanities. The future of human sciences will be marked by the ever-increasing importance of formal models and computational tools, and we believe that an effective communication among the specialists of different fields is crucial for the scientific success every discipline. By presenting this collection of cutting-edge, high-quality papers, we believe we are making a step towards a better definition of the role "Digital Humanities" will play in the next years.

Paola Cotticelli Kurras and Federico Giusfredi

PALAEOGRAPHIC DATABASE FOR THE ANATOLIAN HIEROGLYPHIC SCRIPT 1.0

LORENZO CASTELLANO, LORENZO D'ALFONSO AND ANNICK PAYNE

Abstract

The present article aims to give an overview of the state of development of a prototype database for the palaeography of the Anatolian Hieroglyphic (AH) script.

The Anatolian Hieroglyphic (AH) Script has increasingly moved into the scholarly focus following the publication of new text editions of the Iron Age inscriptions by J.D. Hawkins (2000). This corpus and following editions of further inscriptions (reference collected by Simon 2010), now provide a firm basis for the study of AH inscriptions. While historical and linguistic analysis of AH texts has grown exponentially in the last fifteen years (e.g. Melchert ed. 2003; Yakubovich 2008, Giusfredi 2010; Rieken and Yakubovich 2010; Bryce 2012; Mouton et al. 2013; Weeden 2013; Rieken 2015; Archi 2016), palaeography, with the exception of limited comments on specific sign forms, is still in its infancy. The present authors have identified the need to investigate the development of the script (d'Alfonso 2012; Payne 2015), with a view towards establishing better dating criteria for inscriptions, to study the transmission of the script through time and space and to better understand the influence on it of the socio-political turmoil following the end of the Bronze Age. This is especially relevant as the large majority of texts does not originate from secure archaeological contexts, and inscriptions do not carry dating formulae, so that efforts to date these texts rely on text-internal references to persons or events dateable with the help of outside sources, or stylistic criteria of accompanying artwork (one more area that requires substantial revision), thus carrying a substantial risk of circular argumentation.

Preliminary research by the authors has shown that an in-depth study of AH palaeography does not only provide a tool for dating inscriptions, but also yields new historical information pertaining to the situation of the localities where inscriptions were found (d'Alfonso and Payne 2016). From the beginning of the project 'The Palaeography of Anatolian Hieroglyphic Stone Inscriptions', it was envisaged that the data generated should be made available in an internet-based database. First experiments with a database prototype (AH Pal 1.0) are currently underway, and it is the aim of this article to provide an overview of the state of development of this new tool. Comments by the scholarly community are explicitly invited.

1. Project Architecture and Data Structure

The version of the Anatolian Hieroglyphics Palaeography database here presented (AHpal version 1.0) has been created through MySQL, a Relational Database Management System (RDMS). In order to facilitate the operations of data-entry, consultation, and query, we developed the website through Xataface - a user friendly data driven web application. The database is accessible also from desktop SQL applications, such as HeidiSQL (Microsoft Windows) or Sequel Pro (OS X). By using the package RMySQL, the SQL database is accessible in R studio, allowing graphic representation of the data and their statistical analysis. Because of the limitations of SQL and in general RDBMS (Batra and Tyagi, 2012) it is under development through R2RML mapping a parallel virtual RDF version of the database (graph database).

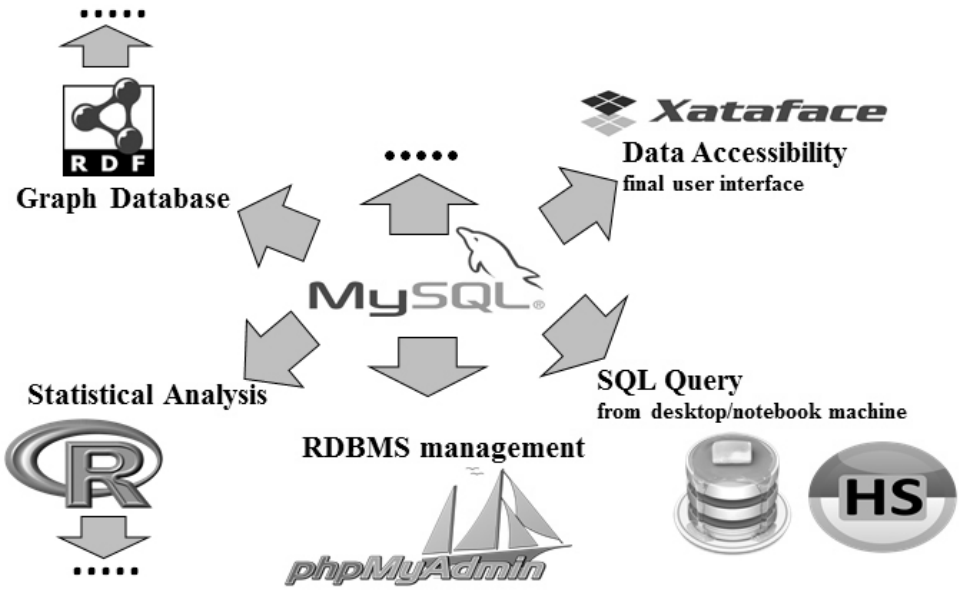

Figure 1. Project Architecture

The architecture of the relational database (figure 2) is organized around five major conceptual groups of information. The inscription metadata are at the center of the database structure, a table in which general data for each inscription, including context and chronology, are stored. Photos and drawings of inscriptions, and publication information are stored in a second group of tables. Two hierarchically structured tables (signs and sign variant tables) manage the palaeographic data; the many-to-many relationship between sign variants and inscriptions is stored in a junction table, in which also the number of attestations of sign variants in each inscription is recorded; a distinction is also made between cursive and monumental writing style (Hawkins 2003: 155).

While the database does not focus on the content of the inscriptions, personal, divine, and geographic names attested in each inscription are collected, because of their relevance in determining spatial and chronological links. Finally, the geographic context of the inscription (find spot) is managed by a site and region table.

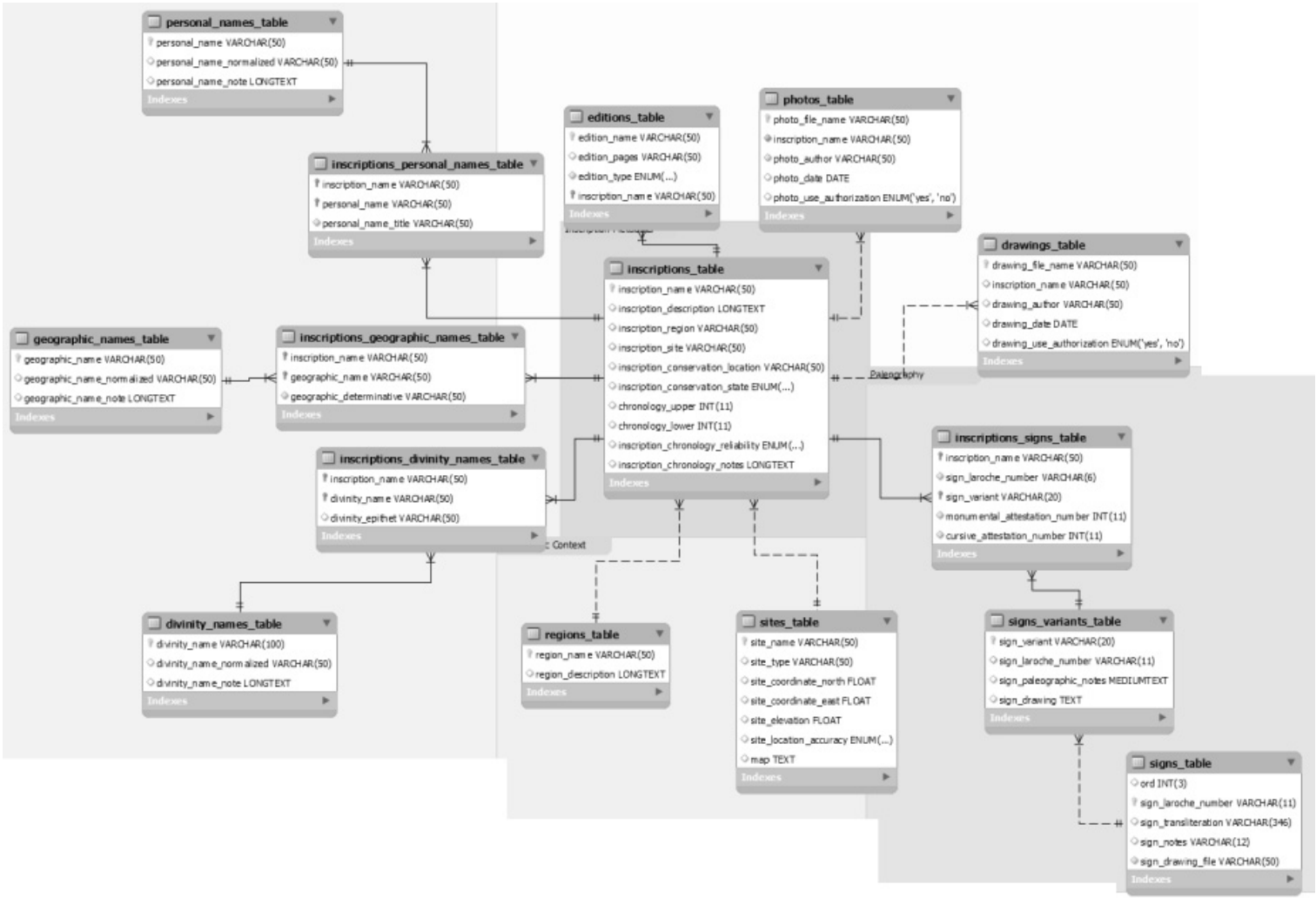

Figure 2. Preliminary EEC diagram

2. Data Display

The database allows the user to navigate between five main pages: **Inscriptions**, **Sign List**, **Palaeography**, **Corpus**, and a query page accessed by the tab **Search Variants**.

2.1. In the inscription table the following data are recorded: name of inscription (following Hawkins 2000 and the editions of inscriptions published after 2000), region of provenience of the inscription (following Hawkins 2000, as modified by d'Alfonso and Payne 2016), site of provenience (if any), and upper and lower limits of date assigned to the inscription, as well as an indication of dating reliability (high/medium/ low); also, if applicable the presence of reliefs or sculpture accompanying the inscription is recorded. More detailed information on each inscription is presented in a page which can be opened by clicking on the inscription name in this table.

2.2. The tab **Sign List** provides the user with a standard table of the AH signs listed by number and transliteration (semantographic and/or syllabic value(s)). List and value follow Marazzi 1998, Hawkins 2000 and 2003, Payne 2010 and further studies and notes on individual signs published after 2000. By clicking on a sign-number, a page with a stereotypical drawing, values and notes on the given sign opens up. Clicking on the tab 'variants', provides access to the list of all variants of the sign.

2.3. Table 1 shows how the category **Palaeography** is organized. The table is organized in four columns. The first column contains the variant number, typically consisting of the Laroche catalogue number introduced by a star sign * (following Marazzi et al. 1998 and Hawkins 2000), and linked to the number of the variant by the low dash; ex. *1_1, for the variant number 1 of the sign *1, EGO. The second column contains the (simple) sign number to which a variant belongs. The third and fourth columns provide a visualization of a drawing of the given variant in both cursive and monumental style (if attested).

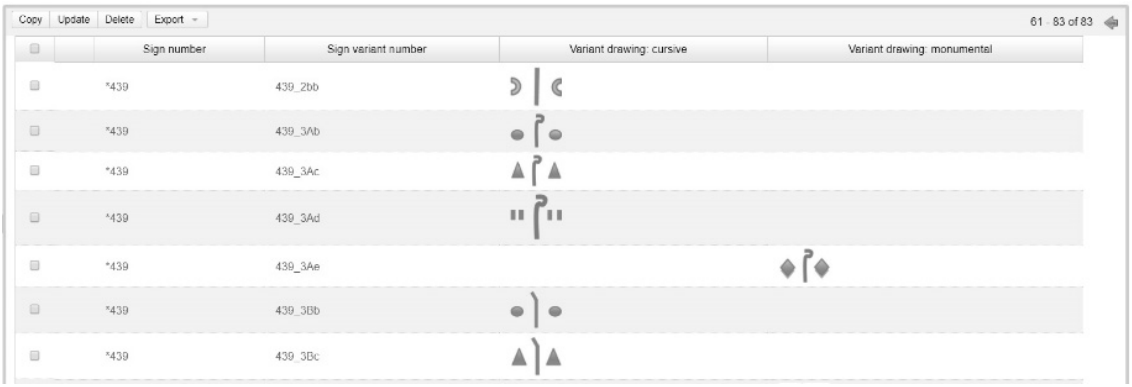

Table 1. Category 'Palaeography'

By clicking on a sign variant number in the left column, the user can access a page exclusively dedicated to that variant. The ensuing page offers a drawing, sign value(s) and notes on the variant; clicking on the tab 'attestations', will open a list of all attestations of the variant by inscription, as registered in the database. We are currently in the process of exploring several ways of tagging attestations of variants on high-definition pictures of inscriptions, so as to offer direct access to the exact realization of any attestation of a variant. That process, however, is very time-consuming, which forces us to explore different strategies in order to establish the most economical manner of data entry. We have therefore decided to dedicate the first phase of the database to collecting palaeographic information, transferring the projected visual access to sign variants to a parallel but independent line of research. Eventually, the database shall link the content of the variant page with physical representations of sign variants on high definition photographs of the inscriptions, as they become available. To achieve this aim, we co-operate with publishers, researchers and research institutions in possession of the suitable photographic materials.

2.4. The **corpus** page menu gives access to the three indices of personal, divine and geographic names recorded in the inscriptions. For each category, a transcription and a normalized form of the name is provided. Further notes to entries are provided as applicable.

3. The Process of Data Entering

Entering data into AH-pal is facilitated by a number of standardised masks, all of which are subsidiary to on the main mask for the entry of a new inscription (Tab. 2). Data entry has been conceived as a process that may generate innovative data which could spread to any and all sections of

the database, and even cause modification of data filed at some earlier stage (see below). Ideally, data recording is conceived as a process involving two active participants, a junior (first) and a senior (second) participant. The first participant exclusively fills the main mask **New Inscriptions**. She/He is asked to collect and enter all data concerning a new inscription, derived from current text edition(s) and studies. This data is divided into four main sections. The first section, **General Data** (table 2) includes the name, and if applicable alias of an inscription, a brief description of the material support, bibliography of edition(s) and type (full/partial).

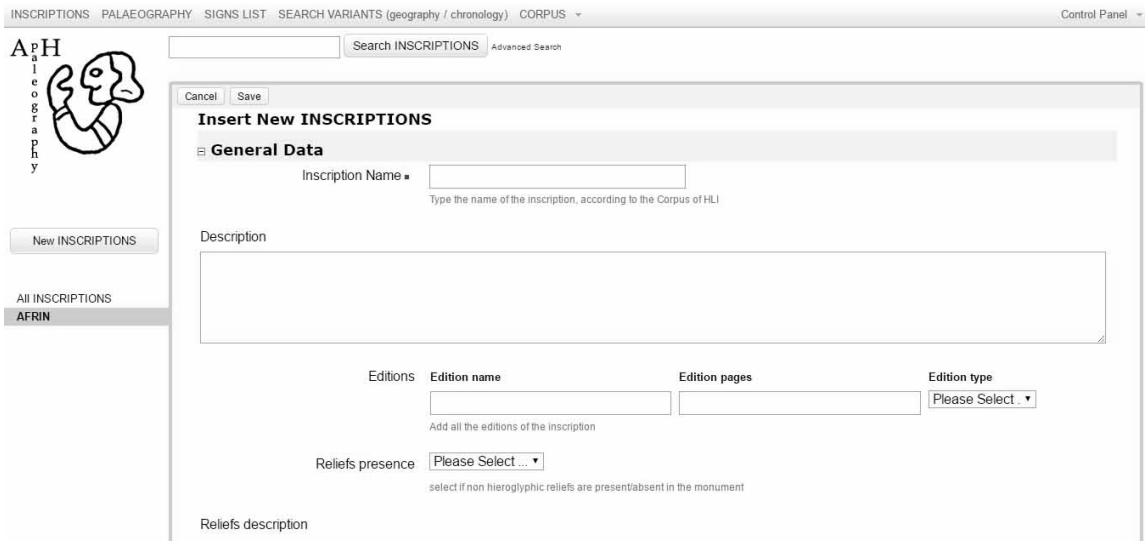

Table 2. Data Entry: General

As illustrated in table 3, the second section records the **Geographic Location** of the inscription according to region and site of provenance. The third section, **Conservation,** records the present location of the inscription, e.g. in a museum, and its state of conservation (poor/average/ good).

Table 3. Data Entry: Geographic and Conservation

The fourth section, **Chronology,** may be considered central to the project, and will be under constant review as work progresses and dating of inscriptions can be improved upon. Initial data for the chronology will be derived from existing literature, and will be used to indicate an upper and lower chronological limit, as well as an evaluation of the level of dating reliability (high/medium/low). For the chronology, it is of special importance to record various non-schematic information on which said dating rests, such as e.g. synchronism between the author of the inscription and a ruler attested in Assyrian annals; or close stylistic links to another inscription. Such is possible in the comment field, see below (table 4).

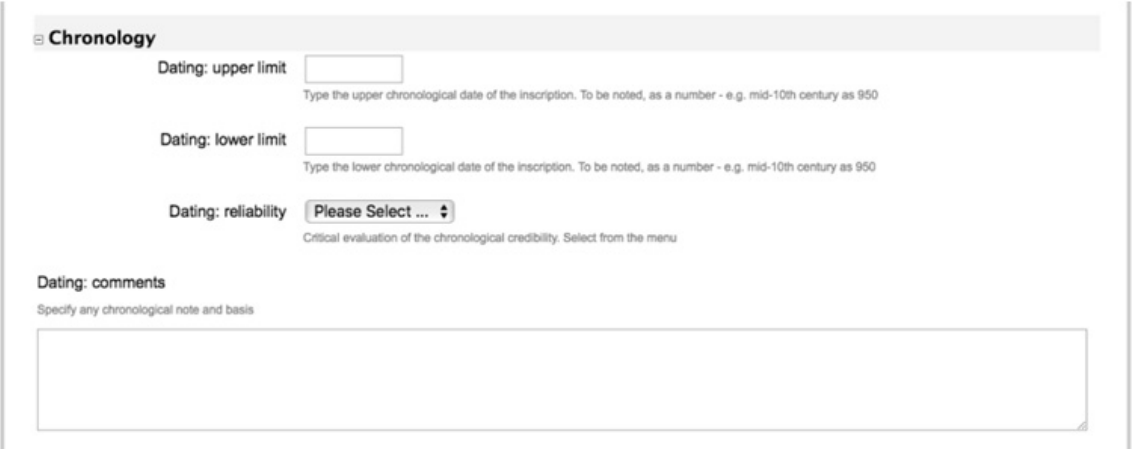

Table 4. Data Entry: Chronology

The fifth section, **Names,** records information for the searchable corpus of personal, divine and geographic names (cf. above, 1). The following sixth section requires entering images of the inscription,i.e. drawings and photographs with the respective information on provenience/authors.

Once the metadata concerning an inscription as a whole is entered, the first participant will collect the number of occurrences of each sign's variants attested in this inscription, and discuss his findings with a senior participant. The latter will check the variants and, as applicable, will assign a new number to a variant not yet present in the database. Data entry of a new variant (essentially, the creation of a new page) is the responsibility of the senior participant. Once all new variants of an inscription are defined and entered into the database, the first participant will enter occurrences of all sign variants of the new inscription in AH-Pal in the final section, **Palaeography**. This last phase of data entry also works as a check, for it is only possible to enter new attestations of variants that have already been entered into the system.

The section records sign-variants according to sign-number, followed by number of variant of that sign, and number of occurrences of each variant, divided between monumental and cursive. This completes the data entry.

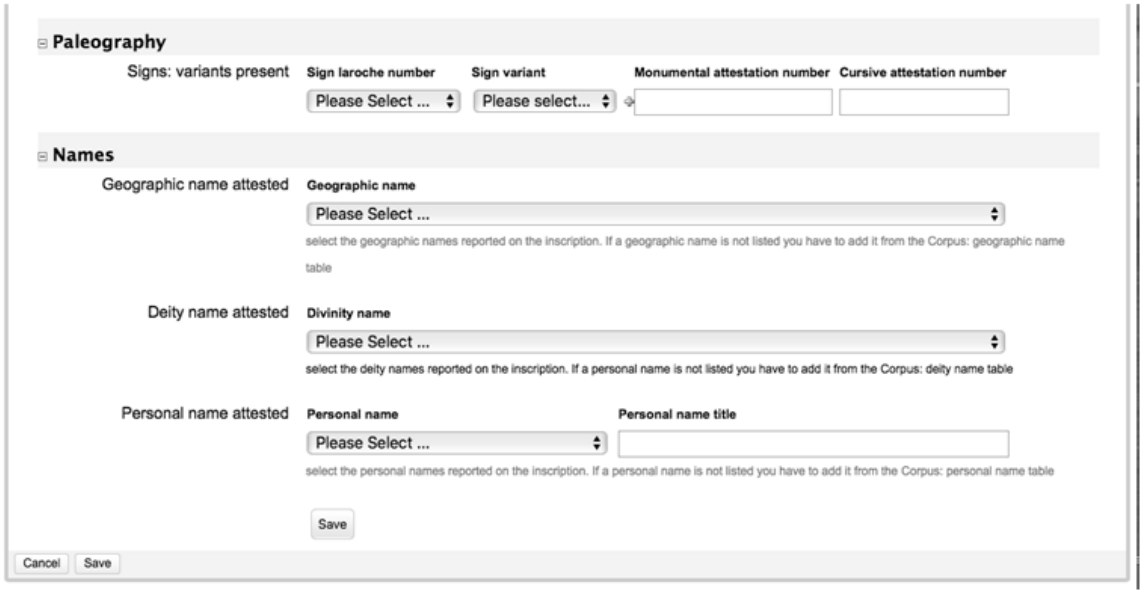

Table 5. Data Entry: Palaeography and Names

4. Data Query

The structure described in the previous section has been developed (and will be further developed) mainly on the basis of queries (‘questions’) that the database should answer. The database is aimed to record and analyse variance in the graphic rendering of the different AH signs throughout time and space. It is, therefore, crucial to correlate this variance with chronological, geographic, and textual parameters. In the frontend application, it is possible to run searches combining the above-mentioned variants. More specific and detailed queries can be run in SQL language, and in the future also in SPARQL - taking advantage of computer reasoning and inferences through a semantic query language.

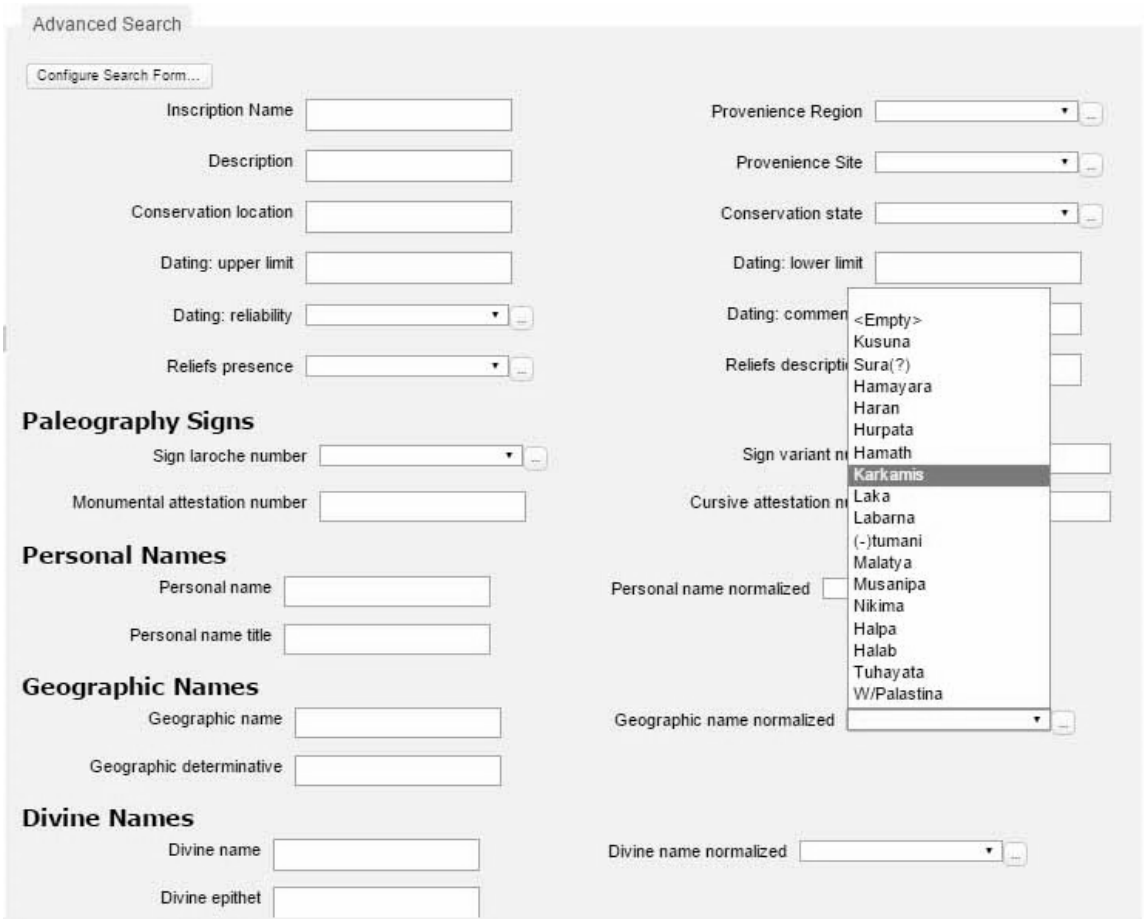

Figure 3. Example of a search form available in the fronted application

Acknowledgments

The project of AH palaeography has been developed over the past four years by L. d'Alfonso and A. Payne: its main scope is to study sign variant distribution in time (14th-8th c. BCE) and space (regions of Central and South Anatolia, Northern Levant and Syria). A prototype of an online database is currently being developed by L. Castellano and tested and implemented by N. Lovejoy, K. Justement and L. d'Alfonso at the Institute for the Study of the Ancient World, New York University. The project is kindly web-hosted by New York University. As senior particpants, L. d'Alfonso and A. Payne are responsible for checking and corrections of the data-entry.

Bibliography

Archi A., 2016, "Luwian Monumental Inscriptions and Luwians in northern Syria", in S. Velhartická (ed.), *Audias fabulas veteres.*

Anatolian Studies in Honor of Jana Součková-Siegelová, Brill, Leiden, 16-47.

Batra S. and Tyagi C., 2012, "Comparative Analysis of Relational And Graph Databases", in *International Journal of Soft Computing and Engineering* (IJSCE), 2(2), May 2012. ISSN: 2231 2307.

Bryce T., 2012, *The World of the Neo-Hittite Kingdoms. A Political and Military History*, Oxford University Press, Oxford-New York.

d'Alfonso L., 2012 "Notes on Anatolian Hieroglyphic Palaeography: An Investigation of the Sign *439, wa/wi", in Cotticelli-Kurras P. et al. (eds.)*, Interferenze linguistiche e contatti culturali in Anatolia tra II e I millennio a.C. Studi in onore di Onofrio Carruba in occasione del suo 80° compleanno*, Italian University Press, Pavia, 87-106.

d'Alfonso L. and Payne A., 2016, "The Palaeography of Anatolian Hieroglyphic: new perspectives", in *Journal of Cuneiform Studies* 68, 107-127.

Giusfredi F., 2010, *Sources for a Socio-Economic History of the Neo-Hittite States*, Winter, Heidelberg.

Hawkins J.D., 2000, *Corpus of Hieroglyphic Luwian Inscriptions, Volume. I: Inscriptions of the Iron Ages*, De Gruyter, Berlin-New York.

—. 2003, "Scripts and Texts", in Melchert H.C. (ed.) 2003, 128-169.

Laroche E., 1960, *Les hiéroglyphes hittites 1. L'écriture*, Paris.

Marazzi M. (ed.), 1998, *Il geroglifico anatolico. Sviluppi della ricerca a venti anni dalla sua "ridecifrazione"*, Istituto Universitario Orientale - Dipartimento di Studi Asiatici - Series minor, Napoli.

Melchert H.C. (ed.), 2003, *The Luwians*, Brill, Leiden-Boston.

Mouton A., Rutherford I. and Yakubovich, I. (eds.), 2013, *Luwian Identities – Culture, Language and Religion Between Anatolia and the Aegean*, Brill, Leiden-Boston.

Payne A., 2015, *Schrift und Schriftlichkeit - Die anatolische Hieroglyphenschrift*, Harrassowitz, Wiesbaden.

Rieken E., 2015, "Bemerkungen zum Ursprung einiger Merkmale der anatolischen Hieroglyphenschrift", in *Welt des Orient* 45, 216-231.

Rieken E. and Yakubovich I., 2010, "The New Values of Luwian Signs L 319 and L 172", in Singer I. (ed.), *ipamati kistamati pari tumatimis – Luwian and Hittite Studies Presented to J. David Hawkins on the Occasion of His 70th Birthday*, Tel Aviv, 199-219.

Simon Z., 2010, "A list of Iron Age Luwian inscriptions since CHLI", in Egedi B. and Simon Z., *Agyagtábla, papirusz. Az ókori Kelet és Egyiptom - kötetlenül és tudományosan*, available at http://agyagpap.blogspot.com/2010/10/list-of-iron-age-luwian-inscriptions.html, 10/31/2010 (accessed march 18th 2017).

Weeden M., 2013, “After the Hittites: The Kingdoms of Karkamish and Palistin in Northern Syria ”, in *BICS* 56-2 (2013), 1-20.
Yakubovich I., 2010, *Sociolinguistics of the Luvian Language*, Brill, Leiden-Boston.

THE DIGITAL PHILOLOGICAL-ETYMOLOGICAL DICTIONARY OF THE MINOR ANCIENT ANATOLIAN LANGUAGES AND ITS CONTRIBUTION TO THE ADVANCEMENT OF DIGITAL ONLINE ENCYCLOPAEDIAS

MARKUS FRANK AND ZSOLT SIMON

Abstract

The first part of this paper introduces the concept of this dictionary and its structure from a linguistic point of view. The second part describes how this concept is realized digitally.

1. Goals and Expectations

The goal of the "*Digital Philological-Etymological Dictionary of the Minor Ancient Anatolian Corpus Languages*" (suggested abbreviation: eDiAna), as expressed in its title, is twofold: synchronic and diachronic. On the synchronic level it will provide a full critical dictionary (thesaurus) of all languages belonging to the Anatolian branch of the Indo-European languages, except Hittite (i.e. Palaic, Luwian, Lydian, Lycian A, Lycian B, Carian, Sidetic, Pisidian) and on the diachronic level an exhaustive critical treatment of the etymology of these lexemes.

There are several reasons that necessitated this enterprise. First, unlike Hittite, which is adequately treated by two ongoing dictionary projects (the *Chicago Hittite Dictionary* and the second edition of the *Hethitisches Wörterbuch*), the existing glossaries of these languages are by far neither exhaustive nor up-to-date. This causes serious problems not only for philologists and linguists working with these languages but also for scholars of fellow disciplines encountering texts in these languages (not to mention the needs of students). Second, due to these circumstances, the currently available etymological treatments of these lexemes are well

below the standards of Indo-European linguistics. This can not only mislead fellow Indo-European specialists, but also prevent access to a valuable source for their research.

Accordingly, this dictionary will present all known lexemes of these languages, together with all of their attestations, where all attestations have been revised (based either on the original medium or on high-resolution photographs, depending on accessibility) and their meaning and morphology have been critically analysed taking into account the full published literature. In "lexemes" we include all parts of speech, but not the onomastic material and the inflectional and derivational morphemes. An exception is made if the onomastic material served to infer the existence of a lexeme: then it will be included and critically evaluated (this occurs for the most part in the case of the alphabetic languages and in the Old Assyrian transmission of Luwian). This will provide a firm, critically evaluated synchronic base that can and should serve for etymological discussion.

The etymological part of this dictionary will provide an exhaustive discussion of the historical phonology, morphology and etymology of the lexemes (if applicable), taking the full published literature account of. All possible levels will be reconstructed (Proto-Indo-European, Proto-Anatolian, and even Proto-Luwic in case of the Luwic lexemes). The etymological treatment follows the standard, school-neutral views of Comparative Indo-European Linguistics (e.g. three laryngeals and no glottalic theory).

It is obvious that the preparation of the entries of the lexemes of these poorly or not very well known languages can lead (and has already led) to new results in a wide variety of fields – especially since it requires painstaking philological work on the texts themselves before any entry can be written. The following listing of some of these possible advances (not exhaustive) presents some preliminary results of our team. Interested parties can thus expect advances not only in lexicographical matters (cf. Simon 2016e, 2016i; Yakubovich 2016) and synchronic grammar (cf. Rieken 2017), but also in understanding the texts (cf. Busse and Simon 2016; Simon 2016-2017) and writing systems (cf. Simon 2016g, 2016h) themselves. Furthermore, new etymologies both within Anatolian and on the Indo-European level (cf. Rieken and Yakubovich 2016; Simon 2016d, 2016f) as well as advances in the historical phonology and morphology of these Anatolian languages (cf. Boroday and Yakubovich in press; Sideltsev and Yakubovich 2016; Simon 2016b; Sasseville 2017) and in

Indo-European linguistics in general (cf. Simon in press-a) can be expected. Finally, the revision of special subcorpora of Luwian (e.g. lexemes attested in or inferred from Old Assyrian texts (cf. Simon 2015, 2016a, 2016c, in press-b) or the so-called *Glossenkeilwörter* (cf. Busse 2016), cf. in general below) may lead to a better understanding of the sociolinguistics of Luwian.

2. The Structure of the Dictionary and its Entries

The lexemes in the dictionary are grouped together on etymological grounds, i.e. those lexemes of the daughter languages are to be found within one entry that continue the same Proto-Anatolian lexeme. Words that are built upon these lexemes at any stage receive their own entries. Needless to say, a large number of words do not have morphologically exact cognates or identified cognates in general and thus they build an entry on their own. This is reflected by the head of the entry, which thus shows either the highest available reconstruction or the lexeme itself. Note that the highest level of reconstruction used for the heads is Proto-Anatolian, its Proto-Indo-European forerunner is discussed within the entry (see below). For these structural reasons and since this is an online dictionary, an elaborated search tool has been built in order to find the lexemes and their reconstructions instead of a table of contents or indices. Readers of course may want to check the related words as well and, accordingly, there is a hyperlink within the entries for derived and compound words (cf. below).

The entries have a fully unified structure. Under the head one can find the two main parts: the data of the daughter languages and the Indo-European etymology. All daughter languages as well as the Anatolian reconstruction have a separate subentry (if there is a Hittite cognate, its short description with references to the standard lexica and handbooks is also added). More precisely, next to Palaic, Lydian, Lycian A, Lycian B, Carian, Sidetic, Pisidian and "Reconstruction" that do not need any comment the following subentries need explanation: the different types of Luwian and the subentry called "Miscellaneous".

Since Luwian has been transmitted in quite different sources, a separation of these categories was inevitable. The existence of the separate groups of "Luwian in Old Assyrian Transmission", "Hieroglyphic Luwian" and "Cuneiform Luwian" does not need any comment. However, the exact definition of Cuneiform Luwian does: contrary to the widespread practice

(exemplified for instance by the Cuneiform Luwian dictionary of Melchert 1993), this category includes *only* those lexemes that are attested in Luwian sentences: Luwian words in Hittite sentences and the so-called *Glossenkeilwörter* are not included. This was necessary for the following reasons: first, the lexemes in both excluded groups require separate investigation regarding their Luwian status (if they are Luwian at all). This is accomplished within the entry (cf. below) and, incidentally, our preliminary results show that many of these alleged words are not Luwian. Second, both groups represent a corpus that requires separate research as such (i.e. what are the *Glossenkeilwörter* exactly? Exactly what kind of Luwian lexical influence reached Hittite and when?). Accordingly, "Luwian in Hittite Transmission" and "Glossenkeilwörter" received separate subentries. The latter includes all instances of lexemes marked with gloss wedges and all those instances of these specific lexemes where the gloss wedges are missing. In other words, "Luwian in Hittite Transmission" includes only those lexemes of Hittite texts that are never marked with gloss wedges and thus supposed to be Luwian for reasons other than the gloss wedges.

"Miscellaneous" includes all lexemes that cannot be attributed to specific languages: they are Anatolian words preserved in external sources as well as the assumed loanwords from Anatolian languages. Needless to say, a critical investigation of all the "Miscellaneous" entries is crucial (cf. below).

The entries of these separate languages and corpora are also fully unified and consist of the following parts: Transmission, Forms, Graphic Features, Meaning, Stem, Compounds, Derivatives, Origin. They will be filled with information only if they are applicable (typically "Compounds" and "Derivatives" may remain empty in lack of such lexemes). Accordingly, "Transmission", "Forms" and "Meaning" appear in all entries, "Origin" only in those where their status needs a critical evaluation (typically in "Luwian in Old Assyrian / Hittite Transmission", "Glossenkeilwörter" and "Miscellaneous"). "Transmission" provides a short overview of when and in which texts the specific lexeme is attested. Under "Forms" all attestations are listed in transcription with grammatical analysis and with the siglum of the text (and usually followed by further comments). This siglum is linked to the textual database, where all texts of these languages can be found digitalized and grammatically annotated (see below). Hovering over the siglum with the mouse, the quoted sentence appears on the bottom of the screen. "Graphic Features" discusses features that do not

immediately affect the forms, "Stem" will be implemented if the stem is not immediately obvious from the head of the subentry.

The etymological entries are unified and consist of five subentries: Indo-European Transmission, Semantic Reconstruction, Morphological Reconstruction, Synctactic Reconstruction and Phraseology, and Additional Literature for Transmission, where the latter two categories will provide information only if applicable. The names of these subentries are mainly self-explanatory; "Indo-European Transmission" means an overview of the related lexemes in other Indo-European languages, where the morphologically identical lexemes and the other lexemes from the same root but with different morphology are discussed separately.

It is important to note that all parts of the subentries are signed by the responsible author(s), unlike in other dictionaries in our field resulting from teamwork (e.g. the above quoted Hittite dictionaries). The advantages and the necessity of this practice are evident: on the one hand, this facilitates the quotation and the communication with the author(s). On the other hand, it is inevitable in case of such hardly understood languages that even the authors of the dictionary disagree in some points, thus it also stresses that the other authors do not necessarily share the views expressed in the subentry and thus they cannot be held responsible for them.

3. Team and Workflow

In order to achieve these goals an international team has been set up from the specialists of Ancient Anatolian languages and Comparative Indo-European Linguistics at the universities of Munich and Marburg (both in Germany, EU) financed by the Deutsche Forschungsgemeinschaft (DFG). The team consists of three linguistic groups and an IT-specialist. Each group is responsible for specific tasks. The first team is responsible for the synchronic entries of Cuneiform Luwian, Luwian in Hittite Transmission, Luwian in Old Assyrian Transmission, *Glossenkeilwörter*, Carian, Sidetic and the "Miscellaneous" entries.[1] The second team is responsible for the synchronic entries of Hieroglyphic Luwian, Palaic, Lycian A, Lycian B, Lydian and Pisidian as well as for the reconstructions up to the Proto-Anatolian level.[2] The third team is responsible for the Proto-Indo-

[1] Jared L. Miller, Zsolt Simon, and Anja Busse (Institut für Assyriologie und Hethitologie, Ludwig-Maximilians-Universität München).

[2] Elisabeth Rieken, Ilya Yakubovich, and David Sasseville (Fachgebiet Vergleichende Sprachwissenschaft und Keltologie, Philipps-Universität Marburg).

European reconstructions.[3] The fourth one is responsible for the digital implementation, conceptualization and design, to be discussed in detail in the second part of this paper.[4]

The workflow is straightforward: first, the synchronic entries of the selected related lexemes are prepared by the responsible teams. After team-internal revision, they are submitted to the team responsible for the Proto-Anatolian reconstruction, whose results are forwarded to the third team, responsible for the Proto-Indo-European reconstruction. All phases are accessible to all members enabling quick and effective cross-checking and improvement. If a group of lexemes is considered to be complete, it will be entered into the online dictionary, which provides one more opportunity to revise and update the entries.

The first round of entries to be written has been selected for practical purposes, i.e. those roots that appear in most daughter languages and have a reasonably assured etymology, since they could thus help us to test and improve all aspects of the digital background of the dictionary and provide a snapshot of it. Since then the cross-links, the linguistic problems encountered, and the ongoing work on the languages and corpora will have led to a continuous flow of new entries to be written, thus the expansion of the dictionary will gain momentum after its initial undertaking.

4. The Digital Implementation

4.1. The Underlying Data Structure

According to its digital conception, the dictionary consists of a series of database modules that are used to create the desired output formats in an "on-the-fly" fashion. At its core lies the lemma database, in which every single lemma chapter is saved in a decentralized manner. Parallel to this core database is the corpus database, which provides all the annotated Ancient Anatolian texts. The corpus database is capable of enriching each single lemma with concrete text information if necessary, but the corpus can also be used independently of the dictionary as a pure source of text. Likewise, parallel to the lemma core database is a bibliographic database,

---

[3] Olav Hackstein, Thomas Steer, and Andreas Opfermann (Lehrstuhl für Historische und Indogermanische Sprachwissenschaft, Ludwig-Maximilians-Universität München).

[4] Markus Frank (IT-Gruppe Geisteswissenschaften, Ludwig-Maximilians-Universität München).

which contains all the bibliographic data used in the context of the project. Just as in the case of the corpus database, the bibliographic database can be used to add bibliographic information to existing dictionary entries but also serves separately as a pure digital bibliography. For internal purposes only an excerpt database can be bound to the bibliographic database in order to accelerate the workflow when creating new lemmata. The overall structure of the dictionary is the result of the combined interaction of each database module and can be diagrammed as displayed in figure 1.

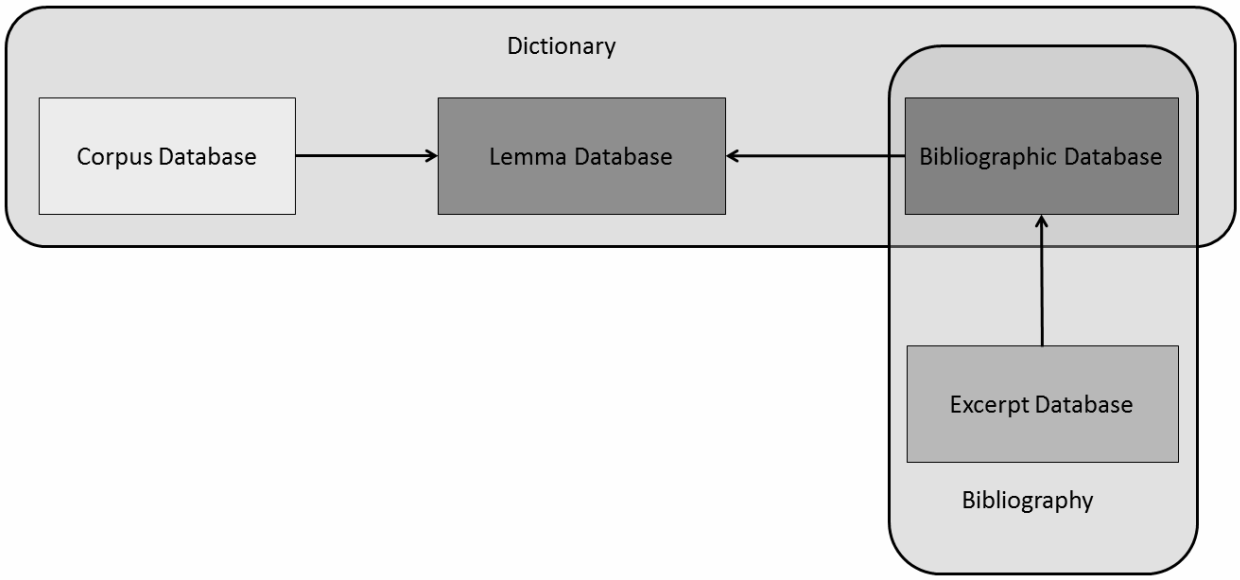

Figure 1. Central database model of the eDiAna

4.1.1 Corpus Database

The Luwian texts of the corpus, including their data structure, are based on the 'Annotated Corpus of Luwian Texts' (ACLT), which was developed and annotated by Ilya Yakubovich and Timofey Arkhangelskiy. Further text corpora of additional Anatolian daughter languages will be integrated into the corpus database as the project continues, thereby maintaining the data structure of the Luwian text corpora for reasons of compatibility.

Within the corpus database two tables are assigned to each language corpus: a corpus table, in which the whole corpus is saved in the form of a token list, as well as a meta table that provides meta information about the specific text in the corpus. Corpus and meta table are synchronised via a common ID row in both tables. Figure 2 exemplifies the table structure of one of the language corpora.

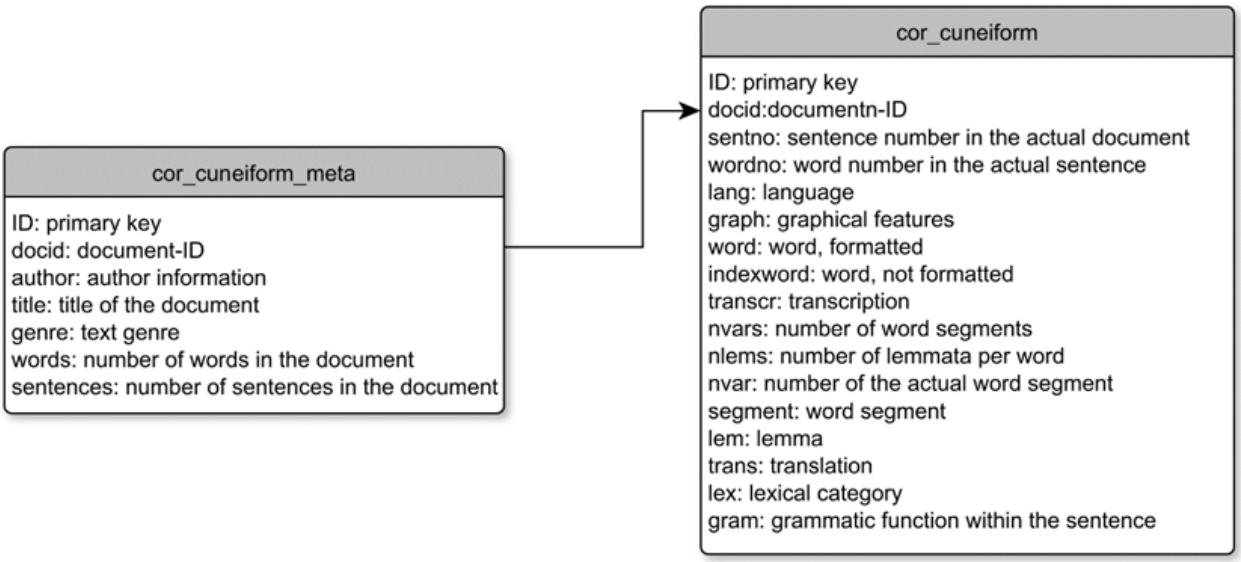

Figure 2. Corpus table structure exemplified by Cuneiform Luwian

The transparency of the data structure makes it possible to search the corpus for text, words, word segments, lemmata, translations, lexical categories and grammatical functions both individually and combined. Additionally, the precise reference system allows for inclusion of documents, sentences and clauses into the dictionary dynamically.

4.1.2 Lemma Database

The data structure of the lemma database represents an exclusively in-house development within the scope of eDiAna. The aim of this database structure is to save the lemma elements as efficiently and also as performative as possible. To fulfil this purpose, lemmata are not saved as a huge continuous XML-text but rather in a decentralised way by chapter and by language. Each lemma is assigned a so called *lemma head*, which provides the centre of each lexicon entry and serves as the suspension structure for the Anatolian daughter languages and the Proto-Indo-European reconstruction. This suspension structure saves the reconstructed (Proto-Anatolian) forms of the lemma along with the grammatical information, a morphological classification and a translation, as well as the existence or, respectively, absence of the lexicon parts of each language/corpus. Once a lemma is retrieved, its scope and the additional data tables that have to be loaded are determined on the basis of the central suspension structure. A parallel table containing each reconstructed (Proto-Anatolian) form is also part of the suspension structure, as these forms cannot be part of the primary suspension structure due to the fact that several alternate reconstructions are permitted per lemma and the